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!

Editorial Board

CONTENTS

LXPER Index: A Curriculum-specific Text Readability Assessment Model for EFL Students in Korea

Bruce W. Lee[1,2]
[1]Research & Development Center
LXPER, Inc.
Seoul, South Korea

Jason Hyung-Jong Lee[1]
[2]Department of Computer & Information Science
University of Pennsylvania
Philadelphia, PA, USA

Abstract—**Automatic readability assessment is one of the most important applications of Natural Language Processing (NLP) in education. Since automatic readability assessment allows the fast selection of appropriate reading material for readers at all levels of proficiency, it can be particularly useful for the English education of English as Foreign Language (EFL) students around the world. However, most readability assessment models are developed for the native readers of English and have low accuracy for texts in non-native English Language Training (ELT) curriculum. We introduce LXPER Index, which is a readability assessment model for non-native EFL readers in the ELT curriculum of Korea. To measure LXPER Index, we use the mixture of 22 features which we prove to be significant in text readability assessment. We also introduce the Text Corpus of the Korean ELT Curriculum (CoKEC-text), which is the first collection of English texts from a non-native country's ELT curriculum with each text's target grade level labeled. In addition, we assembled the Word Corpus of the Korean ELT Curriculum (CoKEC-word), which is a collection of words from the Korean ELT curriculum with word difficulty labels. Our experiments show that our new model, trained with CoKEC-text, significantly improves the accuracy of automatic readability assessment for texts in the Korean ELT curriculum. The methodology used in this research can be applied to other ELT curricula around the world.**

Keywords—Natural language processing; machine learning; text readability assessment; EFL (English as Foreign Language) education

I. INTRODUCTION

Readability Assessment helps quantify the level of difficulty that a reader experiences in comprehending a certain text. Since automatic readability assessment enables the convenient selection of appropriate reading material for readers with different levels of proficiency [1], readability assessment has been an important field of research since as early as the 1940's [11]. Since then, more than 200 readability formulas were developed [3], but most of them concentrated on the general audience in the United States. We argue that there is a need for the development of an improved text readability assessment method for use in English as Foreign Language

(EFL) education around the world.

In China, Japan, and South Korea, many high and middle school students, in addition to their regular classes, also attend English language schools. English education plays an extremely important role in the national educational systems and college entrance exams of the three countries [23], [25]. Furthermore, it is estimated that more than $3 billion is spent annually on English education in South Korea, and in Japan, it is much more [23]. Despite the amount of importance put in English education in such countries, the automatic text readability assessment method has not been in active use. This is mostly because of the low level of accuracy of the traditional readability assessment formulas for use in an international EFL curriculum, which we will prove later in this paper.

Previous work in automatic readability assessment has focused on analyzing the generic features of a text. For example, Flesch-Kincaid readability tests use variables like total words, total sentences, and total syllables to identify the difficulty of a text [20]. Such features are essential, but we argue that more curriculum-specific features are required for use in EFL education of non-native students. In addition to the traditional method of calculating generic features of a text like average number of words per sentence, average number of syllables per word, average number of noun phrases per sentence, we model the cognitive characteristics and the expected level of the vocabulary of a user group. Implementing cognitively motivated features, which operate at the discourse level of a text, has proven to be efficient in predicting readability for adults with Intellectual Disabilities (ID) [14], but no research has been conducted using corpora from an EFL curriculum.

Obtaining well-formatted graded *corpora* is one of the most difficult tasks in conducting modern readability assessment research using Natural Language Processing (NLP) techniques. We managed to collect graded corpora that match the English Language Training (ELT) curriculum in Korea. The results we obtain in this research are mainly based on the Korean ELT curriculum, but the novel techniques and features that we introduce are applicable to any other country with an EFL curriculum.

The contributions of this paper are: (1) we utilize a novel graded corpus of texts from an actual EFL curriculum; (2) we

present the possibility of using a graded corpus of words (that we manually assembled with the help of 20 English teachers) for curriculum-specific optimization; (3) we test the efficiency of discourse-level text analysis for readability assessment in EFL curriculum; (4) we introduce novel readability assessment features for word-level text evaluation; (5) we evaluate the accuracy of our new model in predicting the readability of texts used in non-native ELT curriculum.

The overall structure of our paper is as followed: Section 2 explains some of the past works related to our topic of study, Section 3 describes our hypothesis and briefly outlines the methods we use in our study, Section 4 describes the two corpora that we built, Section 5 describes our linguistic features used in LXPER Index, Section 6 evaluates our index against other popular readability assessment models, and Section 7 includes some of our future plan to improve the accuracy of LXPER Index.

II. RELATED WORK

Many readability metrics are measured by calculating a number of shallow features of a text, which include total words, total sentences, and total syllables [20], [24]. However, as later studies proved, such shallow measures are not directly related to the linguistic components that determine readability [10]. Even though these traditional readability metrics are still being used, they can easily misrepresent the complexity of academic texts [12].

Unlike the readability formulas in the past, most recent studies on text readability assessment using machine learning-based approaches [2], [6], [14]. Reference [29] pioneered the statistical approach to readability assessment but the research stopped at applying simple unigram language models to estimate the grade level of a given text. In contrast, modern readability assessment methods often analyze more than 20 features and explore a broader set of linguistic features [15]. Meanwhile, a variety of machine learning frameworks has been explored, but it was proved that the improvement resulting from changing the framework is smaller than that from optimizing the features [18].

Most work on readability assessment has been directed at estimating the reading difficulty for native English learners. Several efforts in developing automated readability assessment techniques have only emerged since 2007 [32]. Reference [17] proved that grammatical features play a more important role in EFL text readability prediction than native English curriculum. Reference [30] showed that the additional use of lexical features (which we also use in this research) has a significant positive effect on EFL readability assessment. However, the common limitation of the previous research in EFL readability assessment was the use of textual data annotated with the readability levels for native readers of English, not EFL readers. The study of automatic readability assessment for EFL students is still in its early stages.

III. HYPOTHESIS AND METHODS

We hypothesize that the accuracy of EFL text readability assessment can be improved by adding entity calculation and curriculum-specific vocabulary features. EFL readers have limited exposure to English compared to native readers. As a result, EFL readers would have to work harder at connecting each entity to a semantic relationship, compared to the average native readers. In addition, we believe that the biggest difference between native text readability assessment and EFL text readability assessment is that EFL students strictly follow the specific national ELT curriculum. Unlike native students who learn English from a variety of primary and secondary sources, most EFL students learn English as a school academic subject. Thus, we believe that the performance of EFL text readability assessment will greatly improve with the consideration of the specific national curriculum.

To test our hypothesis, we used the following methodology. We collected two corpora (explained in detail in Section 4). The first is CoKEC-text (Text Corpus of the Korean ELT Curriculum), which we created by putting together the texts approved or administered by the Korean Ministry of Education (MOE). We collected the texts that appeared in the National Assessment of Educational Achievement (NAEA), College Scholastic Ability Test (CSAT), and government-approved middle school textbooks. The second is CoKEC-word (Word Corpus of the Korean ELT Curriculum). We manually assembled the corpus with the help of 30 teachers with more than 20 years of experience in teaching EFL students in the Korean ELT curriculum. We classified 59529 words that appeared in the Korean ELT curriculum. Both CoKEC-text and CoKEC-word corpora only contain the texts/words from the official ELT curriculum in Korea. We then analyzed the significance of each feature on CoKEC-text. Finally, we combined the significant features into a linear regression model and experimented with a number of feature combinations.

IV. CORPORA

To test how our linguistic features measure the readability of a text, we collected two English corpora (CoKEC-text and CoKEC-word). Since our goal is to perform a more accurate readability assessment for EFL texts, an ideal corpus for our research must contain texts from a non-native ELT curriculum – in particular, if such texts are labeled with target grade levels. We are not aware of such texts electronically available, so we have decided to collect the texts ourselves. The texts come from government, online, and commercial sources.

A. Graded Text Corpus: CoKEC-Text

It is extremely rare to encounter a corpus in which texts are labeled as being at specific levels of readability. The Weekly Reader corpus [31] is one of the only electronically available text corpus with target grade level information, but the corpus is not available anymore since 2012, as the publisher became a part of the Scholastic Corporation. In addition, the corpus was annotated with the readability levels for native readers of English, so such a corpus is not suitable to this research. We had no choice but to build grade annotated corpora ourselves to continue developing LXPER Index.

Our first corpus, which we refer to as CoKEC-text, is a collection of 2760 unique texts that are officially approved or administered by the Korean Ministry of Education. The texts are from NAEA, CSAT, and government-approved textbooks. We have been collecting the texts for about 10 years, from 2010 to 2020. Each text is labeled with its target grade level

(grade 7: 17 texts, grade 8: 215 texts, grade 9: 80 texts, grade 10: 571 texts, grade 11: 596 texts, grade 12: 590 texts, grade 12.5: 691 texts). Grade 12.5 refers to the English texts that were used in CSAT, which is a college entrance exam for Korean universities. (Korean grades 7 to 12 are for middle and high school students of ages 13 to 19.)

B. Graded Word Corpus: CoKEC-Word

The classification of word difficulty has been a field of research for as long as the text readability assessment. Thus, there are a number of electronically available word corpus, which include the British National Corpus, gathered by Lancaster University and Cambridge [4], and the Corpus of Contemporary American English [9]. Even though these corpora do not contain target grade level classification, a number of methods to identify the word difficulty has been explored using the already available corpora [21].

Reference [13] identified the "easier" words by calculating each word's Kucera-Francis frequency in the psycholinguistic dictionary [28]. Like so, most research on word difficulty has been based on the frequency that a certain word is used. This is under the assumption that the more frequently used words are more familiar to the native readers of English. However, such a method is inapplicable to EFL students because most have comparatively limited exposure to English, which is highly dependent on their national ELT curriculum.

To measure the word difficulty intended by a certain ELT curriculum (we chose the Korean ELT curriculum due to the ease of access), we gathered 30 English teachers with more than 20 years of teaching experience in Korea. Our second corpus, CoKEC-word, is a collection of 59529 words that appeared in the Korean ELT curriculum from 2010 to 2020. Out of the 59529 words, 30608 words are classified into six categories (A: suitable for grade 1 to 4 students – 1315 words, B: grade 5 to 8 students – 1365 words, C: grade 8 to 9 students – 3103 words, D: grade 9 to 11 students – 5269 words, E: grade 11 to 12 students – 7677 words, F: college students – 11879 words). The other 28921 words are unclassified as they are proper nouns or abbreviations. Because we are particularly interested in improving the accuracy of readability assessment for the use of EFL students, we focus on the first section (classified into 6 categories) of CoKEC-word.

V. Linguistic Features and Readability

In this section, we describe the set of features that we used for readability assessment. Table I is the list of the features and their code names. The list is divided into simple features, cognitively motivated features, and word difficulty features.

A. Simple Features

We start by implementing some simple features from Flesch-Kincaid metrics [20]: aWPS (average number of Words per Sentence), aSPW (average number of Syllables per Word), and M3S (percentage of words with more than 3 syllables).

Then we also implemented features from other previous research that have proven to be particularly useful in Machine Learning (ML) text readability evaluation. Reference [26] calculated the following features using the Charniak parser [5]: aNP, aNN, aVP, aAdj, aSBr, aPP, nNP, nNN, nVP, nAdj, aSBr,

nPP (description in Table I). The linguistic features that were identified in [26] are still useful, but there has been a massive improvement in the tree-parsing technology. We used the Berkeley Neural Parser [19], a constituency parser, which proved to identify the linguistic features at a higher accuracy than the Charniak parser.

TABLE I.   Features

Count	Code	Description
Simple Features		
1	aWPS	average number of Words per Sentence
2	aSPW	avg num of Syllables per Word
3	aNP	avg num of Noun Phrases per sentence
4	aNN	avg num of proper nouns per sentence
5	aVP	avg num of Verb Phrases per sentence
6	aAdj	avg num of Adjectives per sentence
7	aSBr	avg num of Subordinate Clauses per sentence
8	aPP	avg num of Prepositional Phrases per sentence
9	M3S	% of words with more or equal to 3 syllables
10	nNP	total num of Noun Phrases per sentence
11	nNN	total num of proper nouns per sentence
12	nVP	total num of Verb Phrases per sentence
13	nAdj	total number of Adjectives per sentence
14	nSBr	total num of Subordinate Clauses per sentence
15	nPP	total num of Prepositional Phrases per sentence
Cognitively-Motivated Features		
16	nUE	total number of Unique Entities
17	aEM	avg num of Unique Entity mentions per sentence
18	aUE	avg num of Unique Entities per sentence
19	nLC	total num of Lexical Chains
20	aLCw	avg num of Lexical Chains per word
21	aLCn	avg num of Lexical Chains per noun phrase
Word Difficulty Features		
22	aCw	avg num of level C (grade 8-9) words per word
23	nCw	total num of level C (grade 8-9) words
24	aDw	avg num of level D (grade 9-11) words per word
25	nDw	total num of level D (grade 9-11) words
26	aEw	avg num of level E (grade 11-12) words per word
27	nEw	total num of level E (grade 11-12) words
28	aFw	avg num of level F (college-level) words per word
29	nFw	total num of level F (college-level) words

B. Cognitively-Motivated Features

The cognitively motivated features used in this research are largely influenced by [14], which proved the usefulness of cognitive features in predicting text readability for adults with Intellectual Disabilities (ID). We believe that EFL students and native adults with ID are similar in the sense that they both face difficulty in the semantic encoding of new information. Among the 10 cognitive features from [14], we implemented 6 features that are applicable to EFL students as well. We test the significance of these features on CoKEC-text in Section 5D.

C. Word Difficulty Features

The biggest difference between EFL students and native English readers is that the respective national ELT curriculum is the only exposure to English for most EFL students. EFL students learn new English words step by step in accordance with the curriculum. On the other hand, native English readers learn vocabulary from a variety of sources. Thus, we believe that the curriculum-specific features related to vocabularies can be particularly useful in predicting the text difficulty for EFL students.

In CoKEC-word, we classified 30608 words into six levels. We focused on the average and total number of vocabularies in levels C, D, E, and F (appropriate for students in grade 8 to college level): aCw, nCw, aDw, nDw, aEw, nEw, aFw, nFw. In

Section 5D, we prove that some of these features have the high Pearson correlations with the target grade level of texts in the Korean ELT curriculum.
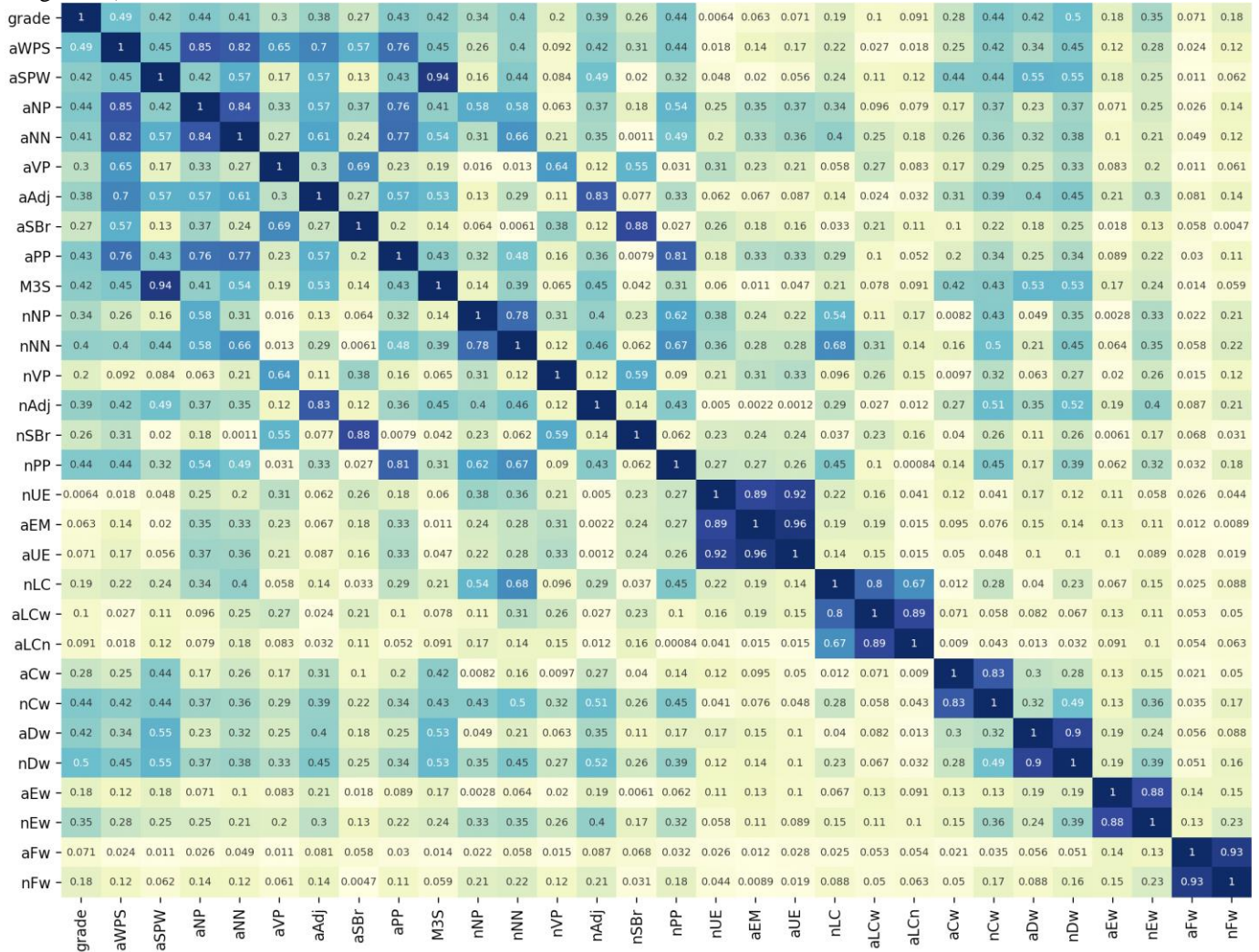
Fig 1.   Correlation Heatmap.

D. Testing the Significance of Features

To select features to include in our readability assessment model, we analyzed the texts in CoKEC-text to make sure that the results that we get are applicable to a non-native EFL curriculum. We calculated the value of each feature in CoKEC-text (the result is shown in Fig. 1), and we used Pearson correlation to test if a feature was significant enough (correlation above 0.05) in predicting the target grade level of a text.

The "Cor" column in Table II contains the correlation value of each feature. We put "Yes" in the "Sig" column if the feature was significant enough in predicting the target grade level of a text in CoKEC-text. Meanwhile, "No" means that the feature will not be included in our final readability assessment model because the feature is not significant enough. The only feature that did not show a significant correlation between the feature and the target grade level was the number of Unique Entity (nUE). The lack of significance for nUE can be

explained by the repeated mentions of similar entities but less unique entities in higher grade texts.

Moreover, discourse-level features (cognitively motivated features) generally had low correlations compared to simple features and word difficulty features. It can be interpreted that the Korean ELT curriculum puts more emphasis on the difficulty of a sentence structure and the difficulty of each vocabulary, rather than discourse-level analysis.

E. Removing Highly Correlated Features

To simplify and stabilize our regression model, we decided to remove the features that are highly correlated with each other (correlation above 0.85). The highly correlated pairs that we found were: M3S & aSPW, nSBr & aSBr, aEM & nUE, aUE & nUE, aLCn & aLCw, nDw & aDw, nEw & aEw, nFw & aFw (also showed in Table II "Pair" column). From each pair, we chose the feature that has a higher correlation with the

target grade level (in Section 5D): aSPW, aSBr, aEM, aUE, aLCw, nDw, nEw, nFw.

The features that we include in our final readability assessment model are shown in the "Include?" column of Table II. In Table III, the selected features are reorganized in the order of importance (from high to low correlation) in predicting a text's target grade level.

TABLE II.   FEATURES SELECTION PROCESS

Count	Code	Cor	Sig	Pair	Include?
Simple Features					
1	aWPS	0.494	Yes		Yes
2	aSPW	0.419	Yes	M3S	Yes
3	aNP	0.445	Yes		Yes
4	aNN	0.410	Yes		Yes
5	aVP	0.302	Yes		Yes
6	aAdj	0.381	Yes		Yes
7	aSBr	0.270	Yes	nSBr	Yes
8	aPP	0.432	Yes		Yes
9	M3S	0.419	Yes	aSPW	No
10	nNP	0.342	Yes		Yes
11	nNN	0.400	Yes		Yes
12	nVP	0.201	Yes		Yes
13	nAdj	0.392	Yes		Yes
14	nSBr	0.261	Yes	aSBr	No
15	nPP	0.442	Yes		Yes
Cognitively Motivated Features					
16	nUE	0.00643	No	aEM, aUE	No
17	aEM	0.0629	Yes	nUE	Yes
18	aUE	0.0705	Yes	nUE	Yes
19	nLC	0.190	Yes		Yes
20	aLCw	0.10196	Yes	aLCn	Yes
21	aLCn	0.0912	Yes	aLCw	No
Word Difficulty-related Features					
22	aCw	0.280	Yes		Yes
23	nCw	0.444	Yes		Yes
24	aDw	0.416	Yes	nDw	No
25	nDw	0.503	Yes	aDw	Yes
26	aEw	0.180	Yes	nEw	No
27	nEw	0.352	Yes	aEw	Yes
28	aFw	0.0714	Yes	nFw	No
29	nFw	0.180	Yes	aFw	Yes

* Features that are not used in our final model are colored red

TABLE III.   CHOSEN FEATURES IN THE ORDER OF IMPORTANCE

Rank	Code	Cor	Type
1	nDw	0.503	Simple Feature
2	aWPS	0.494	Word Difficulty Feature
3	aNP	0.445	Simple Feature
4	nCw	0.444	Word Difficulty Feature
5	nPP	0.442	Simple Feature
6	aPP	0.432	Simple Feature
7	aSPW	0.419	Simple Feature
8	aNN	0.410	Simple Feature
9	nNN	0.400	Simple Feature
10	nAdj	0.392	Simple Feature
11	aAdj	0.381	Simple Feature
12	nEw	0.352	Word Difficulty Feature
13	nNP	0.342	Simple Feature
14	aVP	0.302	Simple Feature
15	aCw	0.280	Word Difficulty Feature
16	aSBr	0.270	Simple Feature
17	nVP	0.201	Simple Feature
18	nLC	0.190	Cognitively-Motivated Feature
19	nFw	0.180	Word Difficulty Feature
20	aLCw	0.101	Cognitively-Motivated Feature
21	aUE	0.0705	Cognitively-Motivated Feature
22	aEM	0.0629	Cognitively-Motivated Feature

VI. READABILITY ASSESSMENT

After testing the significance of linguistic features and removing the highly correlated features, we used a linear regression model and trained it with CoKEC-text to build a readability assessment tool; our model is implemented using Python [27]. To evaluate the model's usefulness for EFL students in the ELT curriculum of Korea, we prepared a separate test corpus. The first part of our test corpus is from the official mock tests (pronounced "moi-go-sa" in Korean), used by the Korea Institute of Curriculum & Evaluation to assess educational achievement of high school students in 2019. There are 264 texts in the first part of our test corpus (grade 10: 88 texts, grade 11: 88 texts, grade 12: 88 texts). The second part of our corpus is from the government-approved middle school textbooks (grade 9: 79 texts). We intentionally collected the texts from two different sources to test how our readability assessment model works on different types of texts. Table IV shows the average number of words per text (aWPT), average number of sentences per text (aSPT), and average number of words per sentence (aWPS) for each grade level in the test corpus.

Next, we calculated the average error for each version of LXPER Index to choose which combination to use. Then we compared LXPER Index with five other popular traditional assessment models: Flesch-Kincaid grade level [20], Coleman-Liau Index [8], Dale-Chall Readability Score [11], Coh-Metrix EFL Index [16], and Lexile Measure [22].

A. Versions

We implemented seven versions of our readability assessment model, which are organized in Table V. The first uses only the simple features, which were studied extensively in previous research (aWPS, aSPW, aNP, aNN, aVP, aAdj, aSbr, aPP, nNP, nNN, nVP, nAdj, nPP). Meanwhile, the second version implements only the cognitively motivated features, which were proved to be useful for readability assessment on adults with ID but haven't been tested on EFL students [14] (aEM, aUE, nLC, aLCw). The third version implements the word difficulty features, which are our novel features in the readability assessment for EFL students (aCw, nCw, nDw, nEw, nFw). The fourth version uses both simple features and cognitively motivated features. The fifth uses cognitively motivated features and word difficulty features, while the sixth version uses simple features and word difficulty features. The seventh version combines all the sets of features. These versions are organized in the "Version" column of Table V as well.

TABLE IV.   TEST CORPUS

Description	Gr 9	Gr 10	Gr 11	Gr 12	All
aWPT	111.725	158.114	164.613	170.126	151.145
aSPT	14.275	8.682	9.409	9.229	10.398
aWPS	7.826	16.599	17.495	18.432	15.088

TABLE V.   TESTING COMBINATIONS

Version	Gr 9	Gr 10	Gr 11	Gr 12	AvgEr*
S	9.817	11.055	11.385	11.639	0.655
CM	10.478	11.019	11.361	11.397	0.865
WD	10.039	10.913	11.402	11.607	0.685
S&CM	9.727	11.056	11.395	11.634	0.636
CM&WD	9.894	10.927	11.455	11.616	0.665
S&WD	9.706	10.995	11.404	11.701	0.601
S&CM&WD	9.629	10.995	11.423	11.693	0.589

*Average Error

By building seven versions of our model, we can check if there's any certain combination that reduces the assessment error as much as possible. We can also measure the relative impact of implementing our word difficulty features.

Table V summarizes the average prediction results of our model for texts with different target grade levels. The average error value of our model decreased more than 0.05 grade level by adding the Word Difficulty features, compared to using only the simple features.

```
===========================================
                LXPER Index
-------------------------------------------
   paragraph1:  11.314610668825809
   paragraph2:  11.325568918902054
   paragraph3:  11.456926496551013
   paragraph4:  11.831492706888628
   paragraph5:  12.555917022434922
   paragraph6:  11.860643021177644
   paragraph7:  12.640597884093928
   paragraph8:  11.655660604408641
   paragraph9:  13.041646893081829
   paragraph10: 12.31356683819909
   paragraph11: 12.23811585459722
   paragraph12: 10.741342273960948
   paragraph13: 12.636533803957006
   paragraph14: 12.012046343461847
   paragraph15: 11.679116422326873
   paragraph16: 10.075361362432751
   paragraph17: 10.218809526336205
   paragraph18: 11.738985102882594
   paragraph19: 10.65425538041245
   paragraph20: 11.626751141441773
-------------------------------------------
   average: 11.680897413318661
   standard dev.: 0.804484551560757
===========================================
```

Fig 2.   Sample Output.

As we can see from Table III and Table V, cognitively motivated features do not seem to have as much effect as we expected in Section 3. Our explanation is that the non-native ELT curriculum simply puts more focus on the difficulty of words than discourse-level analysis. Even though the cognitively motivated features' correlation with target grade level is not as strong as we hypothesized, it seems that cognitively motivated features do improve the accuracy of our model after a few tests – including the one in Table V. We decided to move on with all three categories of features. The sample output of LXPER Index is shown in Fig. 2.

B. Comparison with Other Readability Tools

Like we did in Section 6A, we trained our regression models on CoKEC-text and tested it on the separate test corpus that we prepared. By doing so, we could make sure that our model is working properly on the texts from outside of CoKEC-text.

To comparatively evaluate how our model performs in measuring the target grade level of a text in the Korean ELT curriculum, we ran the same test on five other popular metrics. Using Python [27], we created calculator programs for Flesch-Kincaid [20], Coleman-Liau [8], and Dale-Chall [11] formulas. We used the electronically available tools (from the official source) to calculate Lexile Measure [22] and Coh-Metrix EFL Readability Index [7].

Lexile Measure and Coh-Metrix Index are built based on their unique scales. We initially wanted to rescale Lexile Measure and Coh-Metrix Index and compare all the models in one table, but such comparison could potentially misrepresent the intended results by the initial authors. Thus, we decided to create a separate table for Lexile and Coh-Metrix Index without rescaling. The results are organized in Table VI and Table VII. Columns Gr 9, Gr 10, Gr 11, and Gr 12 contain the average readability predictions of each assessment model for the specific part of our test corpus.

Flesch-Kincaid, Coleman-Liau, and Dale-Chall assessment models are almost entirely dependent on the shallow features like average number of words per sentence and average number of syllables per word. Meanwhile, the grade 10, grade 11, and grade 12 texts in our test corpus had similar aWPT, aSPT, and aWPS, as shown in Table III. As a result, the three assessment models fail to clearly distinguish the readability levels in grade 10 to grade 12 range in Table VI. On the other hand, the values show a sudden drop when predicting grade 9 part of our test corpus, which is also the part where aWPT, aSPT, and aWPS change drastically.

TABLE VI.   TESTING AGAINST OTHER MODELS

Model	Gr 9	Gr 10	Gr 11	Gr 12	AvgEr*
Flesch-Kincaid	5.625	10.572	9.636	9.207	2.026
Coleman-Liau	6.679	9.876	10.183	10.101	1.290
Dale-Chall	5.476	7.924	7.562	7.312	3.432
LXPER	9.629	10.995	11.423	11.693	0.589

*Average Error

TABLE VII.   TESTING AGAINST OTHER MODELS (LEXILE, COH-METRIX)

Model	Gr 9	Gr 10	Gr 11	Gr 12
Lexile(~Gr[*])	644(~Gr 3)	1064(~Gr 6)	1260(~Gr 10)	1120(~Gr 7)
Coh-Metrix	15.432	23.725	13.134	12.462
LXPER	9.629	10.995	11.423	11.693

[*]Expected Grade Level (according to Lexile website, 50th percentile, EOY Spring)

Dale-Chall Readability Score has a variable relating to the number of difficult words. Reference [11] collected a list of 768 words, labeled as "Difficult" or "Not Difficult" by surveying native fourth-grade students. It seems that the formula is unable to identify the difficult words for EFL students, which led to the overall prediction being lower than the target grade level.

Lexile Measure fails to show a constantly increasing trend from grade 9 to grade 12 parts of our test corpus. We believe that this is mostly because it is optimized for native students in the United States. The points where the specific national ELT curriculum defines the "difficulty" of a text are varied. The expected grade level in Table VII is estimated according to a graph on the Lexile website [22]. We followed their standards for the 50th percentile, End of Year (EOY) Spring values.

Coh-Metrix EFL Readability Index, out of the five models that we compared to LXPER Index, is the only model specifically designed for EFL students. It works under the assumption that psycholinguistic and cognitive features have great predictive power in readability assessment for EFL students [10]. The Coh-Metrix EFL Index is almost completely consisted of cognitive features. However, we have shown in Table III that they are not very important in readability assessment for EFL students – at least in the case of the Korean ELT curriculum. In Table VI, Coh-Metrix EFL Index shows a sudden decrease (higher values indicate easier-to-read passages) in the grade 9 section of our test corpus. We believe that this is mostly because the Korean middle school ELT curriculum (grade 7 to 9) mostly consists of conversation-based passages, where a lot of names and locations appear. This led to a sudden increase in the number of entities in the part of the corpus.

LXPER Index was the only assessment model that showed a continuously increasing pattern from grade 9 to grade 12. LXPER Index, with simple features, cognitively motivated features, and word difficulty features all combined, predicts the target grade level of texts in the Korean ELT Curriculum to within 0.589 grade levels on average.

VII. CONCLUSION

There have been several attempts to perform automatic readability assessment for EFL students, but the common limitation was that most past research was based on corpora with target grade levels for native students. In this research, we created LXPER Index, a readability assessment model that incorporates simple, cognitive, and word difficulty features and trained the model on CoKEC-text. As a result, we obtained a much more accurate EFL text readability prediction compared to the other assessment models available now.

We also proved the importance of word difficulty features in an EFL curriculum. On the other hand, cognitive features were not as highly correlated to a text's target grade level as we expected. Most importantly, LXPER Index was the only assessment model that showed a continuously increasing pattern from grade 9 to grade 12 in Table VI, which, we believe, is a significant achievement. The average error of 0.589 grade levels was a better performance than other assessment models, but we believe that the accuracy can be improved with further research.

In our future research, we believe that we can improve the accuracy of our model by implementing grammatical features. The size of CoKEC-text is another part that we should work on. CoKEC-text is currently the biggest collection of texts of the Korean ELT curriculum, but LXPER Index's accuracy can be improved with even more texts. The grade coverage is currently from grade 7 to grade 12. Including texts for lower grades is an approach that we should give an attempt.

We can also attempt regression techniques like logistic regression. In this research, we were not completely sure whether to consider the grade level label in CoKEC-text as a continuous variable or a categorical variable. It is possible that the differences in readability among the texts intended for grade 9, 10, 11, 12 are uneven. Even though we achieved a comparatively good average error value with linear regression, it might be improved by considering the grade level as a categorical variable.

REFERENCES

[1] Alhawiti, K.M. 2014. Natural Language Processing and its Use in Education. *International Journal of Advanced Computer Science and Applications*, 5(12), 72-76

[2] Aluisio, S., Specia, L., Gasperin, C., Scarton, C. 2010. Readability Assessment for Text Simplification. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 1-9

[3] Benjamin, R.G. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1), 63-88

[4] British National Corpus. 2014. http://corpora.lancs.ac.uk/bnc2014/

[5] Charniak, E. 2000. A maximum-entropy-inspired parser. *Proceedings of the 1st Meeting of the NAACL.*

[6] Chen, X. and Meurers, D. 2017. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3), 486-510

[7] Coh-Metrix 3.0. 2017. http://141.225.61.35/cohmetrix2017

[8] Coleman, M. and Liau, T.L. 1975. "A Computer Readability Formula Designed for Machine Scoring," Journal of Applied Psychology, 60(2) 283-284

[9] Corpus of Contemporary American English. 2020. https://www.english-corpora.org/coca/

[10] Crossley, S.A., Greenfield, J. & McNamara, D.S. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3), 475-493

[11] Dale, E. and Chall, J.S. 1949. The concept of readability. *Elementary English*, 26(23)

[12] Davison, A. and Kantor, R. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17(2), 197-209

[13] Devlin, S. and Tait, J. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, 161–173

[14] Feng, L., Elhadad, N., and Huenerfauth, M. 2009. Cognitively Motivated Features for Readability Assessment. *Proceedings of the 12th Conference of the European Chapter of the ACL*, 229-237

[15] Feng, L., Jansche, M., Huenerfauth, M., Elhadad, N. 2010. A Comparison of Features for Automatic Readability Assessment. *Proceedings of the 23rd International Conference on Computational Linguistics*, 276-284

[16] Graesser, A.C., McNamara, D.S., Lourwerse, M.M., and Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36, 193–202

[17] Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M. 2007. Combining lexical and grammatical features to improve readability measures for first and second language text. *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 460-467

[18] Kate, R.J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R.J., Roukos, S. and Welty, C. 2010. Learning to predict readability using diverse linguistic features. *Proceedings of the 23rd International Conference on Computational Linguistics*, 546-554

[19] Kitaev, N., and Klein, D. 2018. Constituency Parsing with a self-attentive encoder. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* 1, 2676–2686

[20] Kincaid, J.P., Fishburne, R.P., Rogers, R.L., and Chissom, B.S. 1975. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. *Research Branch Report,* 8-75

[21] Kucera, H. and Francis, W. 1967. Computational analysis of present-day American English. *University Press of New England*

[22] Lexile. 2020. https://lexile.com/

[23] Mckay, S.L. 2002. Teaching English as an International Language: Rethinking Goals and Perspectives. *OUP Oxford*

[24] McLaughlin, G.H. 1969. SMOG grading – a new readability formula. *Journal of Reading*, 12(8), 639-636

[25] Park, J.K. 2014. Teaching English as a Global Language in Korea: Curriculum Rhetoric and Reality. *Asian Englishes*, 12(1), 124-129

[26] Peterson, S.E., and Ostendorf, M. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23: 89-106

[27] Python Software Foundation. Python Language Reference, version 3.7. Available at http://www.python.org

[28] Quinlan, P. 1992. The Oxford psycholinguistic database, *Oxford University Press*

[29] Si, L. and Callan, J. 2001. A statistical model for scientific readability. *Proceedings of the tenth international conference on Information and knowledge management*, 574-576

[30] Vajjala, S. and Meurers, D. 2014. Readability assessment for text simplification: From analyzing documents to identifying sentential simplification. *International Journal of Applied Linguistics*, 165(2), 194-222

[31] Weekly Reader. 2008. http://www.weeklyreader.com

[32] Xia, M., Kochmar, E., Briscoe, T. 2016. Text Readability Assessment for Second Language Learners. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 12-22

Developing a Radiating L-shaped Search Algorithm for NASA Swarm Robots

Tariq Tashtoush*[1], Jalil Ahmed[2], Valeria Arce[3], Heriberto Dominguez[4], Kevin Estrada[5], William Montes[6],
Ashley Paredez[7], Pedro Salce[8], Alexia Serna[9], Mireya Zarazua[10]

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

Abstract—This paper focuses on designing a search algorithm that the DustySWARM team used in the 2019 NASA Swarmathon competition. The developed search algorithm will be implemented and tested on multiple rovers, a.k.a. Swarmies or Swarm Robots. Swarmies are compact rovers, designed by NASA to mimic Ants behavior and perform an autonomous search for simulated Mars resources. This effort aimed to assist NASA's mission to explore the space and discover new resources on the Moon and Mars. NASA's going-on project has the goal to send robots that explore and collect resources for analysis before sending Astronauts, as the swarm option is safer and more affordable. All rovers must utilize the exact algorithm and collaborate and cooperate to find all available resources in their search path and retrieve them to the space station location. Additionally, swarmies will autonomously search while avoiding obstacles and mapping the surrounding environment for future missions. This algorithm allows a swarm of six robots to search an unknown area for simulated resources called AprilTags (cubes with QR codes). The code was developed using C/C++, GitHub, and Robotics Operation Systems (ROS) and tested by utilizing the Gazebo Simulation environment and by running physical trials on the swarmies. The team analyzed a few algorithms from previous years and other researchers then developed the Radiating L-Shape Search (RLS) Algorithm. This paper will summarize the algorithm design, code development, and trial results that were provided to the NASA Space Exploration Engineering team.

Keywords—*NASA Swarmathon competition; swarm robotics; search algorithm; autonomous; Robot Operating System (ROS); NASA space exploration; simulation; autonomous robot swarm; collaborative robots; autonomous behavior; cooperative robots; swarmies; L-Shaped search; GitHub*

I. Introduction

National Aeronautics and Space Administration (NASA) has been leading the space exploration since before humans even landed on the moon. The first probes and satellites are designed to be unmanned. Nowadays, NASA is looking to utilize robots to their full potential, because there are many habitats in which humans will not be able to explore without costly equipment to protect them from the planets hazardous ecosystem [1-3]. This goal can be achieved by using small and inexpensive rovers to explore new planets surfaces while maintaining low costs. Not only costs would be kept low, but safeguarding our astronauts from potential dangers until planets have been analyzed enough to ensure their safety. Robots help prevent putting humans in hazardous situations and they will reduce the cost related to transporting humans to space. For instance, humans would require a constant supply of water, food, and oxygen. All this would prove to be far too costly to be feasible whereas robots can be crammed anywhere on a spaceship without the need for food or water just energy. This energy could be harnessed from solar panels set up on the planet by the robots themselves.

Not just any robot can be designed and sent into space, for example, if an expensive rover is sent into outer space and was damaged then it would be almost impossible or too costly to send either another robot or a human to do the required repair. With this in mind, NASA decided that the best course of action would be to send inexpensive rovers, which are mostly 3D printed and running on budget-friendly components [2-6]. To make this more effective, NASA is analyzing the possibility of having multiple small robots "swarmies" that are programmed with an ant-like behavior. These rovers traverse exploring multiple areas, collect resources, and communicate with each other, which will allow searching a wider area at a time, a swarmie robot is shown in Fig. 1.
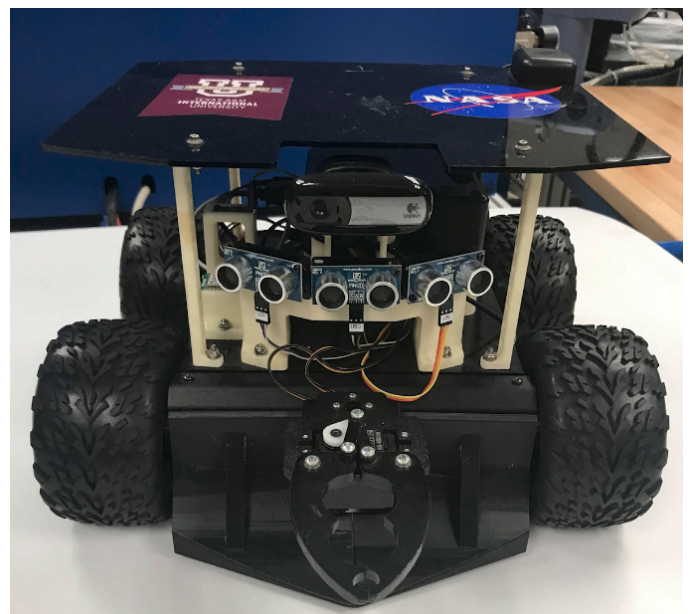
Fig. 1. DustySWARM Swarmie Robot

The University of New Mexico (UNM) were the founders of Swarmathon competition, Dr. Melanie Moses and Joshua

Hecker constructed the SwarmBaseCode-ROS to have a foundation to start developing new search algorithms [2-10]. After founding the competition with the support of Theresa Martinez, NASA manager of Minority University Research and Education Program STEM Management, it was turned into a national level competition. The competition has historically been composed of three parts: Physical Competition, Virtual Competition, and High School Competition.

For the 2019 competition, NASA has decided to have a Virtual competition only. The focus was to develop and simulate a code that will be used on multiple rovers at the same time. One of the most significant changes to this year's challenge was the use of six rovers instead of three. Each rover has around 16 main controller components that control the swarmie's behavior. The most fundamental controllers are the Search Controller, the Pickup Controller, the Drop off Controller, and the Obstacle Avoidance. All coding was conducted using Ubuntu-based C programming and GitHub was utilized to share codes and monitor updates and modifications between the swarmies, team members, and NASA, which was a very valuable tool to implement all Software Engineering principles.

Throughout the literature review, the team was able to analyze the code provided by UNM, Durham Technical College team, Florida International University, and the Japanese Rovers on the asteroid. The team planned to modify the search controller code as it is a major component in a successful search algorithm. The change from the previous DustySWARM team's Square Spiral Search (SSS) into an L-shaped search will allow the robots to move from the square field center to its outside border. Swarmie team consist of six rovers, where a rover will be located at the middle of each side of the squared home base and two extra rovers that each will be placed randomly at the home base corner. This means two extra rovers would be covering same areas, as all rovers must run the same code. Therefore, to solve this overlapping of resources, DustySWARM team decided to set up these two extra rovers with an alternate search route. Essentially, these two extra rovers will go out beyond the scope of the side rovers (using the L-Shape) and cover two quadrants each.

The paper is organized as follows: Section 2 is a background and literature review, Section 3 describes the problem statement and challenge provided by NASA, Sections 4 deals the methodology, while Section 5 illustrates the proposed solutions and implementations, Section 6 summarises the results of both simulations physical runs, and Section 7 concludes the paper and describe the team future plan.

II. LITERATURE REVIEW

The purpose of the project is to create inexpensive improved rover systems that can perform multiple functions, such as image capturing, rock mining, and data collection. There have already been many instances where rovers, as such, are utilized [2-5].

One concurrent example is the Japan Aerospace Exploration Agency (JAXA) small rover that was sent to analyze the surface of an asteroid that is about 280 million kilometers away from Earth. This rover has a designed movement that allow a jumping action, which allowed capturing images of the asteroid and transmit them back to researchers on Earth

[11-12]. This is one of many examples in the breakthrough of space exploration using inexpensive space rovers, which lays the foundation in the lore of how the project began its life.

To achieve the project goal, the team was engaged in the engineering process formulating using past resources. Every year, each new DustySWARM team may have the decision to either revise and improve previous search patterns or create a completely new algorithm that will be implemented from scratch. DustySWARM teams developed a couple of search algorithms for NASA Swarmathon challenge and competed against multiple schools across the nation. They dedicate time to obtain the most optimal path for rovers and alias swarmies to work autonomously and in synchrony. The swarmies must search, retrieve, and deliver AprilTags to the home base within a designated arena and given time limit.

DustySWARM 1.0 team developed a reverse twister search algorithm that depends on Archimedean Spiral using parametric equations, where rovers followed a counter-clockwise rotation spiral shown in Fig. 2 [13-14].
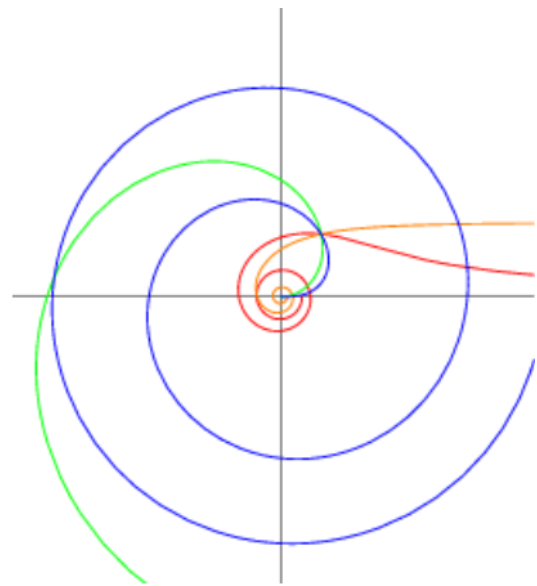
Fig. 2. Archimedean Spiral

DustySWARM 2.0 team analyzed multiple other possible search patterns before they selected their search pattern; Spiral Search was listed as one of their alternatives. Throughout multiple endeavors, DustySWARM team decided on an original path of a spiral path; however, this proved to be inefficient as rovers were not able to explore corners as the team wished. As a result, the team design evolved towards a Spiral Epicycloidal Wave (SEW) path [15-16]. Fig. 3 shows a visual representation of this path. In non-technical terms, the path is essentially a continuous spiral, what would allow maximum coverage. However, there are limitations with circular search patterns, such as the competition field has squared corners. However, the team was well aware of the corner situation and it was decided as a constraint/trade-off to the selected path. They proceeded with the path since it would be easier to implement within three to six rovers with minimum complications.

The DustySWARM 2.0 team divided the effort into three
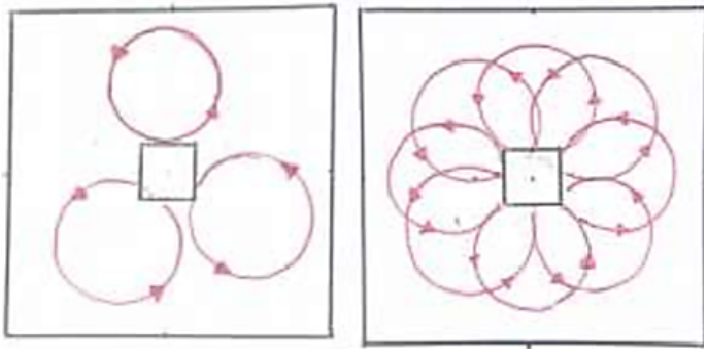
Fig. 3. DustySWARM Spiral Epicycloidal Search Pattern

Fig. 4. The Snake Path a.k.a Square Spiral Pattern

Fig. 5. DustySWARM 3.0 Square Searching Algorithm Path

main concepts, namely, rover task division, square spiral pattern, and spiral motion. The purpose of the rover task division was to dedicate two rovers searching the middle-section of the arena and one of the rovers searching the boundaries. However, this design was discarded since some regulation had been enforced that rovers cannot be individually programmed; meaning that there is only one search algorithm for all rovers. The second concept was the square spiral pattern, where the rovers will scan the area to perform a square spiral path. However, whenever a rover encountered a resource, the rover would return to the home original position and forget it last AprilTags pick-up position it was in. The last design was spiral motion following an Archimedean Spiral through the implementation of parametric equations.

DustySWARM utilized two separate processes in the experimental observations period: a sinusoidal modulation and pulsatory modulation. The sinusoidal modulation consists of exposing the reactor with a rotation period of 30.31 seconds and a wavelength of 1.71 millimeters. The obtained results indicated the formation of hypocycloids and epicycloids. For the pulsatory modulation, the reactor was exposed to the same light luminosity, but with short light pulses, which resulted in a different formation, it was a pronounced linear drift. The shape of the spiral waves resembled a similar result as the sinusoidal modulation: a hypocycloidal and an epicycloid trajectory. DustySWARM 2.0 built the design based on the Spiral Epicycloidal Wave (SEW), and the equations provided the necessary shape after converting them into C++ language [4-6].

DustySWARM 3.0 improved the SEW and developed the Square-Spiral Search (SSS) path shown in Fig. 4. There are numerous advantages of using a square-spiral pattern, it will allow rovers to survey the entirety of the arena including corners and its simplicity will allow future adjustments to maximize the pattern. Additionally, the rovers will be allocated to survey a dedicated area rather than searching collectively, which will increase the coverage percentage as shown in Fig. 5 [17-18].

Florida International University (FIU) Swarmathon team enhanced the rovers' performances through code optimization, by utilizing the sensor data, and compiling it to a single location. This methodology developed a predetermined search pattern in which the rovers would follow. For example, rovers would move a one-meter increment in the X-direction, and then another one-meter in the Y-direction [19-20]. Then the rover's coordinates were stored and a 360-degree radial turn will be initiated to scan for any AprilTags. If none were found, the rover will continue its stair-step motion, but if AprilTags were detected within the rover's proximity, the rover would proceed to retrieve and return the tags to the home base in one motion as shown in Fig. 6. The rover returns to home-based by recording the summation of all X and Y displacements that were traveled [7-8].
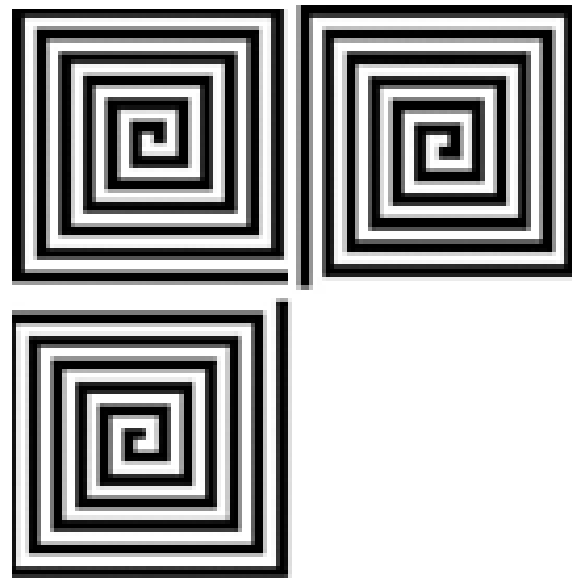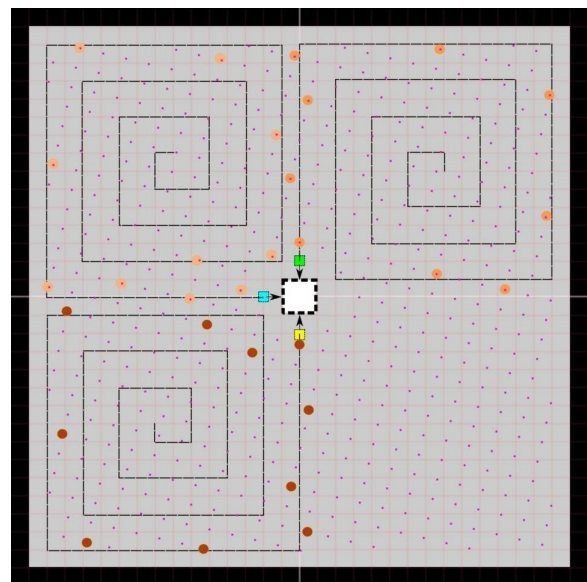
Additionally, FIU team increased the turning increment for the Object Detection to 1.57 radians (90 degrees) from the original value of 0.2 radians (11.45 degree) which was their solution to avoid the frequent rover collisions.

The FIU team simplified the rover collected sensor data,
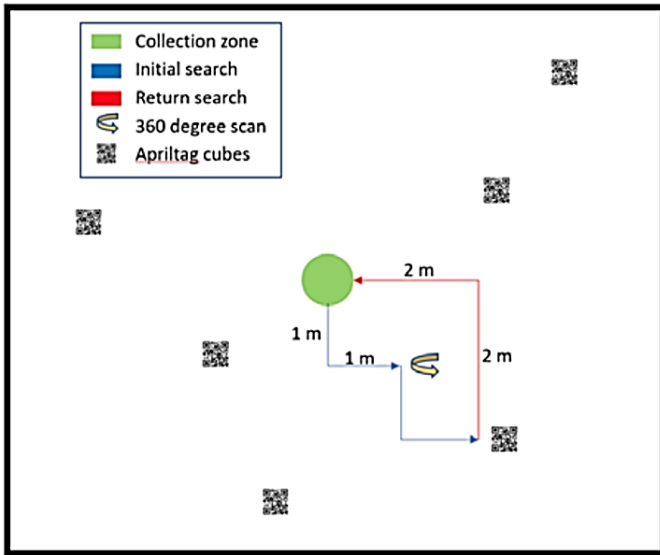
Fig. 6. FIU Rovers Path

which helped in implementing a viable accurate code that will find AprilTags. The definitions of sensor data as stored and generated is shown in Table I. Since the Swarmathon competition prohibits the extraction of the rover's IP addresses, the FIU team assigned random numbers to each rover that was saved in a local variable (R), where the rovers would compare the individual saved variables to each other. The difference in data would determine their rankings. The largest random number created by the rover would be assigned rank one, and subsequently, the rest would follow a numerical order. This approached was discussed in other research and illustrate its success [21-25].

TABLE I. SENSOR DATA

Symbol	Variable	Variable Name	Variable Location
R	Robot#		
T	Time		
CC	Cube Count	Detections[i].pose	Mobility.cpp
X	x-position	currentLocation.x	Mobility.cpp
Y	y-position	currentLocation.y	Mobility.cpp
θ	angular-position	currentLocation.theta	Mobility.cpp
S1	Left Sensor	sonarLeft-rang	Obstacle.cpp
S2	Center Sensor	sonarCenter-rang	Obstacle.cpp
S3	Right Sensor	sonarRight-rang	Obstacle.cpp

Park, et al., reworked the Coverage algorithms to account for time constraints, which is a significant factor as the Swarmathon competition is timed, so rovers must find, collect, and deliver AprilTags quickly and efficiently [26]. The Coverage algorithm was intended and best suited for intelligent robots that are unaware of the surrounding environment, that must be covered by the rovers within a certain period. The most powerful coverage algorithms rely heavily on having a complete grid map of the environment. For this reason, the authors utilized Simultaneous Localization and Mapping (SLAM) algorithms to help their robots to operate efficiently in an unknown environment. In addition to being stationed within an unfamiliar location, the rover robots must be able to adapt to dynamic or static obstacles within the environment. Therefore, a new proposed algorithm by the name of DMax Coverage

was taken into consideration for rover use. However, our competition involves static obstacles and resources to retrieve, so if SLAM were used, adjustments would be necessary [27-30].

DMax algorithm works by first using the SLAM algorithm to find out the boundaries of the unknown environment workspace and the robot's position and orientation. Then, the area will be mapped where the DMax algorithm computes a minimum bounding rectangle (MBR) [27-30]. An MBR is a rectangle that includes all free areas and areas with obstacles. Then this rectangle is simplified into smaller rectangles with no obstacles found, this approach is a common mathematical algorithm and called Rectangle Tiling Scheme, shown in Fig. 7.
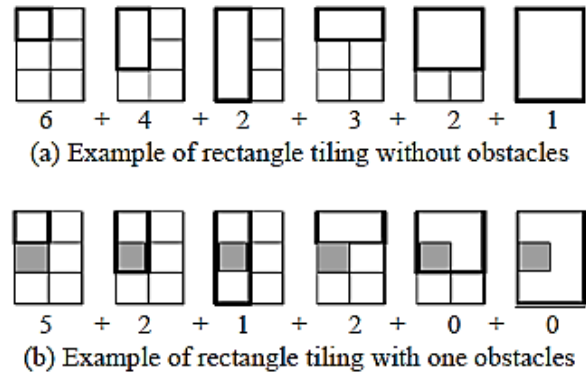
Fig. 7. Examples of Tiling Rectangle Filling

Rectangle Tiling Scheme aims to see how many times the rectangle in the bold outline can fit within the space without being rotated or overlapping an obstacle. After separating the rectangles into sub-spaces, the algorithm must then decide the sequence in which to visit them. The size of the rectangle estimates coverage time, number of turns the robot has to make, number of obstacles, and estimates time to reach the rectangle from its current location while taking into account the maximum coverage [27-30]. While the robot is moving from one point to another, it is continuously mapping the environment and looking for changes that would affect its next path.

Comparing the DMax algorithm to a Random and Boustrophedon algorithms in both simple and complex map environment, showed its superiority by 5% as revealed in Table II. The longer the given deadline, the more drastic the improvement will appear. Although the test conducted is not catered for object retrieval, this algorithm has some useful mapping applications.

TABLE II. DMAX COVERAGE AREA (%) WITH VARIOUS DEADLINES

Deadline (Sec.)	Covered Area (%) without Deadline	Covered Area (%) with Deadline
700	35.28	37.14
800	37.95	38.10
900	39.12	42.61
1000	44.12	48.312

III. PROBLEM STATEMENT

NASA Swarmathon 2019 challenge aimed to to develop cooperative robotics to revolutionize space exploration and further advance technology for future NASA space exploration missions. This competition started with new and unexpected rules and regulations. The first major change was making previous competitions' codes available to all teams, which allowed every team to compare and enhance their code. The second major change was six rovers per team would be used bigger field, which posed the need of creating a search algorithm that allows all six rovers to search equally and efficiently and change parameters to adjust for a larger field. The team goal is to design and test a new algorithm based on all available information that can be used to control 6 swarmies to collect and deliver the max number of simulated resources within the allotted time.

IV. METHODOLOGY

Through rigorous planning and analysis, DustySWARM 4.0 found that the L-shaped pattern might be more viable to use in covering the most area searching for AprilTags. The Gazebo simulator showed that the L-shaped search pattern could run better with few modifications. Although the simulation went well, there were complications during the physical trials. This comparison helped the team to identify the complication sources, namely, the Drop-off controller has the rover driving through the home base, thus knocking out some of the collected AprilTags. Another issue found was rover needs code adjustment to include error measurements to follow the desired search pattern.

The team compared their code with other teams' previous codes. An important observation was that Durham Tech code was based on several switch cases and multiple libraries, thus making their code shorter in lines, which can essentially help the rovers to "think" quicker as it executes code. DustySWARM team decided to build their code utilizing all such libraries' knowledge gained from previous codes.

V. PROPOSED SOLUTIONS AND IMPLEMENTATION

The team planned to develop codes and libraries for every rover motion such as the gripper, the movement, the communication, mapping, and so on. The team focused on five main processes to achieve their successful search algorithm;
A) Basic search and return algorithms,
B) Inter-connectivity and intermediate search algorithm,
C) Mapping,
D) Advanced algorithms using Mapping Techniques and
E) Drop-Off Controller.

All these processes are discussed in the following subsections. Fig. 8 shows the differences between DustySWARM developed SSS and the Radiant L-Shaped search Algorithms.

A. Basic Search and Return Algorithms

Based on the new base code, the team decided to optimize it by addressing two key issues: the search algorithm and the return algorithm. Improvements are suggested based on two criteria: expected distance traveled between the resources' pick
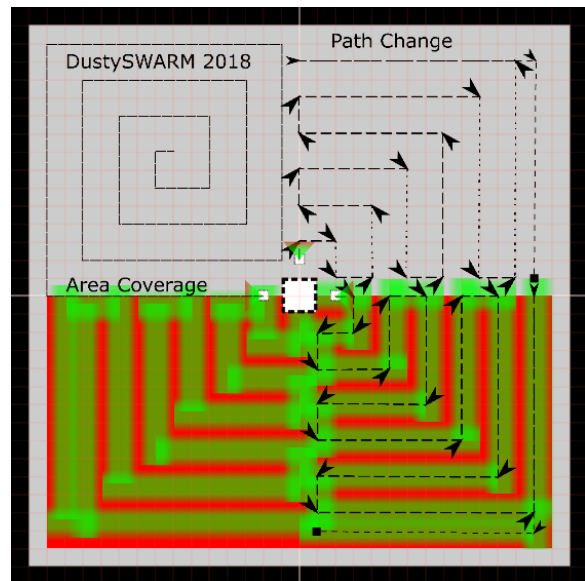
Fig. 8. DustySWARM SSS Vs. Radiant L-Shaped Search Algorithms

up location and the drop off location and the area in which the robot travels.

The square spiral pattern is an efficient method for searching a quadrant, but the implementation of the reverse pattern prioritizes the resources furthest away from the delivery location, which resulted in significant time and distance consumption compared to picking up the closest resources to the delivery location. Additionally, this algorithm allows the understanding of what areas had been covered and potential obstacles had been reduced. DustySWARM 2.0 return algorithm relies on Global Positioning System (GPS) to guide the rover to return to its initial starting spot for drop-off, which was a challenge as the rover small scale resulted in inaccurate GPS readings while determining the correct position, therefore the rovers will drift significantly as shown in Fig. 9.
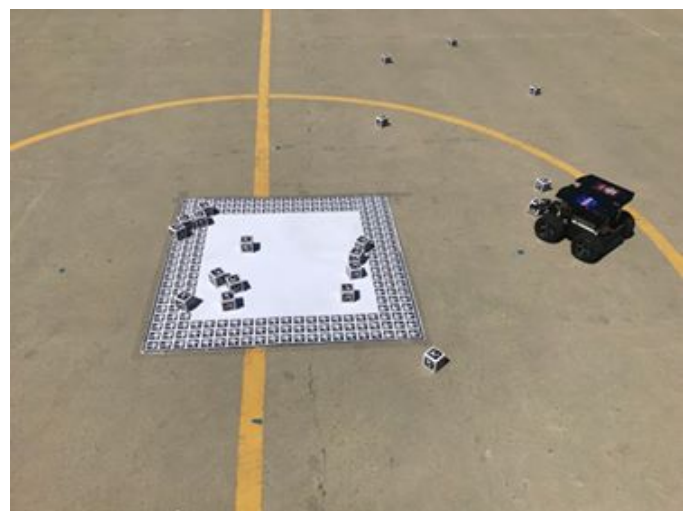
Fig. 9. Swarmie Rover with AprilTags and Home Base

Using the reverse square algorithm resulted in a direct line

of travel between the pickup and the delivery locations, which exposed a problem with resource picked up and ready for delivery but blocking the sensor and front camera. This is a major reduction in the rovers' ability to detect and navigate obstacles in unmapped terrains.

DustySWARM 4.0 search algorithm utilized the L-shaped pattern to search a quadrant in a radiating fashion, which resulted in collecting the closest resources and minimizing the time needed for this task as illustrated in Fig. 10.
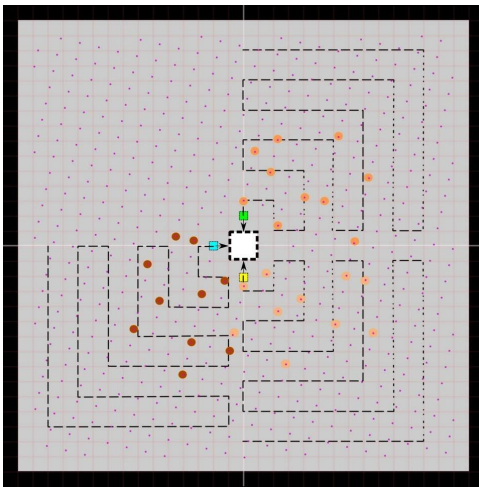
Fig. 10. Three Rovers following DustySWARM Radiating L-Shaped Search Algorithm Path

This method made it easier to potentially map out the covered area and provide a minimal obstacle for deliveries of further resources. The return algorithm had been updated to rely more heavily on the odometery instead of the GPS, as the robot encoders have a more accurate calibration with smaller drift values. In this algorithm, the two corner (extra) rovers will be assigned to search an area outside the highest parameter of the side rovers. This assignment was done by developing a function called (IAmCorner), where the rover will turn around and detect its location based on detecting the surrounding rovers and home base QR codes. Therefore, the side rovers will cover the area between the two corners (0, 0) and (5, 5), while the corner rovers will search the area beyond (6, 0) and explore from there out following a mirrored version of the L-Shape pattern as shown in Fig. 11.

To conduct an accurate test, search pattern for corner rovers was hard coded with IAmCorner value set to true. This IAmCorner value is equal to the return of a function, which at the start of the competition will see how many rovers it has to its left and right. If the corner rovers are placed at opposite corners then the side rovers will only see a maximum of one (1) rover next to them. The corner rovers, on the other hand, will always have two (2) rovers next to them. This function simply returns either true or false and assigns it to IAmCorner, which is shown in Fig. 12. If IAmCorner is false then the rover will run the L-Shape pattern cases, if it is true then the rover will use the extended mirrored L-Shape pattern. This method has proven to work roughly around 80% of the time. There are times when the northeast (north being up on the map) corner rover follows the same pattern as the east side rover.
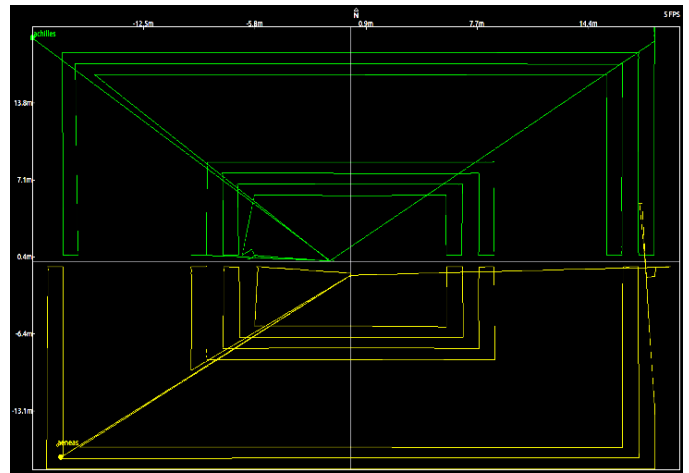
Fig. 11. Corner Rovers Searching Zones

Fig. 12. IAmCorner Function Process to Assign the L-Shape Path

B. Inter-connectivity and Intermediate Search Algorithm

Further improvement was done by implementing Robot Operating System (ROS) to develop actual communication channels between all rovers [30-33]. These communication channels were used to send the position, sensor data, and visible tag information of each rover. The rules for the NASA Swarmathon allowed understanding the initial positioning and orientation of the rovers based on the data provided by the rover's Inertial Measurement Unit (IMU). However, these data did not take into account the orientation of the field. Correction of the field orientation was done by analyzing the delivery location AprilTags and drawing a straight line or corner to establish the delivery zone by comparing the data provided by other rovers.

The search algorithm depended on the ROS communica-

tions information to direct the rovers to the suitable behavior. For example, in a three-rover system, if every rover had been searching a quadrant and one of the rovers finished the area, then that rover will be direct to move to the fourth quadrant to prevent any duplication of the search effort. An active tally of the collected resources allowed the team to change the behavior of the search algorithm by increasing the size of the steps to cover more area and find the resources that exist away from the delivery zone.

C. Mapping

The mapping controller was developed to include two different arrays that are updated continuously. These two arrays represent the Drop-off controller and the Search controller, they are composed of 1's, 5's, and 9's and each number represents a different field representation and condition. Number (1) represents the status of a rover being next to the home base when it is going for a drop-off. Number (5) represents the areas where it has to go; and there are AprilTags and the rovers should head over in that direction, while Number (9) means an area, where the rover cannot go because it represents the home base or obstacles.

Each array had an additional "if/else" and "for" loop statements, to let the rovers know what functions to perform autonomously. In essence, the numbers served as the size of the search nodes, as well as how far the search radius is needed for the rovers to head in those directions to collect AprilTags. The arrays are part of an algorithm that allows the rovers to compute the lowest possible cost to search for and collect AprilTags. Mapping using one rover with few obstacles and AprilTags is illustrated in Fig. 13. Fig. 14 and 15 show the mapping output of one/four rovers exploring the complete field.
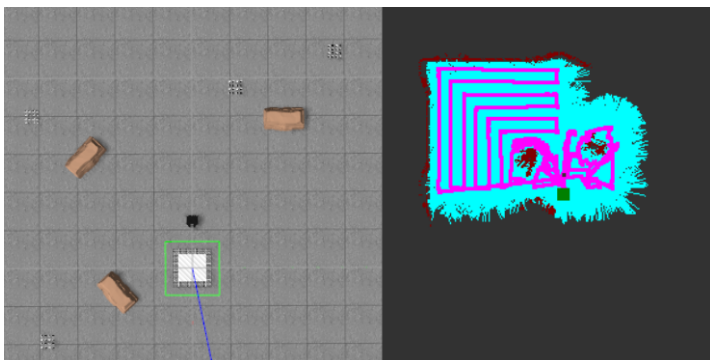
Fig. 13. Single Rover Mapping Simple Field with Obstacles and AprilTags

D. Integrating Mapping Techniques within Advanced Search Algorithms

In this stage, the automated processes of rovers driving to possible resource zones were developed. These zones will be established by using ultrasound detection, mapping techniques to prioritizing areas with probable resources over those that lack resources. All rovers will be programmed to communicate and share the location of any resources or obstacles they discover and any probable resource locations. Based on the percentage of collected resources and/or area covered, this
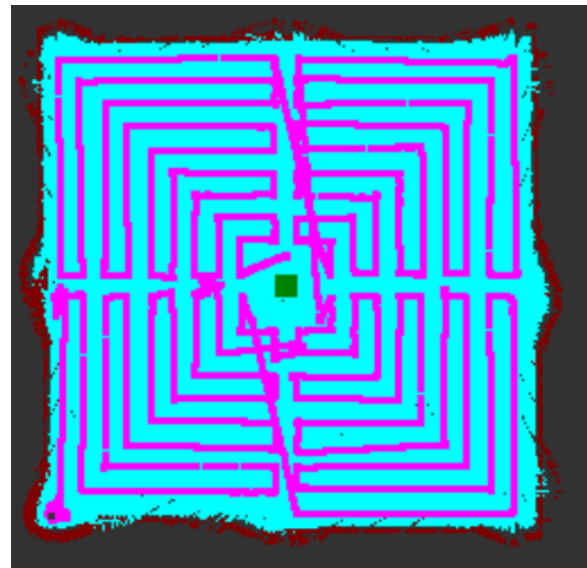
Fig. 14. One Rover Mapping a Complete Field
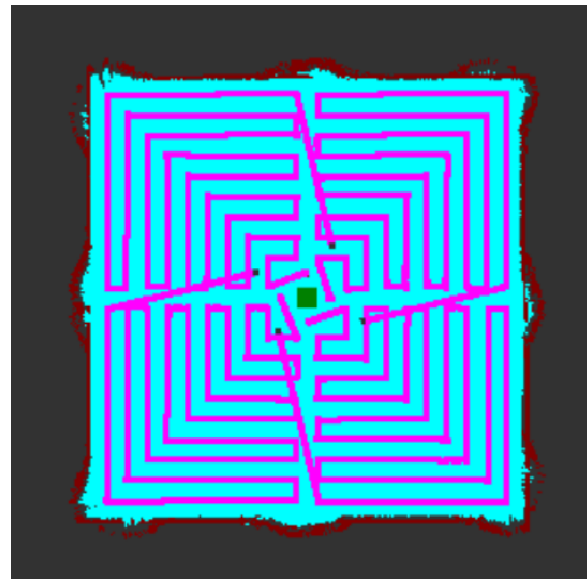
Fig. 15. Four Rovers Mapping a Complete Field

mapping will go into full effect and override the base search sequence.

Depending on trials, the execution conditions of map integration might change to default in the cases of ROS Communication failure, searching in the immediate vicinity of the delivery zone, and valid detection from the ultrasound sensor. The automated process will map out the destination location by prioritizing areas, which have been previously transverse and potentially avoid obstacles. This system can be improved by tracking the time required to move to the location in question, which will allow the prediction of rover motion and paths crossing. This will reduce the collisions chance between rovers. An additional technique that can be implemented before the search period ends is to establish a routine for the rovers with no assigned task to move around

the delivery zone and push back resources that have fallen out.

E. Drop-off Controllers

One of the challenges was the inconsistency of the Drop-off controller performance. In a few instances, once the rover completed the search pattern, and proceeding with the retrieval algorithm, it failed in dropping the AprilTags successfully, where AprilTags will be dropped outside the home base. This situation occurred when the rover would enter the home base at an angle, which would initiate the drop off algorithm by searching left and right for home base QR codes and following only the edge of the base and then proceeded to leave the AprilTag outside of the desired location as shown in Fig. 16 and 17.
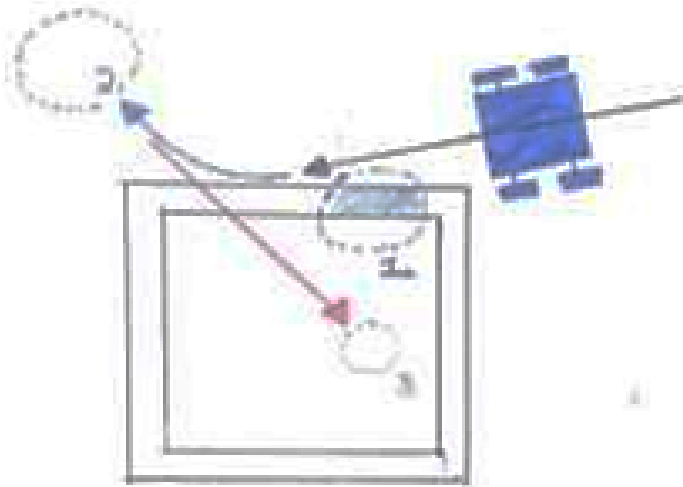
Fig. 16. Illustration of Rover Dropping off the AprilTag Outside Home Base
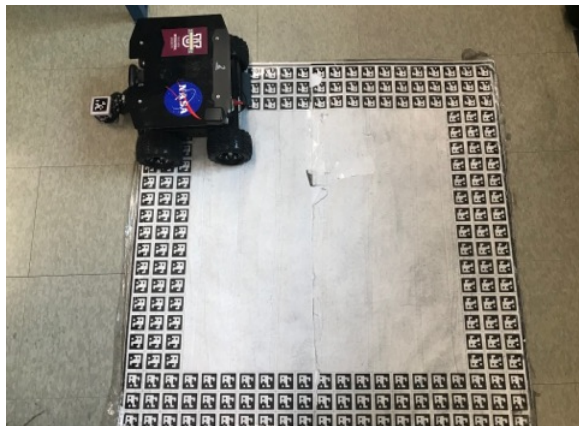
Fig. 17. Rover Dropping off the AprilTag Outside Home Base

The first proposed solution was to assign a quadrant of the home base to every rover, by assigning the quadrant x and y coordinates to each rover. The rover would proceed to go to the nearest cardinal axis of the grid and enter the home base perpendicularly without any angular miscalculation. This idea evolved into another approach that uses waypoints, Fig. 18 shows a segment of the improved drop-off code. The waypoints concept is a type of direction algorithm. It stores

the desired location and current location, which will be used later in matching these two locations coordinates. The code would calculate the nearest cardinal axis relative to the rover current location and then proceed to the axis intercept to enter the base perpendicularly for a drop-off.

```
Point DropOffController::Closest_Drop_off(Point currentLocation)
{
    int XMulti, YMulti;
    double drop_zones[4]={0,.30,.30,0},Min_H_Drop=255,Com_Drop;
    Point Min_Drop;
    switch(inputs.InternVector.Current_Quadrant)
    {
    case 1:
        XMulti=1;
        YMulti=1;
        break;
    case 2:
        XMulti=-1;
        YMulti=1;
        break;
    case 3:
        XMulti=-1;
        YMulti=-1;
        break;
    case 4:
        XMulti=1;
        YMulti=-1;
        break;
    }
    for (int i = 0; i <4; i+=2) {
        Com_Drop=hypot(drop_zones[i]*XMulti-(currentLocation.x+inputs.InternVector.offsetX),d
        cout<<"Drop off Com_Drop is "<<Com_Drop<<" and set ("<<i<<","<<i+1<<")"<<endl;
        if (Com_Drop<Min_H_Drop) {
            Min_H_Drop=Com_Drop;
            cout<<"Drop off Min_H_Drop is "<<Min_H_Drop<<" and set ("<<drop_zones[i]<<","
            Min_Drop.x=drop_zones[i]*XMulti+inputs.InternVector.offsetX;
            Min_Drop.y=drop_zones[i+1]*YMulti+inputs.InternVector.offsetY;
        }
    }
```

Fig. 18. Waypoints Drop-off Code

VI. RESULTS SUMMARY

After several virtual simulation and physical trials, the team was able to develop a new search algorithm that can be implemented in NASA Swarmathon rovers, by programming the swarmies to search a predefined quadrant in an L-shaped radiating fashion. Additionally, These robots had been coded to determined their proximate location based on the number of surrounding robots through the IAmCorner function.

This Algorithm allowed the robots to collect the closest AprilTags/resources while minimizing the search time needed and without interfering or interrupting each other. Utilizing ROS communication, the rovers were able to share the collected AprilTags count that was used for changing the search steps to cover more area and find the resources that exist away from the delivery zone and if rovers finish searching the assigned quadrant, they will be able to select the next search area. Some of the results are shown in Table III.

TABLE III. L-SHAPED RADIATING SEARCH ALGORITHM RESULTS

Target Distribution	Simulation Runs Results	Physical Trials Results
Clustered	32	27
Power Law	118	110
Uniform	72	63

Moreover, the team developed and integrated a mapping technique, that used the ultrasound sensors, GPS and IMU, where rovers will create their own map as an array by identifying and prioritizing the home base, the obstacles and the safe search areas. This map will be used as a fail-safe measurement in case of ROS or WiFi connection failure. Additionally, the

drop-off controller had been improved by including the start location of the rovers to determine the best home base entering point. Fig. 19 shows the simulation run with four rovers to search the predefined field.
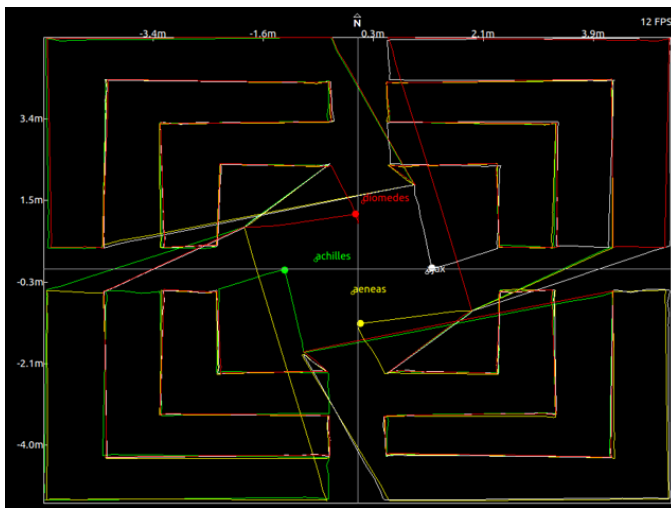
Fig. 19. L-Shaped Radiating Search Gazebo Simulation Results

VII. Conclusion

In conclusion, the team developed a searching algorithm that was implemented to control NASA swarmies to search unknown terrain for simulated resources and retrieve them back to the home base. The current code had multiple improvements and both Gazebo simulation and real-life runs showed an improved consistency of the rovers' performance.

This opportunity was valuable for the team as they practiced the systems engineering and management principals in a real-life project and expanded their computer science techniques to contribute to NASA's mission of exploring space.

Currently NASA Swarmathon had been paused but with any code, there is room for improvement. The team will continue to further improve the algorithm code especially the rovers' behavior when picking-up and dropping-off AprilTags and enhance the rovers' ability to communicate amongst each other. Other areas of future improvements are to programming the rovers with the ability to stack and push AprilTags when dropping-off when swarmies are close to home base.

Acknowledgment

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] Secor, P. (2016). "NASA Swarmathon".

[3] Glenn, T., Ragland, S., Meyer, A., Pulliam, J., Holmes, E., and Riley, N. NASA Swarmathon.

[4] Hecker, J. (2015). "Swarmie User Manual. Quick Start Guide for Physical Robots". University of New Mexico. www.Github.com.

[5] Montague, G. (2014). "Swarmie User Manual: A Rover Used for Multi-agent Swarm Research". www.ntrs.nasa.gov

[6] NASA Swarmathon Home Page. (2015). Retrieved from http://nasaswarmathon.com/

[7] Ackerman, Sarah M., G. Matthew Fricke, Joshua P. Hecker, Kastro M. Hamed, Samantha R. Fowler, Antonio D. Griego, Jarett C. Jones, J. Jake Nichol, Kurt W. Leucht, and Melanie E. Moses. (2018). "The Swarmathon: An autonomous swarm robotics competition". arXiv preprint arXiv:1805.08320.

[8] BCLab-UNM. (2015). BCLab-UNM/SwarmBaseCode-ROS.

[9] Bhattacharya, S., and Agrawal, R. (2017, March). "Development of robot swarm algorithms on an extensible framework". In SoutheastCon 2017 (pp. 1-6). IEEE.

[10] 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

[11] Kubota, Takashi, and Tetsuo Yoshimitsu. "Intelligent rover with hopping mechanism for asteroid exploration." In 2013 6th International Conference on Recent Advances in Space Technologies (RAST), pp. 979-984. IEEE, 2013.

[12] Ulamec, Stephan, Patrick Michel, Matthias Grott, Ute Böttger, Heinz-Wilhelm Hübers, Naomi Murdoch, Pierre Vernazza et al. "A rover for the JAXA MMX Mission to Phobos." In 70th InternationalAstronautical Congress, pp. IAC-19. International Astronautical Federation, 2019.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] FIU Panther Swarm Team (2016). NASA SWARMATHON 2016: FIU Panther Swarm Team Technical Report.

[20] Jagolinzer, S., Larrarte, J., Guerrero, R., and Tosunoglu, S. (2016, May). Development of Swarm Algorithms for Space Exploration. In Proceedings of the 29th Florida Conference on Recent Advances in Robotics, FCRAR 2016.

[21] Richard, W. K., and Majercik, S. M. (2012, July). Swarm-based path creation in dynamic environments for search and rescue. In Proceedings of the 14th annual conference companion on Genetic and evolutionary computation (pp. 1401-1402).

[22] Aguilar, J., Blanchard, A., Sibiski, A., Soto, J., Jagolinzer, S., and Tosunoglu, S.(2017) Performance Optimization of Swarm Algorithm and Sensor Data for NASA Swarmathon Competition.

[23] Miller, P. (2007). The genius of swarms. National Geographic, 212(1), 126-147.

[24] Lim, S. S., and Bouffanais, R. (2019). From Senseless Swarms to Smart Mobs: Tuning Networks for Prosocial Behaviour. arXiv preprint arXiv:1910.01303.

[25] Park, J. K., Jeon, H. S., Noh, S. H., Park, J. H., and Oh, R., (2010, October) "A practical coverage algorithm for intelligent robots with deadline situations" In ICCAS 2010 (pp. 299-303) IEEE.

[26] Kantrasiri, S., Jirakanjana, P., and Kheowan, O. U., (2005) "Dynamics of rigidly rotating spirals under periodic modulation of excitability" Chemical physics letters, 416(4-6), 364-369.

[27] Ramalingam, B., Veerajagadheswar, P., Ilyas, M., Elara, M. R., and Manimuthu, A. (2018). Vision-Based Dirt Detection and Adaptive Tiling Scheme for Selective Area Coverage. Journal of Sensors, 2018.

[28] Monaci, M., and dos Santos, A. G. (2018). Minimum tiling of a rectangle by squares. Annals of Operations Research, 271(2), 831-851.

[29] Idri, A., Oukarfi, M., Boulmakoul, A., and Zeitouni, K. (2017). Design and Implementation Issues of a Time-dependent Shortest Path Algorithm for Multimodal Transportation Network. In TD-LSG@ PKDD/ECML (pp. 32-43).

[30] Martinez, A., and Fernández, E., (2013) "Learning ROS for robotics programming" Packt Publishing Ltd.

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

[32] J O'Kane, J. M. (2014) "A gentle introduction to ROS". www.cs.rpi.edu

[33] Quigley, M., Gerkey, B., and Smart, W. D. (2015). Programming Robots with ROS: a practical introduction to the Robot Operating System. " O'Reilly Media, Inc.".

Impacts of Decomposition Techniques on Performance and Latency of Microservices

Chaitanya K. Rudrabhatla

Executive Director-Solutions Architect
Media and Entertainment Domain, Los Angeles, USA

Abstract—**Micro service architecture (MSA) has undoubtedly become the most popular modern-day architecture, often used in conjunction with the rapidly advancing public cloud platforms to reap the best benefits of salability, elasticity and agility. Though MSA is highly advantageous and comes with a huge set of benefits, it has its own set of challenges. To achieve the separation of concerns and optimal performance, defining the boundaries for the services clearly and their underlying persistent stores is quintessential. But logically segregating the services is a major challenge faced while designing the MSA. Some of the guiding principles like Single responsibility principle (SRP) and common closure principle (CCP) are put in place to drive the design and separation of microservices. With the use of these techniques the service layer can be designed either by (i) Building the services related to a business subdomain and packaging them as a microservice; (ii) or Defining the entity relationship model and then building the services based on the business capabilities which are grouped together as a microservice; (iii) or understanding the big picture of the application scope and combining both the strategies to achieve the best of both worlds. This paper explains these decomposition approaches in detail by comparing them with the real-world use cases and explains which pattern is suitable under which circumstances and at the same time examines the impacts of these approaches on the performance and latency using a research project.**

Keywords—*Microservices; decomposition techniques; single responsibility principle; common closure principle; performance; latency*

I. INTRODUCTION

In the recent past all the applications were built using a monolithic design pattern. This design pattern was a great advancement from the traditional client-server architecture which was prevalent before [1]. Monolithic design pattern was well suited and worked fine in the waterfall software development methodology. But the cutthroat competition and everchanging nature of the businesses demanded a software model which is more agile and nimble enough to cope up with the business needs. This thought process gave birth to the agile methodology. In the agile methodology, the software needs to be developed quickly in smaller pieces and deployed continuously to production. Monolithic systems struggled hard to find their place in the agile methodology. It was soon discovered that monoliths are more complex in nature due to their fundamental design, where the entire application code into a single deployable unit. Due to this design, the code changes need to be well planned in advance and need to be thoroughly tested before deploying to production and thus

leading to longer build cycles. This was found to be anti-agile. These issues have prompted the search for the newer design patterns which involved breaking down the monolith to a more loosely coupled services. Service Oriented Architecture (SOA) is one such pattern which evolved on these lines [2]. In this pattern, backend services are isolated and distributed. But these services are handled by a layer called Enterprise Service Bus (ESB) which is an integration and guarding layer for the entire backend. Though ESB pattern has the advantages of conducting the health checks, performing the routing for the backend services, it was soon found to be a cumbersome layer and a bottle neck as the application services grow in size. To handle these shortfalls, Microservice Architecture (MSA) was introduced with the similar premise of isolation and separation of concerns as proposed by SOA [3], but with a lightweight design.

In MSA, the application services are designed to be heterogeneous, light weight, independent, isolated and highly distributed in nature [20]. The biggest advantage of micro services is that it supports the services to be built in any technology of choice and permits them to be deployed independently from each other. This greatly reduces the efforts and simplifies the development, testing, build and deployment cycles as the changes are limited to a single service rather than the entire monolith. With the proliferation of cloud technologies and advancements in containerization and their orchestration technologies like Kubernetes, Docker swarm etc., the microservice architecture became even more efficient. This has resulted in a continuous integration and continuous delivery (CI/CD) which is the most important feature of agile methodology [6]. But these benefits can only be harnessed if the backend services are carefully examined and decomposed in the most optimal way by considering the big picture of the entire application scope rather than in an ad-hoc way. If not, this design might prove counter-productive and lead to latency, complexity and inefficiency [4]. Rest of this paper is organized as follows – In section (II) various decomposition techniques and their benefits and shortfalls are explained by considering a real-world e-commerce scenario. In section (III) the research project which was conducted to examine the impacts of decomposition techniques on latency and performance [12] of the system is explained and the results are compared. In section (IV) summary, conclusions are provided and an overview of future research work is given.

II. DECOMPOSITION OF MICROSERVICES

For gaining the benefits of MSA, it is very important to strategize the decomposition of micro services [17]. As the

name suggests, the services must be designed to be fine grained and independent. There are two major guiding principles which can drive the decomposition of services. (i) Single Responsibility Principle (SRP) is a guiding principle which states that a service class should exist to serve a single major responsibility and that class should only have one reason to change. Due to the isolation proposed by this theory, it is highly beneficial to apply SRP to build the microservices that implement a small set of closely related business functions. (ii) The second guiding principle states that it is not only in the initial development but also that the design should consider the future changes in such a way that they impact a single service. That is because changes that affect multiple services require more planning, coordination and testing. This slows down development and deployment cycles and would present the same issues of a monolithic application design. This constitutes the fundamental essence of Common Closure Principle (CCP), which is a second guiding principle for service decomposition. As per this principle if there are multiple micro services which serve a business functionality, a change in the business rule should only impact one single service rather than all the microservices involved. Using these two guiding principles, the services can be decomposed into granular microservices [15][16]. The below section examines the ways to implement the decomposition by taking a real-world example of a module from an e-commerce system [19].

A. Decomposition based on Domain Driven Desgin

Domain Driven Design is a decomposition technique which is based on the common closure principle (CCP). As per this principle, all the functions and classes which get impacted by a single business rule change should be packaged together as a single microservice. Some of the key considerations for this implementation are:

1) Services must be designed to be cohesive in nature. Which means that the methods, functions and services included in a microservice should strongly relate to activities related to a granular business functionality.

2) Each service should be designed to be completely autonomous. Which means if the business rule changes, it should be possible to apply a patch and deploy the service independently without impacting anything else.

To explain these designs, let us consider an e-commerce application module [5]. This module has three primary business entities: (i) User entity; (ii) Order entity; and (iii) Item Details entity. Here are some of the important business functionalities around each of these entities:

1) User Entity Responsibilities

- It stores the user details.
- It maintains the user profile settings.
- It has the user address details.
- It stores the payment details like credit card information.

2) Order Entity Responsibilities

- It maintains the Order history.
- It stores the item details of the Order.
- It stores the Order shipment details.
- It stores the payment details for each order.

3) Item Entity Responsibilities

- It contains the Item sku details.
- It maintains the item inventory details.
- It maintains the like score of the other items to give recommendations.

As it can be visualized from the Fig. 1, the business entities in the real world are not independent of each other. They need to overlap with one another for various use cases. For example – User entity relates to Order entity when there is a screen in the UI interface which displays the order history of the user. There will be a foreign key relationship between the User to Order entities. Similarly, an Order entity maintains a relationship with Item entity, such that when the item in a particular order is clicked, it takes to the details page of that particular item. In the same lines, User entity maintains a relationship with Item entity as the application might show the recommended items for purchase for that particular user based on the items purchased by the user in earlier transactions.

Now applying the Decomposition technique based on the Domain Driven Design [18], it can be visualized that the application module might be designed to have three microservices which can deal with the business functionality of each domain as shown in Fig. 2.

Using this principle there is one microservice per each business domain. Applying the Database per Service Pattern [7], there would be one database per microservice to achieve the isolation needed. This model is advantageous for the following reasons: (i) Each business functionality is encapsulated in its own micro service. Thus, it complies to Single Responsibility Principle (SRP). (ii) Changes related to a business functionality are mostly limited to its own micro service. Which means faster time to deploy the changes.
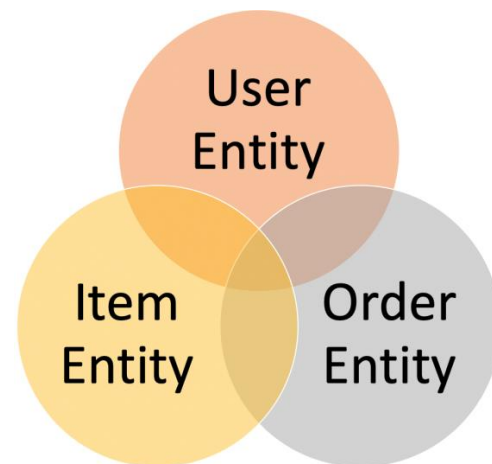
Fig. 1. Overlapping Business Entities in an E-commerce Application.
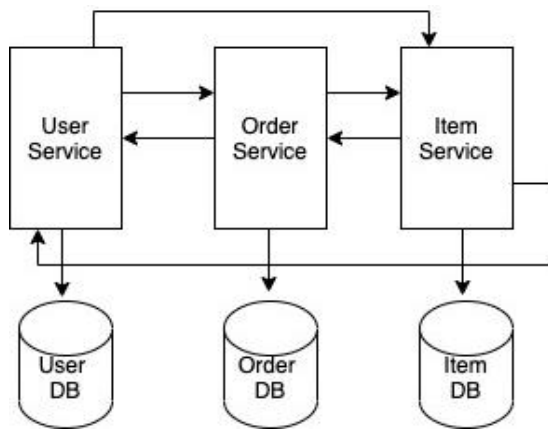
Fig. 2. Domain Driven Model for Microservice Decomposition.

However, it is not always advantageous as this design might lead to cross service events and communications which might complicate the maintenance as the application and business functionality grows. Here are some of the scenarios where the interservice dependency or communication is needed:

- Consider a scenario where the order is in progress and user updates the shipping address in the user profile. In this case, the address for the current order in progress may need to be updated as well upon the user confirmation. In this case, "update address" event on the User DB need to trigger the action for Order Service.

- Consider the other scenario where the order is in progress for shipment and there is a change in inventory because of which the shipment needs to be cancelled. In this case an "update inventory" event on the Item DB needs to trigger an action for the Order Service.

- In the same lines, when user changes the personal preferences, this event needs to trigger an action on the Item Service so that the recommended items are shown as per the new preferences.

There are multiple ways to handle this cross-service events and communication:

1) Inter-Service communication based on the Event Choreography technique [8], where each service can discover the other service using private load balancers and there by trigger the action in the destination service. This can be seen in Fig. 3.

Although this technique works for limited use cases, it has the following challenges:

a) It becomes difficult in the larger applications as the number of events [11] and triggers grow in number due to the intricacies of the business functionalities.

b) Also, the microservices no longer become independent as one team developing the Service1 may need to depend on the other team developing Service2 to call the actions in it.

c) The network traffic would also increase which might be another factor to consider.

d) It might also cause latency due to the inter service communication over network.

That gives rise to the second approach for communication.

2) Central orchestrator approach [9] where services are unaware of each other but post the events to a central orchestrator which takes care of firing the relevant actions in the destination services. This can be seen in Fig. 4.
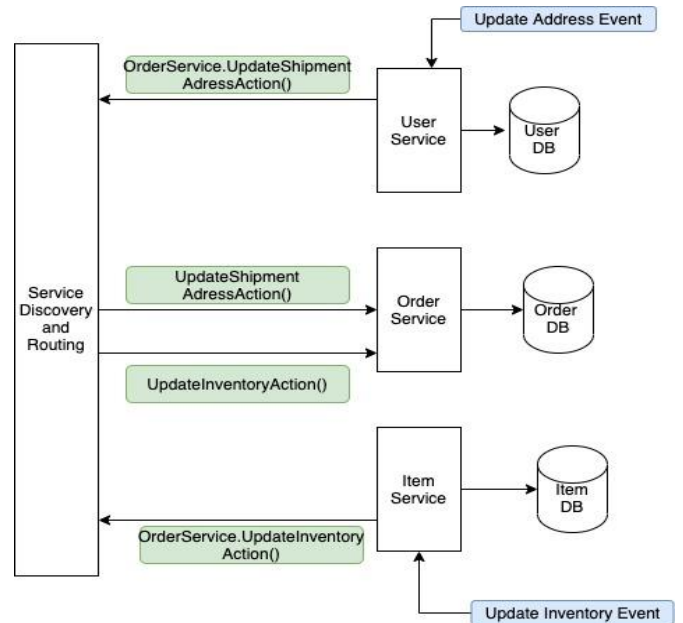
Fig. 3. Inter-Service Communication using Event Choreography and Service Discovery-Routing.
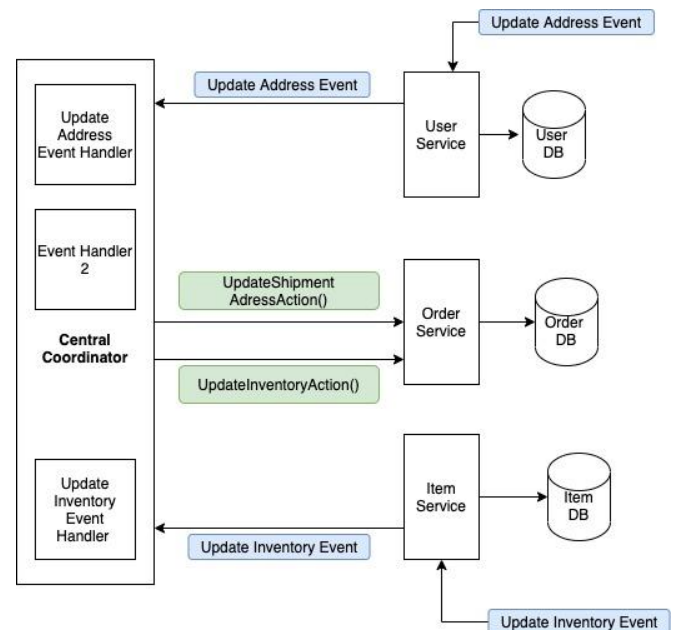
Fig. 4. Inter-Service Communication using Orchestrator or Central Coordinator.

This alternative works fine even in large and complicated applications. But it comes with its own set of challenges:

a) When the number of event handlers grow, the central coordinator becomes a fat layer. This becomes an anti-pattern for micro service architecture where each component needs to granular and light weight.

b) There is a risk that the central coordinator can act as a single point of failure. If this layer fails, entire application functionality might get impacted as it is responsible for multiple service actions.

c) Eventually this pattern might also introduce latency as the logical decisions to trigger the actions to the relevant services based on the incoming events might get tricky.

B. Decomposition based on Business capability

Decomposition based on Normalized Entity Relationship model is a design pattern where the major entities are normalized to the optimal level [14] and services are designed for the management of normalized entities involved in the business transaction. In this design, the services are designed around the activities of the entities. This can be visualized based on the Fig. 5. In this model the entities are broken down to the level where they can be considered standalone and the CRUD operations needed for managing the entities.

This approach is considered as an anti-pattern in the microservice architecture as it might pose a risk where a single transaction might have to flow through lot of management services before persisting finally. In case of a roll back this might get even more complex due to the large number of services which are granular and identifying the roll back steps might become challenging.

C. Decomposition based on the Hybrid approach

This is a third approach where the bigger picture of the application flows is analyzed and a hybrid approach which is a combination of both the approaches – domain based, and capability-based decompositions are used to come up with a hybrid model to gain the efficiencies. Fig. 6 shows the hybrid approach where a combination of subdomains like, user service, item service and order service are used and capability-based services like, address management service and inventory management service are used.
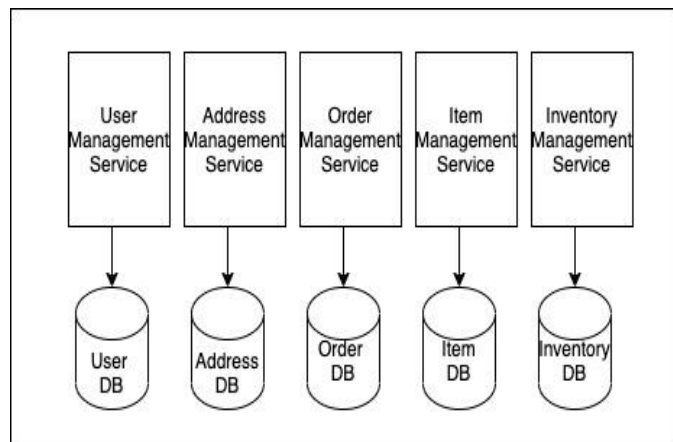
Fig. 5.   Decomposition of Services based on the Business Capabilities.
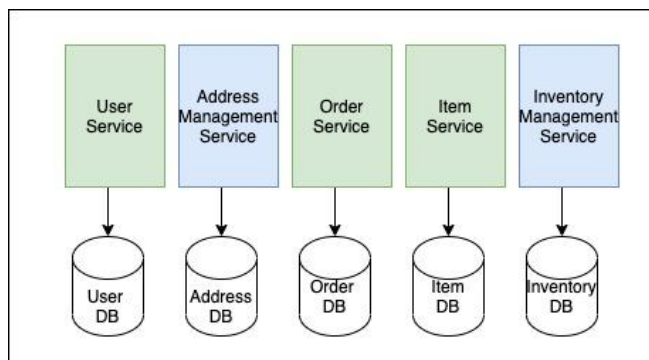
Fig. 6.   Decomposition of Services based on Hybrid Approach.

This approach is beneficial because it combines the benefits of both the approaches. It packages the subdomain in the single service for the most part. But in some cases which lead to actions across multiple entities and warranting interservice communication can be isolated as business capability-based services.

III. RELATED WORK

Related work is done to compare the performance [10] [13] and latency of decomposition techniques by building the micro services using each of the techniques mentioned in the earlier section and response times are observed when a database transaction is invoked. For this work Java based Spring boot microservices are used. MongoDB is used as the backend database.

A. Run 1 – Domain Driven Decomposition

For this simulation, four micro services MS1, MS2, MS3 and MS4 were used each of them having their individual MongoDB databases DB1, DB2, DB3 and DB4. DB1 with 2 dependent collections C11, C12 is used. Similarly, C21, C22 in DB2 C31, C32 in DB3 and C41, 42 in DB4. Eureka is used as the service discovery layer and Zuul as the routing mechanism for this experiment. First run is performed where a small message of size 50 bytes is persisted by calling one micro service MS1 and recorded the time in milli seconds. Later a test is performed where the transaction would persist in multiple collections C12 and C22. For this MS1 would make a call to a method hosted in MS2 to persist the data in C22 via discovery and routing layers. The tests are repeated couple more times where it would persist in 3, 4 and 5 collections in one transaction and time taken is recorded. Fig. 7 shows the performance of this technique.

B. Run 2 – Business capability Decomposition

In this run, the dependent collections were split as individual databases and built as six microservices MS1 to MS6. The reason for this is that in this model the services are built based on the business capability. This is to replicate the manager services for the normalized entities. In this approach a logic is added such that the transaction progresses sequentially in all steps. This is to replicate the behavior where the persistence would wait for the previous step to complete. Though this technique made the complexity of management services simple, the transactions were more time consuming as can be seen in Fig. 8.

C. Run 3 – Hybrid Decomposition

A hybrid approach is simulated where the same number of micro services as Run 1 are used, namely, MS1, MS2 MS3 and MS4. But the number of dependent collections are increased in MS1 and MS2 to 3 from 2 and reduced the collections in DB3 and DB4 to 1. This is to simulate a hybrid approach. The transaction which spans multiple collections is run and the timing is recorded as shown in Fig. 9.
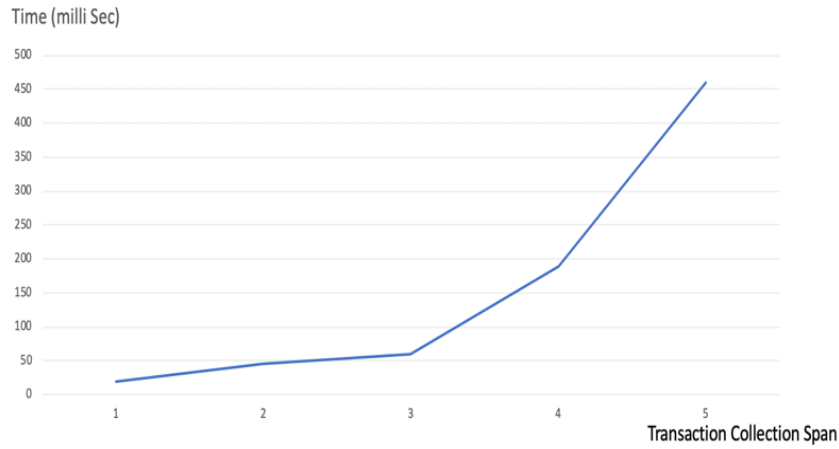
Fig. 7.   Performance of Domain Driven Approach.
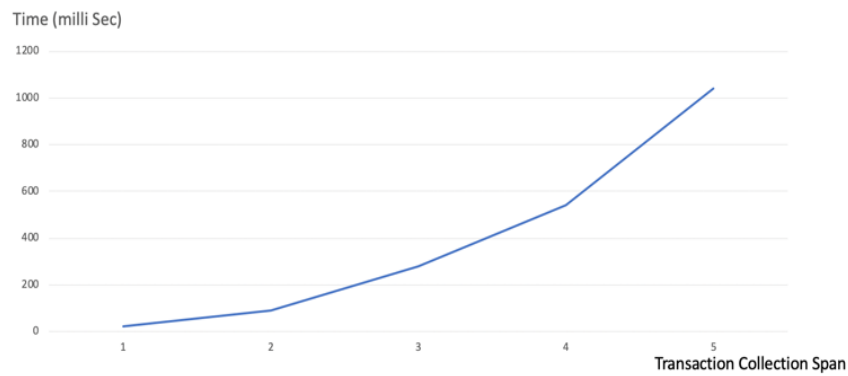
Fig. 8.   Performance of Entity Driven Approach.
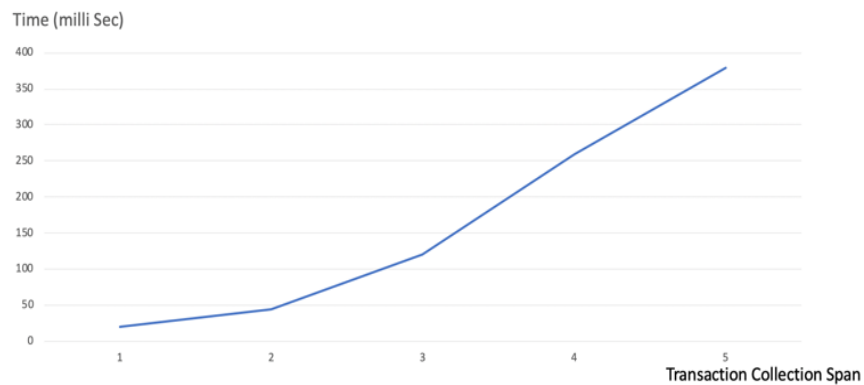
Fig. 9.   Performance of Hybrid Approach.

IV. Conclusion

Though there is no standard approach for decomposing the micro services, it is found that Domain driven decomposition technique to be more superior than the entity model driven business capability-based decomposition. However, in any real time application design, choosing a hybrid approach by considering the big picture of the business functionalities and transactions involved would yield better results in the performance. Future scope of work includes simulating more complex transactions in each of the models and analyzing the resource consumption, throughput and latency.

References

[1] Salah, Tasneem & Zemerly, Jamal & Yeun, Chan & Al-Qutayri, Mahmoud & Al-Hammadi, Yousof. (2016). The evolution of distributed systems towards microservices architecture. 318-325. 10.1109/ICITST.2016.7856721.

[2] K., Chaitanya. (2018). A Systematic Study of Micro Service Architecture Evolution and their Deployment Patterns. International Journal of Computer Applications. 182. 18-24. 10.5120/ijca2018918153.

[3] S. Newman, Building Microservices. " O'Reilly Media, Inc.", 2015.

[4] Zhu, Yuhao & Richins, Daniel & Halpern, Matthew & Reddi, Vijay. (2015). Microarchitectural implications of event-driven server-side web applications. 762-774. 10.1145/2830772.2830792. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] Asrowardi, Imam & Putra, S & Subyantoro, E. (2020). Designing microservice architectures for scalability and reliability in e-commerce. Journal of Physics: Conference Series. 1450. 012077. 10.1088/1742-6596/1450/1/012077.

[6] Balalaie, Armin & Heydarnoori, Abbas & Jamshidi, Pooyan. (2016). Microservices Architecture Enables DevOps: an Experience Report on Migration to a Cloud-Native Architecture. IEEE Software. 33. 1-1. 10.1109/MS.2016.64.

[7] Messina, Antonio & Rizzo, Riccardo & Storniolo, Pietro & Tripiciano, Mario & Urso, Alfonso. (2016). The Database-is-the-Service Pattern for Microservice Architectures. 9832. 223-233. 10.1007/978-3-319-43949-5_18.

[8] Rudrabhatla, Chaitanya. (2018). Comparison of Event Choreography and Orchestration Techniques in Microservice Architecture. International Journal of Advanced Computer Science and Applications. 9. 10.14569/IJACSA.2018.090804.

[9] Malyuga, Konstantin & Perl, Olga & Slapoguzov, Alexandr & Perl, Ivan. (2020). Fault Tolerant Central Saga Orchestrator in RESTful Architecture. 278-283. 10.23919/FRUCT48808.2020.9087389.

[10] Jayasinghe, Malith & Chathurangani, Jayathma & Kuruppu, Gayal & Tennage, Pasindu & Perera, Srinath. (2020). An Analysis of Throughput and Latency Behaviours Under Microservice Decomposition. 10.1007/978-3-030-50578-3_5.

[11] Dayarathna, Miyuru & Perera, Srinath. (2018). Recent Advancements in Event Processing. ACM Computing Surveys. 51. 1-36. 10.1145/3170432.

[12] Dayarathna, Miyuru & Suzumura, Toyotaro. (2013). A Performance Analysis of System S, S4, and Esper via Two Level Benchmarking. 8054. 225-240. 10.1007/978-3-642-40196-1_19.

[13] G. Mazlami, J. Cito and P. Leitner, "Extraction of Microservices from Monolithic Software Architectures," 2017 IEEE International Conference on Web Services (ICWS), Honolulu, HI, 2017, pp. 524-531, doi: 10.1109/ICWS.2017.61.

[14] Wuxia Jin, Ting Liu, Qinghua Zheng, Di Cui, Yuanfang Cai, "Functionality-Oriented Microservice Extraction Based on Execution Trace Clustering", Web Services (ICWS) 2018 IEEE International Conference on, pp. 211-218, 2018.

[15] Justas Kazanavičius, Dalius Mažeika, "Migrating Legacy Software to Microservices Architecture", Electrical Electronic and Information Sciences (eStream) 2019 Open Conference of, pp. 1-5, 2019.

[16] Francisco Ponce, Gastón Márquez, Hernán Astudillo, "Migrating from monolithic architecture to microservices: A Rapid Review", Chilean Computer Science Society (SCCC) 2019 38th International Conference of the, pp. 1-7, 2019.

[17] Sara Hassan, Rami Bahsoon, Rick Kazman, "Microservice transition and its granularity problem: A systematic mapping study", Software: Practice and Experience, 2020.

[18] A. Krause, C. Zirkelbach, W. Hasselbring, S. Lenga and D. Kröger, "Microservice Decomposition via Static and Dynamic Analysis of the Monolith," 2020 IEEE International Conference on Software Architecture Companion (ICSA-C), Salvador, Brazil, 2020, pp. 9-16, doi: 10.1109/ICSA-C50368.2020.00011.

[19] W. Hasselbring and G. Steinacker, "Microservice Architectures for Scalability Agility and Reliability in E-Commerce", Proceedings of the IEEE International Conference on Software Architecture Workshops, pp. 243-246, 2017.

[20] Holger Knoche, Wilhelm Hasselbring, "Using Microservices for Legacy Software Modernization", Software IEEE, vol. 35, no. 3, pp. 44-49, 2018.

A Home Intrusion Detection System using Recycled Edge Devices and Machine Learning Algorithm

Daewoo Kwon[1], Jinseok Song[2], Chanho Choi[3], Eun-Kyu Lee[4]*

Department of Information and Telecommunication Engineering
Incheon National University, Incheon, Korea

Abstract—**This paper proposes a home intrusion detection system that makes the best use of a retired smartphone and an existing Wi-Fi access point. On-board sensors in the smartphone mounted on an entrance door records signals upon unwanted door opening. The access point is reconfigured to serve as a home server and thus it can process sensor data to detect unauthorized access to home by an intruder. Recycling devices enables a home owner to build own security system with no cost as well as helps our society deal with millions of retired devices and waste of computing resources in already-deployed IT devices. In order to improve detection accuracy, this paper proposes a detection method that employs a machine learning algorithm and an analysis technique of time series data. To minimize energy consumption on a battery-powered smartphone, the proposed system utilizes as few sensors as possible and offloads all the computation to the home edge server. We develop a prototype and run experiments to evaluate accuracy performance of the proposed system. Results show that it can detect intrusion with probability of 95% to 100%.**

Keywords—Security; intrusion detection; edge computing; Internet of Things; recycling

I. Introduction

Home security has been a social problem that severely threats lives and activities of the public. These days, people have installed a smart lock at a front door and a web camera at home to enhance their security levels. However, unauthorized intrusion also uses more intelligent and various ways. For instance, it is reported that many web cameras are vulnerable to attacks like malware and that people do not change default passwords set from a factory [1]. Moreover, these devices only protect home in a passive way. A home owner can see the inside of the house, only when she opens a corresponding mobile application; she may be busy on real intrusion events. Registering a professional home security service is another option for her. But, this require service users to pay fees and sometimes to purchase additional equipment, which does not attract people in an economical sense. Moreover, sensitive data such as activity records at home is stored and managed at a server of the service company, introducing privacy issues.

To tackle the problem, this paper proposes a new intrusion detection system that composes of recycled devices that have not been used for a while and/or already exist at home. More specifically, the system (i) uses a retired smartphone and (ii) reuses a Wi-Fi access point device, which saves an additional cost as well as solves other issues.

With fast advancement of a smartphone technology, people tend to upgrade to new smartphones frequently. A tons of old smartphones are abandoned every year or remain unused at home. Unfortunately, we have not found the right way to take advantage of retired smartphones. It is observed that such smartphones are still with advanced on-board sensors such as acceleration sensors, gyro sensors, and humidity sensors. The proposed system tries to make full use of the sensors. A retired smartphone in the system is mounted on a door, and on-board sensors recognize the door opening, which detects unauthorized accesses to a house by intruders. A Wi-Fi access point is a ubiquitous device installed at almost every houses, or even at every rooms; it is hard to imagine home without wireless connectivity today. Such a device is often with computing capability, and latest ones are equipped with powerful CPUs. However, it is observed that the device operates only as a router delivering network packets. Most computing resources remain unused and wasted. The proposed system tries to reuse surplus computing resources at the existing home device. This paper builds a prototype of a Wi-Fi access point server that processes sensor data transmitted from the retired smartphone as well as provides direct wireless connectivity to the smartphone. Placing the server at home also gives benefits. A local data processing implements an edge computing paradigm [2, 3], which makes it possible to analyze data in real time and thus improves latency performance. An event of home intrusion can be immediately captured and appropriate reaction can be taken quickly. Since data is stored at the local server and under a full control of a home owner, the privacy issue is also resolved. We note that this approach also affect our society in a good way beyond personal advantages. Recycling retired, existing devices enables us to save cost for disposing old devices and to make the best of surplus IT resources, otherwise wasted, on devices deployed widely today.

The proposed intrusion detection system faces technical challenges. First, a smartphone is with cheap on-board sensors mainly because of a cost issue. Therefore, it is hard to expect high accuracy in sensor data. Next, the sensors react to a variety of physical phenomena and generate millions of data accordingly. Out of the flood of big data, the system should be able to identify an event of interest (unauthorized door opening in our system) with high probability and to filter out unrelated noise signals. Last, the smartphone runs on a battery, and thus the system should be designed to minimize energy consumption and to maximize its lifetime. In order to resolve the first and second challenges, this paper develops an intrusion detection method that employs the *k*-nearest neighbors

algorithm, a machine learning algorithm used for classification and regression. Our method also employs the dynamic time warping algorithm to avoid the curse of dimensionality in the *k*-nearest neighbors algorithm. In this way, the proposed method is able to detect an intrusion event in a lightweight manner. For the third challenge, our system tries to utilize as small numbers of on-board sensors as possible. In a prototype, it only turns on two sensors, an accelerometer and a magnetometer in the retired smartphone. In addition, the smartphone offloads all the computation to the home edge server, minimizing its duty to save energy. Since the server is AC-powered, it is safe from unexpected power outage.

While our research makes the best use of the retired smartphone, there are still limitations on it. One example is battery condition of the smartphone. In general, the smartphone is likely to be with a battery of bad health. This implies that it should be recharged frequently, which is a big troublesome. An emerging energy harvesting technology may reduce the battery issue. Turning off unnecessary jobs in the smartphone can be another solution approach. By default, the smartphone runs multiple tasks both in a foreground and a background. Many of them are not directly related to our system. Thus, a proper setup can minimize power consumption on the smartphone and thus extend its lifetime. Finding an optimal configuration becomes an interesting research topic that this paper does not address.

The rest of the paper is composed as follows. Section 2 reviews related works. The proposed intrusion detection system is described in detail with background knowledge in Section 3. Next section develops a prototype of the system, which is followed by experiments and performance evaluation of detection accuracy in Section 5. The last section concludes the paper.

II. Related Works

Wu et al. [4] developed a method that detected events of door opening/closing by using a barometer of a smartphone. The sensor measured fluctuation in air pressure inside a house when a door is opening and closing. However, the method turned to be ineffective if the house has another open window(s). Dissanayake et al. [5] proposed an algorithm that recognized events of multiple doors in an indoor environment using a microphone and sound recording capability on a smartphone. It employed an active sound probing and analysis of the Doppler shift that captured acoustic characteristics telling door states (open/close) via impulse response. Mahler et al. [6] developed SecureHouse that used sensors of a smartphone mounted on a wall near a door. The system captured unique vibration signatures of door events, opening and closing, as well as the rotation of a door. Gong et al. [7] proposed an infrastructure-free door event detection approach that utilized built-in magnetic sensors of a smartphone. A key observation is that magnetic signals change patterns especially when a smartphone passes through a door. Behringer et al. [8] proposed a car security system that used GPS and an accelerometer of a smartphone. It could detect a variety of car-related events such as door opening/closing, engine start, and movement. Unlike these research, our system makes use of more than two sensors at the same time, which helps accuracy

of event detection. Moreover, it adopts the concept of edge computing, which is implemented by reusing a Wi-Fi access point device.

Wireless signal data has been also used to recognize door events. Ohara et al. [9] proposed a method to detect events on indoor objects. Their system consisted of a Wi-Fi access point and a special Wi-Fi signal receiver that captured changes on Wi-Fi signal propagation during events and analyzed it to detect door opening. Similarly, Shi et al. [10] developed a system that deployed FM-radio signal receivers inside a room and used the same principle that any changes in an indoor environment affected the propagation of radio waves. Hnat et al. [11] developed Doorjamb, an object tracking scheme. It used many ultrasound sensors on doors that enabled to detect moving object in an indoor environment.

There are a list of research projects that make use of open source hardware platforms. Jabbar et al. [12] developed IoT@HoMe, an IoT based home automation system that monitored home conditions and automatically controlled home appliances. It utilizes a node microcontroller unit (NodeMCU) to implement a microcontroller / an Internet gateway that obtains data from sensors and forwards it to a cloud server. Ozeer et al. [13] designed a framework for Fog-IoT applications. Raspberry Pi devices were extensively used in the framework on which lights are controlled appropriately according to owner's modes. Kaur et al. [14] developed a home automation system using Arduino devices and GSM. These research enables to build cost-effective smart home systems, but scarcely addressed how to use retired smartphones that our system adopts to maximize cost benefit of a home intrusion system.

III. Proposed Intrusion Detection System

A. System Architecture

The proposed home intrusion detection system consists of a retired smartphone, a home edge server, and a user carrying a smartphone as shown in Fig. 1. The retired smartphone, mounted on a door, senses the door opening and closing and transmits sensor records to the home server using its built-in Wi-Fi connectivity. The server, placed inside home, analyzes the received data to determine occurrence of a security event. It also sends an alarm message to the user in the case of event occurrence. We note that the goal of our system is to develop an intrusion detection system at a reasonable cost. To this end, the sensors and the server are developed by recycling devices. The sensors are from a retired smartphone, and the server can be made by an edge device like a home Wi-Fi access point.

B. Operation Scenario

Our system operates in two steps: an initialization phase and an intrusion detection phase.

1) Initialization phase: After mounting the retired smartphone, a user registers it to the edge server so that the server recognizes sensor data from the smartphone as authorized inputs. The on-board sensors should be calibrated properly to guarantee accuracy of sensor data. Since our system uses a machine learning algorithm, training data should be collected and stored at the server initially. To this

end, the user repeats door opening and closing and sends base data to the server. Finally, the user activates the retired smartphone on. Fig. 2 shows illustration of the initialization phase.

2) Intrusion detection phase: Once our system is activated on, the sensors keep measuring and sending data to the server. The system consumes data from two on-board sensors: accelerometer and magnetometer. An accelerometer measures the rate of change of the velocity of an object (acceleration) with the unit of meters per second squared. It has enjoyed a variety of applications; especially it is useful for sensing vibrations or orientation in systems. In a modern smartphone, it measures acceleration on three axes to detect changes in orientation and to tell the phone. In this way, the smartphone knows itself up from down and rotates its own screen. A magnetometer measures magnetic fields along three perpendicular axes X, Y and Z. It produces voltage proportional to the strength and polarity of the field along the axis each sensor is directed. By varying its voltage output to a smartphone, it can tell the smartphone's orientation (rotation) relative to the Earth's north. The sensor reports data in the unit of micro Tesla [μT].

By using the sensor records, the server runs a detection algorithm developed based on *k*-Nearest Neighbors and Dynamic Time Warping algorithms to determine a door opening event. When an intruder tries to enter into home by opening the door, therefore, it is able to detect an abnormal event. If the data is deemed to be intrusive, the server sends a push alarm to the user, and the user can check video stream generated from a web camera installed at home. Fig. 3 demonstrate steps of the intrusion detection phase.

C. Background Knowledge

1) k-Nearest Neighbor (k-NN): *k*-NN is a machine learning algorithm for classification [15,16]. Once data is received, it searches k most similar data out of all the base data stored in it and determines which category the new data belongs to. Fig. 4 describes how it works.

The following is a simple example of the *k*-NN process where we assume that data progresses with two components [x, y].

a) Store base data in own database. Each base data is already classified based on own components x and y. Fig. 4(a) shows all the base data that is classified into two groups, A and B.

b) A new data (star marked) comes, and *k*-NN is asked to decide which group the data belongs to.

c) The *k*-NN algorithm first computes distances from the new data to all other data. A distance represents how similar the new data is to an existing base data (see Fig. 4(c)).

d) Select k most similar data (having short distance) out of all the base data. In Fig. 4(d), k is 3, and two of them are from group A and one from B. The new data is then classified into group A by the decision of the majority.
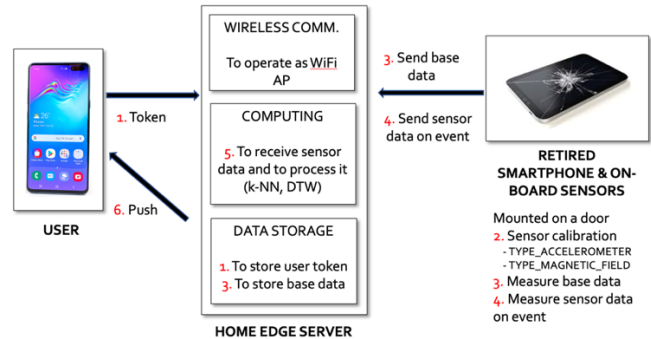
Fig. 1. The Proposed Home Intrusion Detection System Consists of a Retired Smartphone, a Recycled Home Edge Server, and a user. The Numbers, Red Colored, Represent Operation Steps.
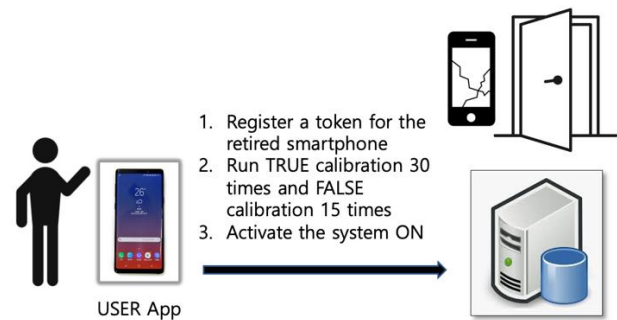
Fig. 2. An Initialization Phase Requires a Sensor Calibration Process.
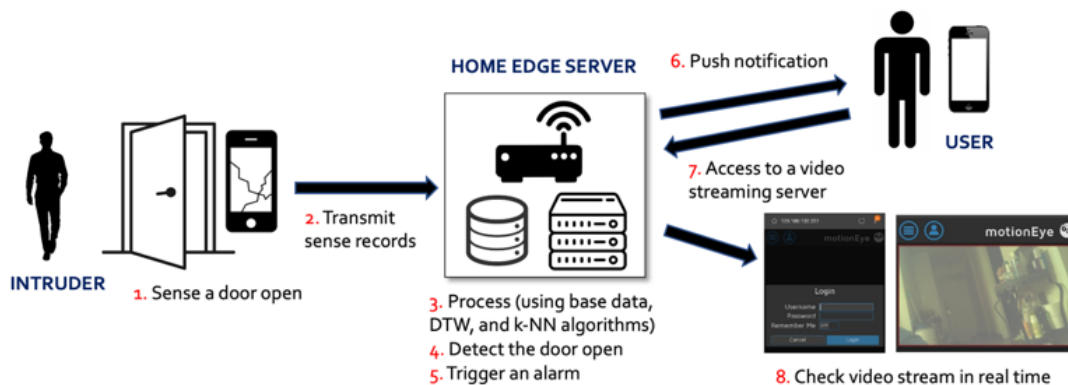
Fig. 3. The Detection Phase Senses an unauthorized Door Open, Pushes an Alarm Notification to a user, and Provides Video Stream in Real Time.
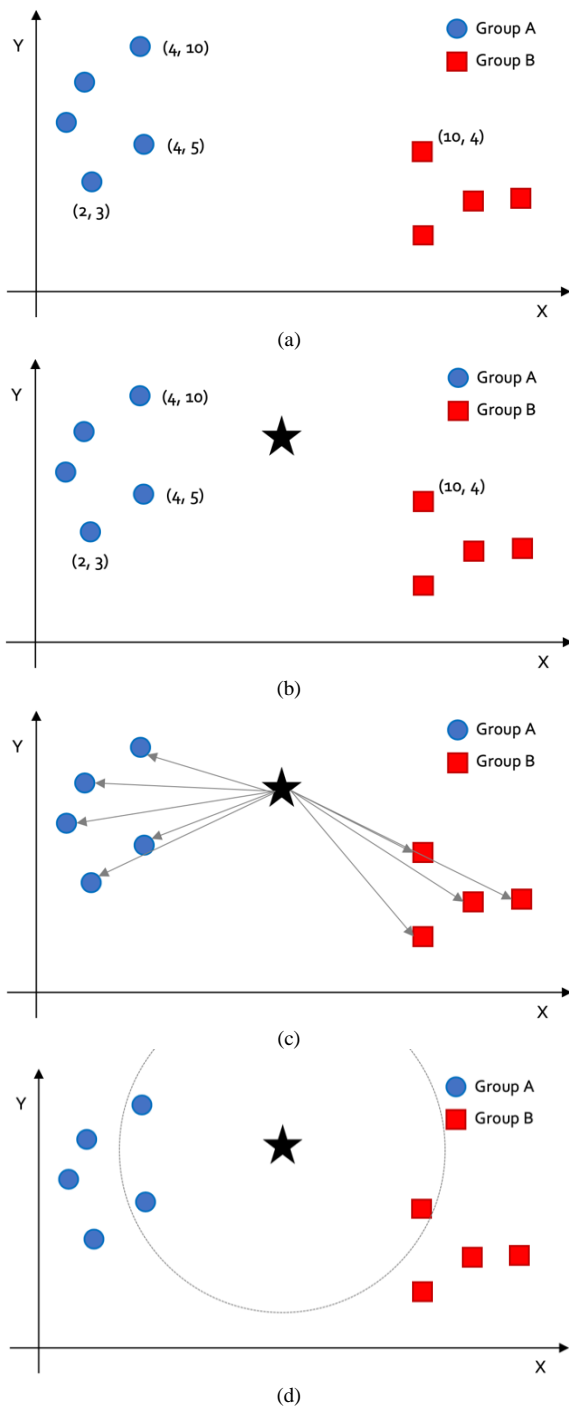
Fig. 4. Graphical Description of the Computation Process of *k*-NN Algorithm.

There are a few things to consider when applying *k*-NN. In the case of high-dimensional data, it is necessary to reduce the dimension because the curse of the dimension occurs when using the Euclidean distance formula. The proposed system measures the distance between high-dimensional data by using

a Dynamic Time Warping algorithm without reducing the dimension. Therefore, the *k* value should be chosen as an odd number approximating the square root of the underlying data.

2) Dynamic Time Warping (DTW): DTW is an algorithm for determining similarity between time series data [17]. In addition to comparing the same temporal index as itself, the neighboring indexes are also compared to select more similar elements as their pairs. Below is an example illustrating the DTW process. Suppose two time-series data *A* and *B*, each of which have six components as shown in Fig. 5.

DTW computes similarity of *A* and *B* in the following ways.

(i) Create a 6x6 matrix *M* and fill out all the elements with a predefined initial value, named max.

(ii) Set $M(0,0) = |A0 - B0|$. In rows of the 1st column, i.e., $M(1,0) \sim M(5,0)$, $|Ax - B0|$ is added to the previous row. For instance, $M(1,0) = M(0,0) + |A1 - B0| = 6 + |2 - 7| = 11$. For columns of the 1st row, $M(0,1) \sim M(0,5)$, $|A0 - Bx|$ is added to the previous column. $M(0,1)$ is computed as $M(0,0) + |A0 - B1| = 6 + |1 - 8| = 13$. In this way, the first rows and columns are filled out as shown in Fig. 6(a).

(iii) The rest of the elements, $M(1,1) \sim M(5,5)$, is computed as follows. For $M(i,j)$, pick three values, $M(i-1,j-1)$, $M(i-1,j)$, and $M(i,j-1)$, and find the minimum (say, $M(i,j,min)$). Then, $M(i,j) = M(i,j,min) + |Ai - Bj|$. For instance, $M(2,3) = M(2,3,min) + |A2 - B3| = min(13, 12, 17) + |2 - 5| = 12 + 3 = 15$. Fig. 6(b) shows outputs of this computation.

(iv) The last element, $M(5,5) = 34$, represents a DTW value. The last step routes back to the first element from this element. For $M(i,j)$, it pick the minimum out of three elements, $M(i-1,j-1)$, $M(i-1,j)$, and $M(i,j-1)$, and then jump to it. Repeat this until we reach to the first element. For instance, DTW is at $M(5,5)$ at first. Then, it jumps to $M(4,4)$, $M(3,3)$, $M(2,2)$, $M(1,1)$, and $M(0,0)$ sequentially as shown in Fig. 6(c). The distance is 5 in this example.

(v) Similarity between two data *A* and *B* is computed as DTW / distance = 34 / 5 = 6.8. The figure also shows that the route makes the main diagonal. This indicates that two data change constantly.

DTW in our system is used to derive similarity between newly received sensor data and underlying base data.

A	1	2	3	4	2	3
B	7	8	5	9	11	9

Fig. 5. An Example of DTW. This Assumes that there are Two Time Series Data, *A* and *B*, each of which Contains 6 Components, [A0, A1, …, A5, B0, B1, …, B5].

6	6+7 (13)	13+4 (17)	17+8 (25)	25+10 (35)	35+8 (43)
6+5 (11)	max	max	max	max	max
11+4 (15)	max	max	max	max	max
15+3 (18)	max	max	max	max	max
18+5 (23)	max	max	max	max	max
23+4 (17)	max	max	max	max	max

(a)

6	6+7 (13)	13+4 (17)	17+8 (25)	25+10 (35)	35+8 (43)
6+5 (11)	6+6 (12)	12+3 (15)	15+7 (22)	22+9 (31)	31+7 (38)
11+4 (15)	11+5 (16)	12+2 (14)	14+6 (20)	20+8 (28)	28+6 (34)
15+3 (18)	19	15	19	26	31
18+5 (23)	24	18	22	28	33
23+4 (17)	28	20	24	30	34

(b)

6	6+7 (13)	13+4 (17)	17+8 (25)	25+10 (35)	35+8 (43)
6+5 (11)	6+6 (12)	12+3 (15)	15+7 (22)	22+9 (31)	31+7 (38)
11+4 (15)	11+5 (16)	12+2 (14)	14+6 (20)	20+8 (28)	28+6 (34)
15+3 (18)	19	15	19	26	31
18+5 (23)	24	18	22	28	33
23+4 (17)	28	20	24	30	34

(c)

Fig. 6. DTW Computation makes a Matrix and Calculates Elements from Time Series Data.

IV. DEVELOPMENT

This section describes our development of the proposed intrusion detection system.

A. Retired Smartphone and On-Board Sensors

A retired smartphone, mounted to a door, serves as a group of sensors. That is, it senses events and sends measurement data to the server. We use Samsung SM-G900L that is with CPU of Qualcomm Snapdragon 2.5Ghz Quad-Core Krait 400, GPU of Qualcomm Adreno 330 578Mhz, and 2GB of memory [18]. It runs on Android 6.0.1 and API 23 versions. Out of a full list of on-board sensors, our system makes use of an accelerometer and a magnetometer. In order to control the retired smartphone and sensors, we develop a mobile application as shown in Fig. 7.

Using the application, a user is able to perform the initialization phase. First, a self-calibration measures values from the accelerometer and the magnetometer for 5 seconds when a door is closed and computes average values (see Fig. 7(a)). These values are used as references in a general situation (i.e., *stationary state*); they are stored in the x, y, and z reference variables of each sensor and are used in our detection algorithm to determine future events. We note that x, y, and z are from the accelerometer and represents three motion axes (components) related to the orientation of the retired smartphone. Next, the application starts a TRUE/FALSE calibration process (Fig. 7(b)). When the door is opened through a TRUE correction button, an array of 150 data is transmitted to the server 30 times, each of 50 x, y, and z coordinates. Thereafter, a FALSE correction button transmits an array of 150 data to the server 15 times. These values are stored in the server and used as reference when judging whether the door is on *intrusion state* (i.e., door opening) or on *noise state* (e.g., door knocking). Last, the user is able to activate and deactivate the retired smartphone ON and OFF through the buttons. Because sensor measurements during activation take place in the background service, it works even if the screen of the smartphone is turned off after installation is complete. The accelerometer and the magnetometer start sensing when activated. The retired smartphone sends accelerometer values to the server when an event occurs based on the mean value of a magnetic force value.
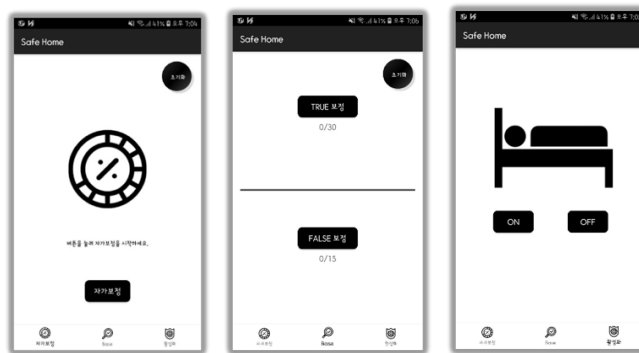
(a) Self-calibration   (b) Sending base data   (c) Activation

Fig. 7. A user makes an Initial Setup of the Retired Smartphone via a Mobile Application.

B. Recycled Home Edge Server

One of the keywords in our research is recycling; that is, our home server can be developed using existing home devices. For instance, a Wi-Fi router can be modified with OpenWrt [19] and support computing resources. Alternatively, an old PC at home can be used. Our system develops a home server using Raspberry Pi 3 Model B+ that is on Broadcom BCM2837B0, Cortex-A53 (ARMv8) 64-bit SoC @ 1.4GHz with 1GB memory [20]. The server provides three main functions; Wi-Fi access point, edge computing, and local database.

1) Wi-Fi Access Point (AP Mode): The server is configured so as to serve as a Wi-Fi access point, Access Point (AP) mode [21]. Thus, it provides an Internet connectivity service to all the home devices. With this configuration, the server is able to make a direct communication with the retired smartphone, which reduces transmission delay.

2) Data Processing for Intrusion Detection (Edge Computing Mode): Our server provides edge computing capability [2,3]. Because of performance and privacy issues, the edge computing is becoming a new computing model. Since it is not easy for non-IT people to use cloud computing, moreover, it is quite important to develop a fully localized system. In our system, the retired smartphone transmits raw data without pre-processing to the server. Instead, the server is responsible for data processing for intrusion detection, which helps save battery life of the retired smartphone.

The server converts a registration token received from the user and base data sent from the retired smartphone into a string and stores it in Redis database [22]. Upon receiving accelerometer data from the retired smartphone, it runs the *k*-NN algorithm to determine whether it is noise or meaningful data implying a door is actually opened. Noise can be either natural noise or environmental noise. Fine vibrations that a human being notices hardly are called natural noise, whereas impacts on doors such as knocking and vibrations from surrounding objects generate environmental noise. More technically, the retired smartphone computes the absolute value of difference between sampling data xN, yN, zN out of raw data and x, y, and z values obtained by self-calibration process at the initialization phase. Output of this computation, [x1, x2, x3, … , y1, y2, y3, ... , z1, z2, z3, ...], is then transmitted to the server. Then, the server normalizes both all 45 sets of base data and the received data; a random value X is divided by the maximum value out of the data, multiplied by 2, and minus 1. The server computes distance (similarity) between base data and the received data using the DTW algorithm. It then sorts the derived similarities in ascending order and select k values that are similar (the closer to zero, the similar). The majority vote is performed on the selected *k* value to determine whether the received data is noise or represents a door opening. When the reading results are not noise, the user is notified via push notification.

C. Web Video Service: Video Streaming in Real Time

Our system also implements a web video service by connecting a camera to the server and by using MotionEye open source library [23]. With this feature, a user can enjoy video streaming in real time by accessing the server as shown in Fig 8. For the convenience of service, the user can check for intrusions once receiving a push alarm. The library also offers a variety of functions including motion detection and storage. When using a motion detection function, the server automatically saves a video clip to a file upon detecting movement. Therefore, depending on a user's needs, a variety of functions can be set, such as setting images or storing images.

D. User

Our system also implements a simple mobile application that can be controlled by a user. By clicking ON or OFF button in the application, the user can easily register or unregister herself to the home server. On registration, a token is generated and transmitted to the server. Push notification can be received through the server when a door is opened upon activation. A camera button allows the user to access the web streaming server and watch real-time video.
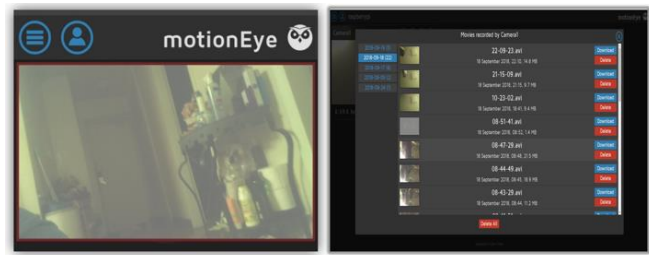
Fig. 8. A user can Double-Check Situation at the Door via Video Stream.

V. EXPERIMENTS AND PERFORMANCE EVALUATION

An eventual goal of this section is to evaluate detection accuracy of the proposed system. For experiments, the retired smartphone is mounted near a handle on the steel door of a classroom, as shown in Fig. 9, so that on-board sensors can spin when the door is opening. The first experiment measured base data that was then used as a training set. The next experiment recorded sensor data under various conditions, processed it using the proposed detection algorithm, and computed detection accuracy of the system.

Fig. 9. In our Experiment, the Retired Smartphone is Mounted on a Steel Door near the Handle.

A. Measurement of Base Data

The first experiment collected base data (i.e., ground truth data sets) that was classified into to two groups – TRUE event (the door is opening or intrusion state) and FALSE event (stationary state and noise state). All the base data included 30 TRUE events and 20 FALSE events. TRUE data represented 10 events of the door opening slowly, 10 events of the door opening at a normal speed, and 10 events of the door opening quickly. FALSE data represented 10 events of the door remained stationary (unopened) and 10 events of door knocking. We also ran experiments on a glass door that recorded 20 FALSE data; 10 events with a gentle wind and 10 events with the door shaking heavily.

Fig. 10(a) draws sensor records collected from the accelerometer when the door remains stationary. X-axis represents normalized sampling data, and Y-axis indicates acceleration values in the unit of meter per second per second. We note that all the graphs in this subsection are drawn on the same X-Y plane. Most values in the figure remain zero, representing tiny vibrations on the door. Two spikes are

measured; further investigation found that people were passing by the door at the moment. Fig. 10(b) shows acceleration values in the event of door knocking. Vibration is measured each time we knock the door. In the experiment, we knocked the door 4 times with high strength that were captured by 4 spikes at each axis. The smallest (the first) spike occurred when knocking at the point 1 meter away from the smartphone, whereas the biggest (the last) one when 10 cm away.

Fig. 11 shows raw data collected from the accelerometer when the door were opening at three different speeds. When the door were opening slowly, we observed that values at the X and Y samples were ignorable as shown in Fig. 11(a). Their average values are 0.022 and 0.021, respectively. But, variation at the Z samples is noticeable, up to 0.228. When speeding up the door opening, the values at the Z samples become higher; almost 5 times bigger at the peak, 1.023, than that at a slow speed (Fig. 11(b)). Unlike the Y samples, values at the X samples increase gradually and go up to 0.41 at max. This indicates that the door vibrates slightly from side to side. This movement becomes quite remarkable when opening the door at a high speed (Fig. 11(c)). The maximum value reaches to 1.057. We also note that values at the Z samples double and draw the same shape when comparing to those at the normal speed.

Our experiments also measured another 20 FALSE data, in which we set up a wobbly commercial glass door. The first experiment recorded sensor values on a gentle wind, and its results are demonstrated in Fig. 12(a). Values range from 0 to 0.09 that are under an internal threshold. When the door shook heavily, values at the Z samples fluctuated accordingly (Fig. 12(b). While they goes over 1.6 on the Y-axis, the up-and-down pattern is quite different from them in the case of door opening. The proposed detection algorithm is able to distinguish the patterns after a learning phase.

(a)

(b)

(c)

Fig. 11. Representation of Raw Data from the Accelerometer when a Door is Opening at a Slow Speed (a), A Moderate Speed (b), and A High Speed (c).

(a)

(b)

Fig. 10. Representation of Raw Data from the Accelerometer when a Door Remains Stationary (a) and when a Door is Knocked (b).

(a)

(b)
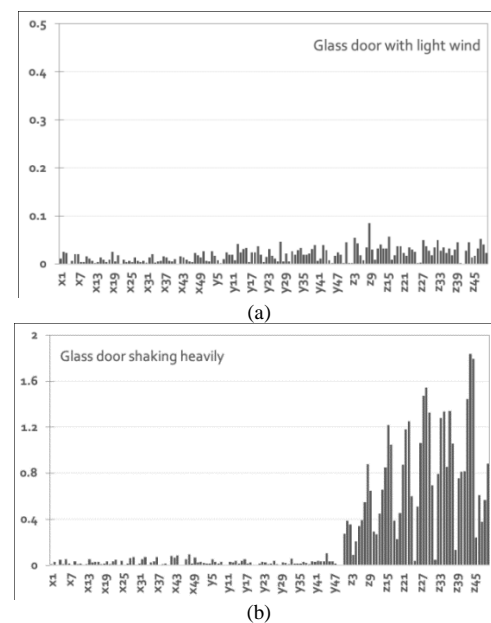
Fig. 12. Representation of Raw Data from the Accelerometer when there is a Light Wind (a) and when a Door is Shaking Heavily (b) on a Glass Door.

B. Evaluation of Accuracy of the Proposed System

This subsection describes our experiments to evaluate that the proposed system detects door opening under various conditions. We (i) knocked the steel door multiple times lightly and heavily, (ii) opened the door at 2 and 5 seconds of speeds, and (iii) approached a magnet to the door. The proposed detection algorithm made use of on-board sensors differently; (i) using accelerometer data only, (ii) using magnetometer data only, and (iii) using both of them together. We repeated the experiments both on a wooden door and on a glass door. The experiments were conducted 20 times each by producing various situations, and results are summarized in this subsection.

1) Using accelerometer data only: The detection algorithm in the first experiment that collected the accelerometer data only and processed it using two internal algorithms, *k*-NN and DTW. Its result of detection accuracy is shown at the first three columns from the left in Table I. *True Positive (TP)* represents the state that the proposed detection algorithm recognizes the door opening correctly, whereas *True Negative (TN)* implies the state that the algorithm correctly classifies input data as noise or recognizes that the door remains stationary.

When the door was knocked lightly, one event were falsely recognized as detection. As the strength of door knocking increased, the number of false detection also increased; 85% of detection accuracy with heavy knocking. Further investigation found that all the false detection occurred when knocking the door very fast. This condition drew a curve of sensor samples similar to that when opening a door slowly. When the door was opening at a moderate speed for about two seconds, detection accuracy recorded 70% with 14 correction detection and 6 false detection. When slowing down the speed (about five seconds), the number of false detection increased up to 8 with detection accuracy of 60%. The results indicate that the accuracy of door opening is quite low when accelerometer data is used only. In particular, when opening a door slowly Z-axis data was only sampled from the sensor and its values were small as demonstrated in the previous subsection. Therefore, the detection algorithm is likely to recognize the situation as noise with high probability. In the case of a magnet approaching, 100% detection accuracy was shown because the magnet did not affect movement of the door. Thus, the sensor did not capture meaningful data for detection of door opening.

2) Using magnetometer data only: The next three columns in the table represent results of the second experiment that uses the magnetometer data only in the detection algorithm. Detection accuracy in the cases of knocking and door opening shows 100%. These results are quite straightforward. As noted, the sensor can tell a device's rotation relative to the magnetic north, and here, the device is the door. Therefore, it can detect the door's movement accurately regardless of opening speeds, which results in high detection rate. However, it should be noted that the magnetometer also reacts to a magnet. When we approached a magnet to the door and manipulated it properly, the detection algorithm falsely recognized all the events as detection. Thus, the table shows 0% of accuracy. In summary, while magnetometer data is able to detect intrusion with high probability, it is not recommended to use it only because of critical weakness.

3) Using data from both accelerometer and magnetometer: Previous experiments show that it is not possible to detect intrusion correctly when using data from a single sensor. At the third experiment, thus, the detection algorithm makes use of data both from the accelerometer and the magnetometer. The last three columns in the table summarizes the experimental results.

In the case of door knocking, the strength of knocking generated tiny vibration on the door. Moreover, a small rate of false detection was corrected by the magnetometer. In this way, detection accuracy achieves 100%. In the case of door opening, detection performance improves remarkably, comparing to results from the experiment using the accelerometer data only; from 70% to 100% in slow opening and from 60% to 95% in fast opening. These values indicate that one can advance performance of a system by coordinating more data in general. But, we note that using many sensors and heterogeneous data from them may make a system complicated and introduce unwanted processing delay. Therefore, a special attention must be payed to. In the experiment of magnet approaching, detection accuracy of 100% was achieved. This result is mainly attributed to the role of the accelerometer; a magnet did not affect movement of the door, and thus the detection algorithm successfully recognized that the door was not opened.

4) Various door materials: The last experiment aims to evaluate performance of our detection algorithm on a wooden door and on a glass door in a residential building. To this end, we mounted the retired smartphone to them and set up the detection algorithm to use data from both the accelerometer and the magnetometer. The experiment ran three situations: door knocking, door opening, and magnet approaching.

Table II summarizes experimental results. In the experiments of door knocking (the first two rows from the top in the table), the algorithm recognized all the events correctly, resulting in 100% of accuracy on both materials. In the experiment of quick door opening, it also detected the events with 100% of probability. However, when opening the door slowly, 3 and 4 false detection were observed on the wooden and the glass doors, respectively. The 85% and 80% of accuracy are mainly attributed to the speed of door opening. Since the door was opened physically, the magnetometer sensed it clearly. The accelerometer recorded small values of acceleration because of the slow speed, which led our detection algorithm to classify them into noise.

TABLE I. DETECTION ACCURACY (ACC, %) IN DIFFERENT SETTING UNDER VARIOUS SITUATIONS. TP: TRUE POSITIVE AND TN: TRUE NEGATIVE

		Using accelerometer data only			Using magnetometer data only			Using both data		
		TP	*TN*	*Acc*	*TP*	*TN*	*Acc*	*TP*	*TN*	*Acc*
Knocking (multiple)	lightly	1	19	0.95	0	20	1.0	0	20	1.0
	heavily	3	17	0.85	0	20	1.0	0	20	1.0
Door opening	2 sec of speed	14	6	0.7	20	0	1.0	20	0	1.0
	5 sec of speed	12	8	0.6	20	0	1.0	19	1	0.95
Magnet approaching		0	20	1.0	20	0	0	0	20	1.0

TABLE II. DETECTION ACCURACY (ACC, %) ON DIFFERENT DOOR MATERIALS. TP: TRUE POSITIVE AND TN: TRUE NEGATIVE

Using both accelerometer and magnetometer data		On a wooden door			On a glass door		
		TP	*TN*	*Acc*	*TP*	*TN*	*Acc*
Door knocking (multiple)	lightly	0	20	1.0	0	20	1.0
	heavily	0	20	1.0	0	20	1.0
Door opening	2 sec of speed	20	0	1.0	20	0	1.0
	5 sec of speed	17	3	0.85	16	4	0.8
Magnet approaching		0	20	1.0	0	20	1.0

5) Discussion: Experimental results showed that detection accuracy reached up to 99% on average. This results can be compared to average accuracy of 93%~97% in previous works. This confirms that better performance is attributed to more data from multiple on-board sensors. We note that our experiments included a scenario that a door was opening at a very slow speed. Such slow movement could make some sensors hard to capture the door event. Since previous research never considered the scenario, it is not possible to compare performance directly. Our system could detect the event with high probability mainly thanks to using multiple sensors. With this approach, however, the retired smartphone generates/processes more data and consumes more energy. By offloading all the computing processes to a local home edge server, our system could reduce overhead on the smartphone.

Our experiment on various door materials is another highlight that may attract readers' interests. A primary role of a sensor in the emerging Internet of Things (IoT) is to capture phenomena (or events) occurring in physical environment. Different door materials represent different environment. In this sense, the sensor may recognize the same event (i.e., door opening) differently once it is on different physical materials. IoT research of future may want to consider such subtle interactions between sensors and physical environment of interests.

We also believe that our solution shows feasibility of recycling IT devices to our society as well as enables a home owner to build own security system with no cost. Recycling retired, existing devices enables us to save cost for disposing old devices and to make the best of surplus IT resources, otherwise wasted, on devices deployed widely today.

VI. CONCLUSION

This paper proposes a home intrusion detection system that is composed of recycled IT devices. It mounted a retired smartphone on an entrance door and used particular on-board sensors to records signals upon unwanted door opening. A Wi-Fi access point was reconfigured to serve as a home server, and thus it processed sensor data to detect unauthorized access to home by an intruder. The proposed system faced three technical challenges: inaccurate sensor data, detection of an event of interest out of the flood of big data, and energy consumption of the battery-powered smartphone. To solve the first two challenges, this paper proposed a lightweight detection method that employed the *k*-nearest neighbors algorithm and the dynamic time warping algorithm. The last challenge was resolved by using only two sensors and by offloading computation to the home edge server.

We developed a prototype where Samsung SM-G900L was reused and the server was implemented using a Raspberry Pi that was also equipped with a web camera. The first experiment measured base data to understand accuracy of on-board sensors in the smartphone as well as to calibrate sensor data. The next experiment considered three scenarios: (i) knocking a steel door multiple times lightly and heavily, (ii) opening the door at 2 and 5 seconds of speeds, and (iii) approaching a magnet to the door. It also set three options on the proposed detection algorithm. Experimental results demonstrated that detection accuracy could reach 95%-100% when the algorithm used data from an accelerometer and a magnetometer. The last experiment showed that our system was able to detection intrusion with high probability when using various doors of different materials.

REFERENCES

[1] Webcams vulnerable to attack. Available online: https://blog.malwarebytes.com/hacking-2/2019/09/15000-webcams-vulnerable-how-to-protect-webcam-hacking/ [accessed on 7/24/2020].

[2] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, " Fog computing and its role in the internet of things," in Proceedings of ACM workshop on Mobile Cloud Computing, Helsinki, Finland, August 2012.

[3] E. Hamilton, "What is Edge Computing: The Network Edge Explained," 2018. Available online: https://www.cloudwards.net/what-is-edge-computing/ [accessed on 7/24/2020].

[4] M. Wu, P. Pathak, and P. Mohapatra, "Monitoring Building Door Events using Barometer Sensor in Smartphones," in Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), Osaka, Japan, September 2015.

[5] T. Dissanayake, T. Maekawa, D. Amagata, and T. Hara, "Detecting Door Events Using a Smartphone via Active Sound Sensing," Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies., vol. 2, no. 4, article no. 160, 2018.

[6] M. Mahler, Q. Li, and A. Li. "SecureHouse - A Home Security System Based on Smartphone Sensors," in Proceedings of IEEE International Conference on Pervasive Computing and Communications (PerCom), Hawaii, US., March 2017.

[7] L. Gong, Y. Zhao, C. Xiang, Z. Li, C. Qian, and P. Yang, "Robust Light-Weight Magnetic-Based Door Event Detection with Smartphones," in IEEE Transactions on Mobile Computing., vol. 18, no. 11, pp. 2631-2646, 2019.

[8] R. Behringer, M. Ramachandran, and V. Chang, "A Low-Cost Intelligent Car Break-In Alert System Using Smartphone Accelerometers for Detecting Vehicle Break-Ins," in Proceedings of International Conference on Internet of Things and Big Data, Rome, Italy, April 2016.

[9] K. Ohara, T. Maekawa, and Y. Matsushita, "Detecting State Changes of Indoor Everyday Objects using Wi-Fi Channel State Information," in Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies., vol. 1, no. 3, article no. 88, 2017.

[10] S. Shi, S. Sigg, and Y. Ji, "Passive detection of situations from ambient FM-radio signals," in Proceedings of ACM Conference on Ubiquitous Computing (UbiComp), Pittsburgh, US., September 2012.

[11] T. Hnat, E. Griffiths, R. Dawson and K. Whitehouse, "Doorjamb: Unobtrusive room-level tracking of people in homes using doorway sensors," in Proceedings of ACM Conference on Embedded Network Sensor Systems (SenSys), Toronto, Canada, November 2012.

[12] W. Jabbar, T. K. Kian, R. Ramli, S. Zubir, N. Zamrizaman, M. Balfaqih, V. Shepelev, and S. Alharbi, "Design and fabrication of smart home with internet of things enabled automation system," in IEEE Access, vol. 7, pp. 144059-144074, 2019.

[13] U. Ozeer, L. Letondeur, F.-G. Ottogalli, G. Salaün and J.-M. Vincent, "Designing and implementing resilient IoT applications in the fog: A smart home use case," in Proceedings of Conference on Innovation in Clouds, Internet and Networks, Paris, France, February 2019.

[14] S. Kaur, R. Singh, N. Khairwal and P. Jain, "Home automation and security system", in Advanced Computational Intelligence: An International Journal (ACII), vol. 3, no. 3, pp. 17-23, 2016.

[15] T. Cover and P. Hart, "Nearest neighbor pattern classification," in IEEE Transactions on Information Theory., vol. 13, no. 1, pp. 21-27, 1967.

[16] N. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," in The American Statistician., vol. 46, no. 3, pp. 175–185, 1992.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," in IEEE Transactions on Acoustics, Speech, and Signal Processing., vol. 26, no. 1, pp. 43–49, 1978.

[18] Samsung Galaxy S5 (SM-G900L) Specification. Available online: https://www.phonemore.com/specs/samsung/galaxy-s5/sm-g900l/ [accessed on 7/24/2020].

[19] The OpenWrt Project. Available online: https://openwrt.org/ [accessed on 7/24/2020].

[20] Raspberry Pi 3 Model B+. Available online: https://www.raspberrypi.org/products/raspberry-pi-3-model-b-plus/ [accessed on 7/24/2020].

[21] Setting up a Raspberry Pi as a routed wireless access point. Available online: https://www.raspberrypi.org/documentation/configuration/wireless/access-point-routed.md [accessed on 7/24/2020].

[22] Redis. Available online: https://redis.io/ [accessed on 7/24/2020].

[23] MotionEye. Available online: https://github.com/ccrisan/motioneye/wiki [accessed on 7/24/2020].

Development and Verification of Serial Fault Simulation for FPGA Designs using the Proposed RASP-FIT Tool

Abdul Rafay Khatri[1], Ali Hayek[2], Josef Börcsök[3]
Department of Computer Architecture and System Programming,
University of Kassel, Kassel, Germany

Abstract—**Fault simulation is the critical approach for many applications such as fault detection & diagnostics, test set quality measurement, generation of test vectors, circuit testability, and many others along with the help of fault injection technique. The fault simulation approach is divided into many types. The most straightforward approach among them is a serial fault simulation. In the simulation process, the circuit under test is faulted, and a faulty copy is achieved by either using a simulator command technique or instrumentation technique. A fault simulator must examine the behaviour of specified target fault in design and classified as detected or undetected by the applied test patterns. To modify the original code is a very challenging and time-consuming task. Therefore, the RASP-FIT tool is developed, which alters the fault-free FPGA design, which is under investigation, at the Verilog HDL code level. It produces the copies of faulty design along with the top design file for several fault simulation methods. Using this tool, a serial fault simulation environment can easily be created with no much effort. In this work, a serial fault simulation method is verified and validated using the RASP-FIT tool for an ISCAS'85 benchmark design as an example.**

Keywords—*FPGA design; fault injection; fault simulation; RASP-FIT tool; Verilog HDL*

I. INTRODUCTION

Fault simulation and fault injection approaches are the most widely used techniques for confirming the functionality of Register Transfer Level (RTL) design and provide an approach to check the quality of test-benches and test patterns [1]. In the design environment, fault simulation is also practised for validation of test quality [2], [3]. By definition, fault simulation is the technique used to simulate a design in the presence of faults.

In comparison with logic simulation, a fault simulation method has an additional complexity due to the modelled faults in the design and their behaviour during the simulation. When performing simulation for a design, the measurement of Central Processing Unit (CPU) computations is almost proportional to these three things, which are the size of the circuit, the number of test vectors applied, and the number of modelled faults injected in design. Fault simulation methods are separated into five main methods, namely, [4], [5], [6]:

- Serial fault simulation

- Parallel fault simulation

- Deductive fault simulation

- Concurrent fault simulation

- Differential fault simulation

The most straightforward approach among them is the serial fault simulation in which an individual simulation is performed at any one time. There are two primary components of serial fault simulation, i.e. fault-free designs and faulty designs (the design which is consisted of faults and is termed as faulty design). Firstly, the fault-free logic simulation is executed on the fault-free design to achieve the fault-free output responses. After that, faulty designs are simulated with faults and responses are also obtained. Both responses are saved and compared to determine whether an applied test pattern can identify a fault or not [4], [7], [6], [8].

In serial fault simulation, one fault is simulated at a time. To generate the faulty design, fault injection is first introduced, which alters the original circuit, and the circuit behaviour is evaluated in the presence of the fault. The faulty circuit is simulated to determine the inadequate responses for the currently activated fault for the given test patterns applied to a fault simulation. This process repeats until all faults in the fault list have been simulated [5], [4], [9]. In serial fault simulation, one fault is activated at a time, and test patterns are applied until the fault is detected or all test patterns have been applied. After that, another fault is selected and activated as a new fault, the circuit should be at an initial state, and then the new faulty design is simulated. Repeat this process until all faults are tested [5]. Serial fault simulation needs multiple simulations runs on a standard gate-level simulator using built-in simulator commands. In this method, the complex data structures are not required. There are certain advantages and disadvantages of serial fault simulation technique and are addressed in the sequel.

- Advantages:- A few advantages of serial fault simulation are described below:
  1) Easy to implement.
  2) Any true value simulator can be used.
  3) Less memory is required if the concepts of fault dropping, fault collapsing etc. are introduced in process.

- Disadvantages:- A few disadvantages of serial fault simulation are described below:
  1) Much repeated computation.
  2) Large CPU time required for very large scale integration design.
  3) Not feasible for large design with the large number of inputs.

In previous research, Khatri et al. proposed and developed a fault injection tool and named it RASP-FIT (RechnerArchitektur und SystemProgrammierung-Fault Injection Tool) tool [10], [11], [12], [13], [14], [15], [16]. This tool is used to instrument the Field Programmable Gate Array (FPGA) based designs. FPGA-based designs are written in Hardware Description Languages (HDL). During the last few decades, HDLs have been involved in promoting several methods and techniques regarding digital system testing. These methodologies reduce the technological passage among the tools and techniques used by design and test engineers. Using HDL, the design engineers can verify, validate and test the design at an early step at the code level [5]. In this work, the serial fault simulation is performed for FPGA-based designs at the code level. Serial fault simulation is executed at the code level of the designs. To generate the faulty copies of the original designs, the RASP-FIT tool is used, which modifies the design for different fault models. The RASP-FIT is designed to perform different functions. In this paper, serial fault simulation is carried out, which proved that the RASP-FIT tool can be used to develop any fault simulation schemes/applications.

The organisation of the paper is as follows: Section II describes the brief introduction about the RASP-FIT tool and an environment for serial fault simulation. Section III presents the usage of the RASP-FIT tool in fault simulation applications, and the evaluation of the result is shown. Results are discussed in the Section IV. In the end, Section V concludes the paper and presents some future directions.

II. The RASP-FIT Tool and Simulation Environment

The RASP-FIT tool is proposed and developed by Khatri et al. using Matlab graphical user interface development environment at the University of Kassel, Germany. The RASP-FIT is a fault injection tool, which is developed to perform fault injection, testing, hardness analysis for the FPGA-based designs at the code level. At the code level, the design and test engineers can perform testing and verification at an early stage of the development cycle. The main advantage of fault injection at the code level is to create the state of the art methods and also develop the new methods with little effort. More details about this tool can be found in [11], [12], [13], [15], [16].

A. RASP-FIT Verilog Code Modifier

The Verilog code modifier function under the RASP-FIT tool consists of approximately 563 lines of code in Matlab having 20 functions. The RASP-FIT is a tabbed based tool. Verilog code modifier is tabbed under fault injection analysis and is shown in Fig. 1. To modify the design, the user needs to apply three inputs for code modification and generates compilable faulty design. These inputs are:

1) A Synthesizable Verilog design file.
2) Type of fault model for injection in the design from a drop-down menu.
3) A number of faulty copies the user wants to generate.

By clicking on the *Generate* button, faulty modules are created along with the top file. In order to differentiate one
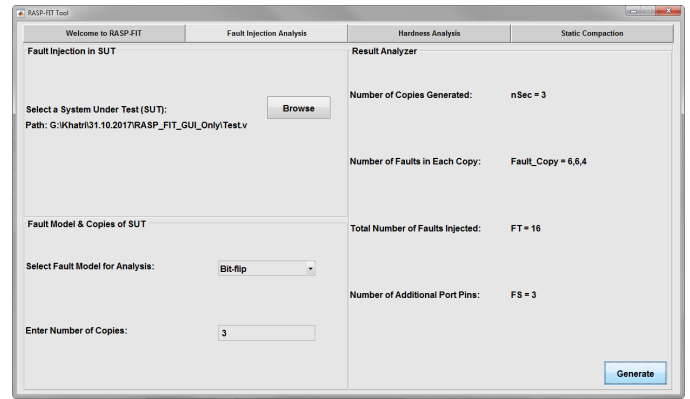
Fig. 1. Verilog HDL Code Modifier Tab under RASP-FIT Tool.

file from other files, the RASP-FIT tool saves the faulty modules under the names (moduleName_faultycopy1.v, moduleName_faultycopy2.v, and so on). The RASP-FIT also generates *Top.v* file which consists of the comparator logic, fault detection logic and digital logic Verilog code for storing the responses in the memory. These modified designs are now used for the fault simulation/emulation, digital testing and dependability analysis, with FPGA tools, without much effort. The development of this tool is presented in previous research [11], [12], [15], [13], [16], [17], [18]. In this paper, serial fault simulation is validated for ISCAS'85 benchmark designs.

Verilog HDL code describes the design at several abstraction levels, e.g. gate-level, data-flow, and behavioural levels. The way of modification of the code is different for each abstraction level, and also fault models are coded and modified the design at that abstraction level. There are two main components for serial fault simulations.

B. Fault-Free Design

The original FPGA design under investigation is called a fault-free module or golden module. It is a reference design for the comparison between the responses of faulty SUT and the fault-free design. Fault-free design is taken from the ISCAS'85 benchmark designs written in Verilog HDL. Fig. 2 (left) shows the original (fault-free) design code of `c17.v` benchmark design as an example.

C. Faulty Design

Fault-free designs are modified using fault injection technique and are called faulty design. The proposed RASP-FIT tool is used to create faulty designs for performing the serial fault simulation. In this work, FPGA-based designs written in Verilog HDL are considered. This tool injects various fault models in the design such as bit-flip, stuck at 1/0 fault models. Fig. 2 (right) shows the modified compilable code by the proposed tool.

1) FISA Unit: A fault control unit is an important component for fault simulation applications. Khatri et al. proposed a demultiplexer based fault control unit [15] and called it the FISA unit. The term FISA unit stands for Fault Injection, Selection and Activation unit. It is designed for fault injection investigation to examine the injection of faults, as shown in

```
// Verilog Design
// c17 Benchmark Circuit ISCAS'85

module c17 (G1,G2,G3,G6,G7, G22,G23);

input G1,G2,G3,G6,G7;
output G22, G23;
wire G10,G11,G16,G19;

nand G_1 (G10, G1, G3);
nand G_2 (G11, G3, G6);
nand G_3 (G16, G2, G11);
nand G_4 (G19, G11, G7);
nand G_5 (G22, G10, G16);
nand G_6 (G23, G16, G19);
endmodule
```

```
module c17_1 (select,G1,G2,G3,G6,G7,G22_f1,
    G23_f1);
input G1,G2,G3,G6,G7;
output G22_f1, G23_f1;
wire G10,G11,G16,G19;
input select;
wire fis =1;
reg f0;
always @ (select) begin
if (select == 1'd1) begin
f0=fis;end
else begin
f0=0;end
end

nand G_1 (G10,f0 ^ G1, G3);
nand G_2 (G11, G3, G6);
nand G_3 (G16, G2, G11);
nand G_4 (G19, G11, G7);
nand G_5 (G22_f1, G10, G16);
nand G_6 (G23_f1, G16, G19);
endmodule
```

Fig. 2. Fault-Free Design Code (Left) & Instrumented Compilable Code for Serial Fault Simulation (Right) by RASP-FIT.

Fig. 3. The FPGA design under test is written in Verilog HDL. Therefore, a fault control unit must be described in HDL code in the design. For that purpose, the function under RASP-FIT is included, which generates the code for the proposed FISA unit in each faulty copy. For large designs, as the number of injected faults increased, fault selection lines are also increased. De-multiplexer can be designed in Verilog HDL in various formats, e.g. using a case or if-else statements. De-multiplexer is a component which contains one input port (in this case Fault Injection Signal (FIS)), selecting one of many data-outputs, which is attached to the input port.


```
wire FIS = '1 '; //Declaration Part
input [1:0] select;
reg f0, f1, f 2, ... ;
always @ ( select ) begin
if ( select == 2'd0 ) begin
f0=FIS ; f1=0; f2=0; end
else if ( select == 2'd1) begin
f0=0; f1=FIS ; f2=0; end
else if ( select == 2'd2 ) begin
f0=0; f1=0; f2=FIS; end
else begin
f0=0; f1=0; f2=0; end
end
```
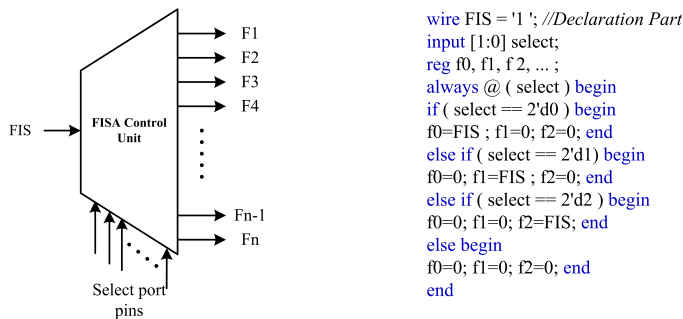
Fig. 3. Fault Selection and Activation Unit Schematic (Left), and Verilog Code (Right).

D. Experimental Set-up for Serial Fault Simulation

In serial fault simulation approach, one fault is inserted into the circuit at a time, and the fault effect is observed. The number of simulations is equal to the total number of faults plus one simulation (the initial simulation of the golden circuit) [9]. The block *c17.v* (original) is the fault-free circuit and considered as an example from ISCAS'85 designs. These FPGA designs are written in Verilog HDL code at the gate-abstraction level. The other blocks *C17(f1)* to *C17(fn)* are

faulty copy of the *c17.v* original design with faults *f1* through *fn* permanently instrumented as shown in Fig. 4. This method is an alternative technique for serial fault simulation to reduce CPU computation time.

The schematic diagram for the example *c17.v* is shown in Fig. 5. Fig. 6 shows the various locations of the design where faults can be injected to develop faulty designs to perform serial fault simulations. The mark (x) points out the location of the faults. Each (x) represents the fault models. Faulty copies are modified using instrumentation technique, and faults are injected in the design, for example, stuck-at 1/0 and bit-flip fault models.
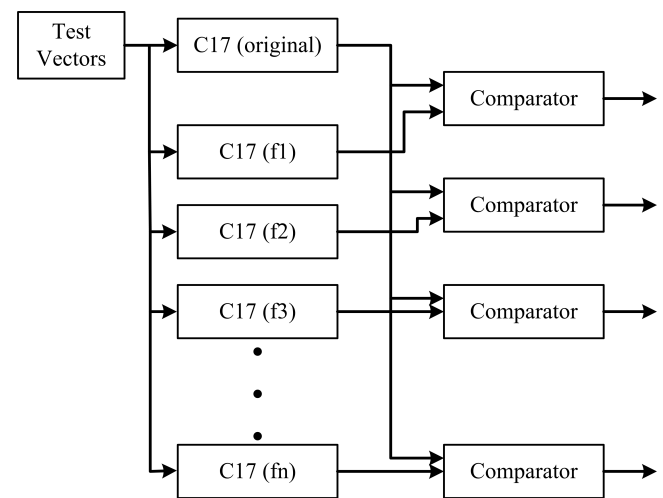
Fig. 4. Experimental Setup for Performing Serial Fault Simulation.

1) Fault Dictionary: When the fault simulation is performed by applying test vectors, and the result is obtained by comparing it to the responses of golden design. These results are organised in a simple two-column table where
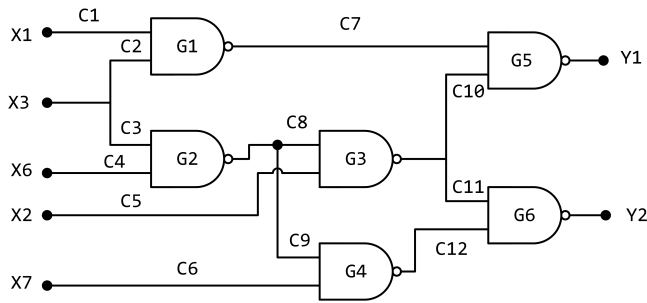
Fig. 5. Example of Design from ISCAS'85 Benchmark Circuits (Original Design).
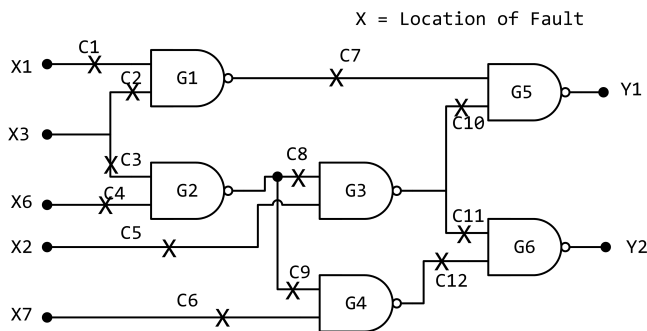
Fig. 6. Example of Design from ISCAS'85 Benchmark Circuits (with Faulty Locations).

every column correspond to circuit's faults and test vectors that detect them. This arrangement of data is called a fault dictionary. One fault can be detected by many test vectors and is called a detectable fault, whereas some faults cannot be detected by any of the test vectors, which is called undetectable faults.

Fault simulation is performed on fault-free and faulty designs. The responses are gathered in the presence of faults and presented in the tabular form. When a fault is injected in the design, then those test vectors which can detect that fault are placed in the fault dictionary [5]. For the large design with many input ports, this is not possible to simulate the design with all possible combination of inputs. Therefore, serial fault simulation is a simple approach which can be applied to the smaller designs with few faults and few test vectors. Table I depicts the example of fault dictionary. In the table, three faults {f1, f2, f3} are used during the fault simulation method, and test vectors (2,5,9) detect the f1 fault, test vectors (5,12,6) find f2 whereas f3 is not undetected with any test vector. This information helps design and test engineers to calculate Fault Coverage (FC).

TABLE I. EXAMPLE OF FAULT DICTIONARY

Fault No.	Test Vectors
f1	2,5,9
f2	5,12,6
f3	-

III. EXPERIMENTAL EVALUATION USING THE RASP-FIT TOOL

Serial fault simulation is a straight forward technique. It requires the fault-free circuit and the faulty circuits. The RASP-FIT tool is a fault injection tool which modifies the fault-free design and generates faulty modules. Along with the fault-free design, the user also needs two more inputs. The first one is a type of fault model used in serial fault simulation. The RASP-FIT can inject three fault models, i.e. bit-flip, stuck-at 0 and stuck-at 1 models separately and modifies the code accordingly. The second input is the number of copies required by the user. The RASP-FIT tool can evenly distribute the number of faults in different copies of the design, as shown in Fig. 7 (Step 1). In the second step, all these files are used to create a project using the Xilinx ISE tools, and Modelsim or Xilinx ISIM tools are utilised for simulation. After the simulation, results are stored in a text file. This text file is applied as an input to the Matlab script, which develops the fault dictionary and calculates the fault coverage for the design using Eq. 1. It is defined as the ratio of fault detected to the total fault injected.

$$FC = \frac{F_D}{F_T} \times 100\% \qquad (1)$$

where F_D is the number of detected faults during the serial fault simulation and F_T shows the total faults injected during the experiment. In the last step, the test bench must be added to the project to perform the simulation. For large designs, serial fault simulation technique is not a feasible solution, but the proposed RASP-FIT tool helps design and test engineers to perform serial fault simulation.
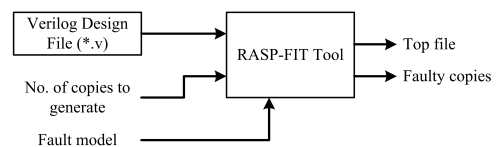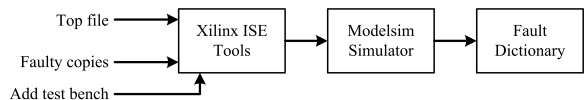
Fig. 7. Experimental Set-up using the RASP-FIT and FPGA Tool.

A. Top File Structure

The top file is generated under the RASP-FIT tool, which is used to perform serial fault simulation. The top file contains various elements, which are used to perform serial fault simulation, listed below:

1) Instantiations (golden and faulty copies).
2) Comparator logic.
3) Memory for storing responses.
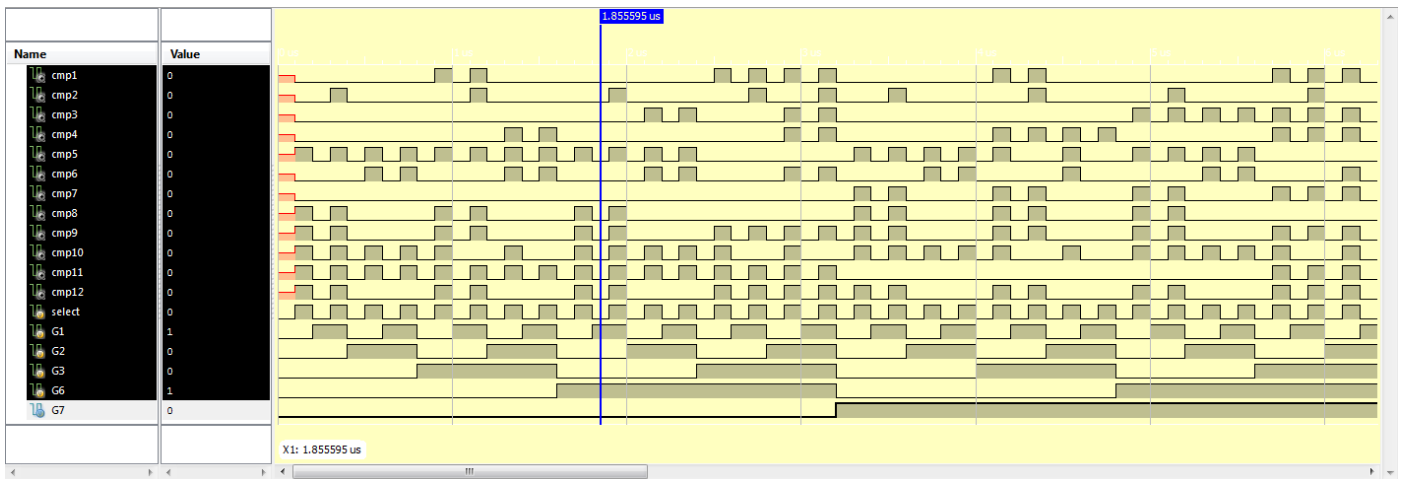
Fig. 8. Waveform for Serial Fault Simulation of *c17.v* Design.

IV. Result and Discussion

In this work, the RASP-FIT tool is used to perform and verify serial fault simulation applications. The objective is to validate the claim that, using RASP-FIT tool, the user can perform various fault simulation methodologies. An example from ISCAS'85 benchmark designs is considered to explain the serial fault simulation operation. As explained earlier, in serial fault simulation, one fault is injected at a time and input is applied to see whether the fault is detected or undetected. To save the time of simulation, authors use the experimental set-up as shown in Fig. 4. One fault is injected into the each copy of the design. The schematic of the c17.v design is shown in Fig. 5 and the Verilog code is illustrated in Fig. 2 along with the faulty copy containing one fault. The fault location is also marked on Fig. 6.

The simulation is performed using the Xilinx ISIM or Modelsim tool and waveform is shown in Fig. 8. Fault dictionary is constructed and shown in Table II. This example design is a simple design which consists of only 12 faults, and a total of 32 test vectors are applied as input to the design. Therefore, it is not a feasible idea for large designs having thousands of faults and many input ports to apply all possible combination of input. As every fault is detected at least one time hence fault coverage is 100%.

TABLE II. Fault dictionary for the example under simulation

Fault No.	Test Vectors	Detected (X) / Undetected (-)
f1	4,5,12,13,14,15,20,21,28,29,30,31	X
f2	1,5,9,13,15,17,21,25,29,31	X
f3	10,11,14,15,24,...,31	X
f4	6,7,14,15,20,21,23,28,...,31	X
f5	0,...,11,16,...,20,22,24,...,27	X
f6	2,3,6,7,10,11,14,15,19,22,26,27,30,31	X
f7	16,17,20,21,24,28,29,30,31	X
f8	0,1,4,5,8,9,16,17,20,21,24,25	X
f9	0,1,4,5,8,9,12,...,17,20,21,24,25,28,...	X
f10	0,...,4,6,8,...,12,14,16,...,20,22,24,28,30	X
f11	0,...,15,28,29,30,31	X
f12	0,1,4,5,8,9,12,...,17,20,21,24,25,29,...,31	X
F_T = 12	Total TV = 32	FC = 100%

V. Conclusion

Fault simulation technique assists designers and test engineers in several applications such as design's verification, test patterns generations and many other applications. The RASP-FIT is a simple, automatic and user-friendly fault injection tool, which works at the code level of the designs at several abstraction levels. This tool can inject faults in the whole design, and produce the compilable code. This tool helps the design and test engineers to perform various fault simulation applications. In this work, a serial fault simulation method is verified and validated. It is shown that, with the help of this tool, serial fault simulation can easily be performed. In future, parallel, deductive, concurrent and differential fault simulation approaches will be implemented and validated using the RASP-FIT tool.

References

[1] N. Bombieri, F. Fummi, and V. Guarnieri, "FAST-GP: An RTL functional verification framework based on fault simulation on GP-GPUs," in *2012 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 562–565, IEEE, Mar 2012.

[2] A. Parreira, J. P. Teixeira, and M. Santos, "A novel approach to fpga-based hardware fault modeling and simulation," *Design and Diagnostics of Electronic Circuits and Syst. Workshop*, pp. 17–24, 2003.

[3] S. Misera and R. Urban, "Fault Simulation and Fault Injection Technology Based on SystemC," in *Design and Test Technology for Dependable Systems-on-Chip*, pp. 268–293, 2011.

[4] L.-T. Wang, Cheng-Wen Wu, and Xiaoqing Wen, *VLSI TEST PRINCIPLES AND ARCHITECTURES*. 2006.

[5] Z. Navabi, *Digital System Test and Testable Design*. Boston, MA: Springer US, 2011.

[6] M. L. Bushnell and V. D. Agrawal, *Essentials of Electronic Testing for Digital, Memory and Mixed-Signal VLSI Circuits*. Kluwer Academic Publishers, 2002.

[7] F. Corno, G. Cumani, M. Sonza Reorda, and G. Squillero, "Effective techniques for high-level ATPG," in *Proceedings 10th Asian Test Symposium*, pp. 225–230, IEEE, 2001.

[8] S. Devadze, *Fault Simulation of Digital Systems*. PhD thesis, TALLINN UNIVERSITY OF TECHNOLOGY, 2009.

[9] Ian A. Grout, "Test Pattern Generation and Fault Simulation," in *Integrated Circuit Test Engineering*, ch. Chapter 10, pp. 235–255, London: Springer-Verlag, 2006.

[10] A. R. Khatri, M. Milde, A. Hayek, and J. Börcsök, "Instrumentation Technique for FPGA based Fault Injection Tool," in *5th International Conference on Design and Product Development (ICDPD '14)*, (Istanbul, Turkey), pp. 68–74, 2014.

[11] A. R. Khatri, A. Hayek, and J. Börcsök, "ATPG method with a hybrid compaction technique for combinational digital systems," in *2016 SAI Computing Conference (SAI)*, (London, UK), pp. 924–930, IEEE, Jul 2016.

[12] A. R. Khatri, A. Hayek, and J. Börcsök, *Applied Reconfigurable Computing*, vol. 9625 of *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2016.

[13] A. R. Khatri, A. Hayek, and J. Börscök, "Validation of the Proposed Fault Injection, Test and Hardness Analysis for Combinational Data-Flow Verilog HDL Designs Under the RASP-FIT Tool," in *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, (Athens, Greece), pp. 544–551, IEEE, Aug 2018.

[14] A. R. Khatri, A. Hayek, and J. Börcsök, "RASP-TMR: An Automatic and Fast Synthesizable Verilog Code Generator Tool for the Implementation and Evaluation of TMR Approach," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 590–597, 2018.

[15] A. R. Khatri, A. Hayek, and J. Börcsök, "RASP-FIT: A Fast and Automatic Fault Injection Tool for Code-Modification of FPGA Designs," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, pp. 30–40, 2018.

[16] A. R. Khatri, A. Hayek, and J. Börcsök, "Fault Injection and Test Approach for Behavioural Verilog Designs using the Proposed RASP-FIT Tool," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 4, pp. 57–63, 2019.

[17] A. R. Khatri, A. Hayek, and J. Börcsök, "Validation of the Proposed Hardness Analysis Technique for FPGA Designs to Improve Reliability and Fault-Tolerance," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 12, pp. 1–8, 2018.

[18] A. R. Khatri, "A Technical Guide for the RASP-FIT Tool," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, 2019.

Study of K-Nearest Neighbour Classification Performance on Fatigue and Non-Fatigue EMG Signal Features

W. M. Bukhari[1]*, C. J. Yun[2], A. M. Kassim[3], M. O. Tokhi[4]

[1,2,3]Centre of Excellence for Robotic, and Industrial Automation (CERIA)
Rehabilitation and Assistive Engineering Technology Research Group (REAT)
Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, Durian Tunggal, 76100 Melaka, Malaysia
[4]Cognitive Systems Research Centre, School of Engineering
London South Bank University, SE1 0AA, London, United Kingdom

Abstract—**For our body to move, the muscle must activate by relaxing and contracting. Muscle activation produces bio-electric signals that can be detected using Electromyography or EMG. The signal produced by the muscle is affected by the type of contraction done by the muscle. The eccentric contraction generating different EMG signals from concentric contraction. EMG signal contains multiple features. These features can be extracted using MATLAB software. This paper focuses on the bicep brachii and brachioradialis in the upper arm and forearm, respectively. The EMG signals are extracted using surface EMG whereby electrical pads are placed onto the surface of the muscle. Features can then be extracted from the EMG signal. This paper will focus on the MAV, VAR, and RMS features of the EMG signal. The features are then classified into eccentric, concentric or isometric contraction. The performance of the K-Nearest Neighbour (KNN) classifier is inconsistent due to the EMG data variabilities. The accuracy varies from one data set to another. However, it is concluded that non-fatigue signal classification accuracy is higher than fatigue signal classification accuracy.**

Keywords—Electromyography; surface electromyography; k-nearest neigbour classifier; feature extraction; dynamic contraction

I. INTRODUCTION

The movement of the body is a complicated process that requires a specific group of muscles to contract and relax in a specific order. Continuous contraction of the muscle over a prolonged period of time would eventually result in fatigue. Fatigue is the state of the muscle during the onset of fatigue caused by muscle contraction, whereas Non-Fatigue refers to the state of muscle before the onset of fatigue [1]. In order to extract information from the muscle to assess the Non-Fatigue and Fatigue condition of the muscle, sEMG, or surface electromyography is often used. sEMG is a method whereby electrodes are placed on the surface of the muscle to acquire the bioelectrical activity of the muscle. The signals acquired can then be analyzed for distinctive features that can be observed quantitatively such as the size of the signal or the shape of the signal: smooth or spiky, and fast or slow [2]. As the muscle enters a fatigued state, the ability of the muscle to do work would decrease as the muscle struggles to produce enough force to do the work. This, in turn, would affect the EMG signal produced as the muscle behaves differently.

Furthermore, the change in muscle state would also affect the behaviour of the muscle when doing the same type of work as in the previous state.

Generally, muscle contractions are divided into two groups: (1) Isometric and (2) Dynamic. These contractions are divided based on the movement of the muscle, which is lengthening, shortening, or static. Dynamic contractions are further divided into Eccentric and Concentric contractions. Contraction of the muscle when the muscle is lengthening is called the Eccentric Contraction, whereas contraction of the muscle when the muscle is shortening is called Concentric Contraction. Isometric contraction is when the muscle exerts a force, but the muscle length does not change [3].

A muscle uses ATP (Adenosine Triphosphate) to contract and shorten, producing a force on the objects it is connected to [4]. The upper arm is located between the shoulder joint and the elbow joint. It contains four muscles: three in the anterior compartment (biceps brachii, brachialis, coracobrachialis), and one in the posterior compartment (triceps brachii) [3]. When a muscle contracts in a minimal interval, it is called a muscle twitch. There usually are two forms of muscle fibres, the slow-twitch fires and fast-twitch fibres. Slow-twitch fibres are muscle fibres that twitch for a shorter time. They are more unaffected to fatigue and have higher endurance, but they cannot generate rapid force. In contrast, high-twitch fibres are muscle fibres which can twitch for a longer time [5]. They have higher endurance but low resistant to fatigue, but they can generate rapid force. The biceps muscle is consists of 46% slow-twitch fibre and 54% fast-twitch fibre. The composition of the muscle fibre will have some effect on the sEMG signal obtained as a different type of fibre will produce different types of neuromuscular activity [6].

"Fatigue" is generally a term which is used to describe the decrease in the physical performance of the muscle group in performing a particular task related to the muscle group. Relating to exercise, fatigue can be defined as the failure to maintain the exercise in the correct form. Fatigue is caused by multiple physiological phenomena such as "central fatigue" and "peripheral fatigue". "Central fatigue" indicates the decrease in the voluntary contraction of the muscle, whereas

*Corresponding Author

"Peripheral fatigue" indicates the decrease in the contraction strength of the muscle. During exercise such as the "bicep curl", the fatigue onset is usually caused by the "Central Fatigue" whereby fatigue is induced progressively by the exercise performed and degrading the voluntary contraction of the muscle [7]. Another way to look at muscle fatigue as seen in [8],[9]. It can be described as the reduction in the maximal force a muscle can produce in a related task that utilizes the specific muscle.

The steps to reduce skin-to-electrode impedance is divided into several types [10]. Firstly is the surface preparation. The most effective method, according to [11], was found to be applying abrasion on the surface with electrolytic gel. Next is the electrode metal. Ag is chosen over Cu as Cu tends to corrode faster, making it unsuitable for most electrode applications. After that is the electrode size. After extensive comparison by W. Besio et al. [10], it was discovered that a smaller diameter of concentric ring electrodes yields better and more consistent. Other than that, applying pressure onto the electrode will help to keep it in place, reducing motion artefact.

Time statistics based features are preferred over frequency domain and time-frequency domain features due to the better accuracy it provides [12]. Muscle endurance is measured with time. Thus, the time-domain feature is useful in detecting muscle fatigue. From the amplitude of the EMG, we can observe the pattern of muscle contraction. These amplitudes are affected by the rate of firing of motor units [13]. Therefore, the time-domain feature is chosen to be used in this research. Other than that, time-domain features have lower computational complexity and can be implemented faster.

"Inherent noise" is a type of electrical noise that is present in all electronic equipment. This inherent noise has a frequency range of 0 Hz to several thousand Hz. Surface EMG uses non-invasive or adhesive type electrodes which can be directly applied on the skin of the subject. Silver/silver chloride (10×1 mm) have an adequate signal-to-noise ratio and is very steady electrically. This makes it suitable to use as surface electrodes [14]. The impedance is inverse with the electrode size, the larger the electrode, the smaller the impedance. However, the electrode size must be in accordance with the SENIAM standard. Therefore, both parameters should be taken into consideration. Other than that, the inherent noise can also be eliminated with suitable circuit design or higher quality equipment. The motion artefact is created when there is the movement of the cable connecting the electrode to the amplifier. The association between the detection surface of the electrode and the skin also creates motion artefacts. When the muscles activate, it generates electrical activity or bioelectrical signals. To record the EMG signal, electrodes are placed on the skin covering the target muscle group. During the contraction and relaxation of the muscle, the muscle changes in length and stretches the skin. This will cause the electrode on the skin to move. This undesired movement of the electrode will cause motion artefact. Typically, the motion artefact has a range of 1 Hz to10 Hz. The voltage generated from this noise has an amplitude that is comparable to the amplitude of the EMG signal. This type of motion artefact can be removed by using recessed electrodes. Recessed electrodes are electrodes which have a conductive layer of gel between the skin surface and the electrode. This helps the electrode to detect EMG signals despite the motion of the skin provided the electrode is in contact with the skin. However, recessed electrodes cannot remove motion artefact caused by the potential difference between skin layers. Thus, this type of artefact must be removed by reducing skin impedance [15]. Scratching the skin reduces these artefacts [10]. Alcohol or any alcohol rubbing solution or gel can be used to apply abrasion on the skin to reduce skin impedance.

For dynamic contraction, acquiring an EMG signal can be difficult as noise can be quickly introduced into the signal from the swaying of the electrical wire used for EMG. These noises can be removed in a study done on the bicep brachii by Kuthe and co. [16], the muscle strength in terms of muscle force and fatigue for biceps brachii, was computed using supervised and unsupervised subjects in static contraction (isometric contraction). The raw EMG signals were filtered using a band-pass filter (second-order Butterworth) with a frequency range of 5-500Hz. Using this method, they were able to filter out most of the noise.

Specific features such as Mean Absolute Value (MAV), Root Mean Square (RMS), and Variance (VAR) can be extracted from the EMG signal. These specific features are useful in detecting muscle fatigue as it is a time-domain feature, and muscle endurance is measured with time [12].

In this study, the performance of the KNN classifier in classifying fatigue EMG signal will be compared to the non-fatigue EMG signal. Section 2 will describe the methods used to drive the muscle into fatigue condition from the non-fatigued condition while doing dynamic and isometric contraction. The KNN classification performance will be shown in Section 3. Section 4 discusses the observations made in Section 3, which later a conclusion will be made from the observation made in Section 4.

II. METHODOLOGY

A. Subject

Six healthy male volunteers without any history of neuromuscular skeleton disorder participated in this study. A consent form was filled by the volunteers before participating in the study. The subjects were asked to avoid any activity that may overexert their body 12 hours before the experiment. The volunteers were given instruction and training without load before starting the exercise.

B. Equipment Selection

The equipment chosen to be used in this study is:

- Vernier EKG-BTA sensor
- LabQuest Mini
- Vernier Disposable electrode

The EKG-BTA Sensor evaluates cardiac electrical potential shown as waveforms (voltages formed during the contraction of the heart). It can be used to make standard 3-lead EKG tracings to record electrical activity in the heart, or to collect surface EMG recordings to study contractions in muscles in your arm, leg or jaw. From Fig. 1, the EKG-BTA sensor is

equipped with a connector which connects to the interface (LabQuest Mini) and three leads (Red, Green, and Black) which connect to the electrodes. The Red, Green and Black leads represent positive, negative, and ground, respectively. The leads will relay the signal detected by the electrodes to the sensor.

The Vernier LabQuest Mini is a powerful (Fig. 2), affordable, and easy to use sensor interface for computer-based data collection. It receives the data from the EKG-BTA sensor and converts it into a format that can be easily interpreted. The example of the format is shown in Fig. 4 and 5.

The Vernier Disposable electrode is adhesive type electrodes (Fig. 3). They have adhesive of the sensor face to attach to the surface of the subject. The adhesive surface is covered out of the box, and the cover can be removed when needed, much like a sticker.

C. Data Collection

Before proceeding with the experiment, the skin of each subject must be cleaned with alcohol rub in order to reduce skin impedance and acquire more precise EMG signals. Isopropyl alcohol was used in this study. According to SENIAM or Surface Electromyography for the Non-Invasive Assessment of the Muscle, the inter-electrode distance must be 2cm from the centre point to centre point. The electrodes are applied parallel to the muscle fibre direction. In this study, the muscle group that will be focused on is the Biceps Brachii (long head and short head) and the brachioradialis.

Fig. 1. The Vernier EKG-BTA Sensor that Connects Electrodes to the LabQuest Mini.

Fig. 2. The LabQuest Mini Collects Data from Electrodes and Transfers to PC.

Fig. 3. The Vernier Disposable Electrode. These Electrodes are Disposable and must be Replaced after usage.
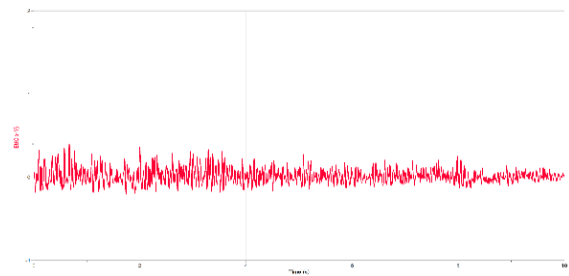
Fig. 4. EMG Signal of Biceps Brachii.
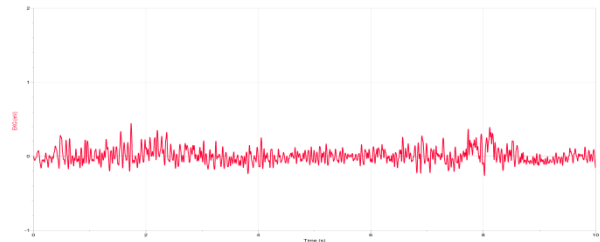
Fig. 5. EMG Signal of Brachioradialis.

There are a total of six subjects that participated in this study. The average age, height, weight, and BMI (Body Mass Index) of the subjects are listed in Table I.

TABLE I. AVERAGE AGE, WEIGHT, HEIGHT AND BMI OF SUBJECTS

Average Age	Height (m)	Weight (kg)	BMI (kg/m2)
24	1.68	64.7	22.9

The subjects are required to do the "bicep curl" exercise using a standard 5kg dumbbell. For the movement of the bicep curl exercise, the subjects are to lean against a wall with knees bent as shown in the image below, to limit the involuntary involvement of other muscles during the exercise. When doing the exercise, the subjects are to keep their elbows close to their body with palms facing forward. In the initial position, the subjects are to keep the dumbbell at the side of their body, perpendicular to the floor, as shown in Fig. 6.

The elbow is tugged to the side of the body to avoid the involuntary involvement of other muscle groups such as the front deltoid, especially during the fatigue state. The workout is separated into two stages, with three types of contraction: Eccentric, concentric, and isometric. The first stage is to acquire the non-fatigued condition of the muscle, and the second stage is to acquire the fatigue condition of the muscle.

Fig. 6. The Initial Position of the Bicep Curl Exercise.

The steps of the biceps curl exercise for isometric contraction are as shown below:

1) Subjects start in the initial position with elbow flexion at 90° empty-handed. Their back lay flat on the wall in the upright position with knees bent slightly at about 45°, and elbows tugged tightly to the side of the body. Throughout the exercise, only the elbow is allowed to move while the other parts of the body are to remain in the same position.

2) The 5 kg dumbbell is then placed in the palm of the subject for the subject to hold.

3) The subjects are to hold the position for 10 seconds.

4) After 10 seconds, the dumbbell is removed, and the subject rests for 2 minutes before starting the next set.

5) Step 1 to 4 is repeated for the next set.

6) A total of 3 sets were done.

For dynamic contraction, the initial position in step 1 is replaced with max muscle flexion with the dumbbell at shoulder height for eccentric contraction and minimum muscle flexion at waist height for concentric contraction. The subject then moves the dumbbell from shoulder level to waist level for eccentric contraction and from waist level to shoulder level for concentric contraction within 10 seconds in step 3. Step 4 to 6 is the same for all three types of muscle contractions.

D. EMG Signal Preprocessing

Pre-processing must be done in order to remove unwanted information such as crosstalk, motion artefacts and power line interference (PLI) from the raw EMG signal. To remove the motion artefacts in the EMG signal, a high pass filter needs to be used. The frequency range of motion artefacts is 10-20Hz, whereby the 10Hz represents walking artefacts while 20 Hz for the rapid movement artefacts.

In this case, we used 20 Hz to eliminate sudden rapid movements when executing the bicep curl exercise. The low pass Butterworth filter has been applied to a raw EMG signal to optimize the frequency of the EMG signal and to remove high-frequency components. The cut off frequency is set at 1000Hz. Besides filtering, the EMG signal was segmented before feature extraction according to the contraction. EMG signal was segmented into the 1-second window with 0.5 seconds overlap between segments to reduce data loss during feature extraction.

III. RESULTS AND DISCUSSION

From the result, three types of graphs are obtained. The scatter plot shows the predictions made by the classifier. Besides the plot, there is a legend for the predictions made. A dot means correct guess whereas a cross means the wrong guess. Each of the dots and crosses has its own colour. Blue for Concentric contraction "CON", orange for Eccentric contraction "ECC" and yellow for Isometric contraction "ISO". The plot provides a view on the distribution and range of the data as well as the region of error for the prediction.

The confusion matrix shows the performance of a classifier. The matrix shows the True Positive Rate and False Negative rate of the classifier based on the predictors "ECC", "CON", and "ISO". True Positive is indicated by green while False

Negative is indicated by red. The hue of the colour changes depending on the percentage of prediction. The more the predictions are True Positive or False Negative, the higher the percentage, the deeper the hue and vice versa. This shows which predictors have the most errors and vice versa.

The Receiver Operating Characteristics (ROC) curve shows the performance of the classifier. The AUC or 'Area Under Curve' is a measure of the degree of separability. It shows how well the classifier can categorize the data. The closer the AUC is to 1.0, the better the degree of separability, the better the classifier performs. The examples of the graphs obtained are shown in Fig 7, Fig. 8, and Fig. 9.
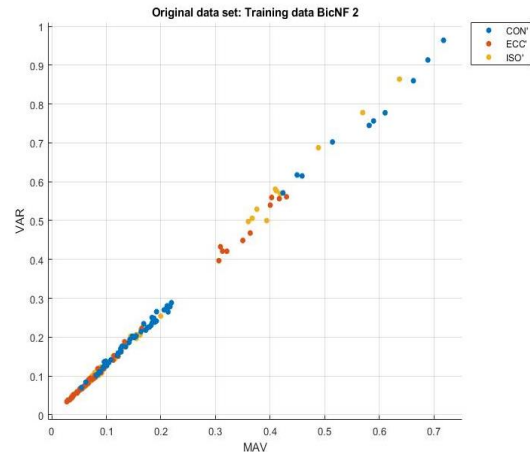
Fig. 7.   The Scatter Plot of the Data.
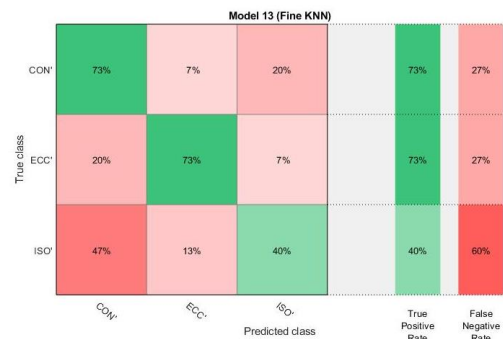
Fig. 8.   The Confusion Matrix of the Data.
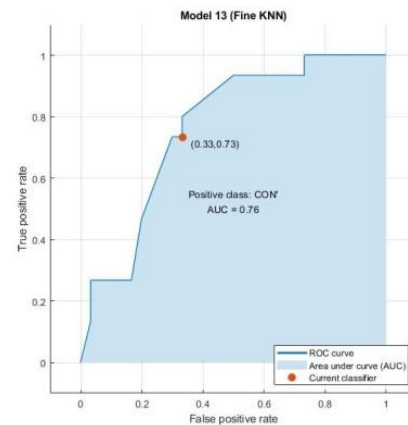
Fig. 9.   The ROC Curve of the Data.

The classifier is divided into four types. Each type of classifier corresponds to each data type:

- BicNF (Bicep Non-Fatigue)
- BicF (BicepFatigue)
- BraNF (Brachioradialis Non-Fatigue)
- BraF (Brachioradialis Fatigue)

Table II to Table V shows the performance of each classifier. Tables II and III in section A show the non-fatigued EMG signal classification performance, whereas Tables IV and V show the fatigue EMG signal classification performance. Analysis and discussion for each table are written below the respective table.

A. Non-Fatigue EMG Signal KNN Performance

Tables II and III show the summary of the result obtained from the Scatter Plot, Confusion Matrix, and ROC curve. These results are for the Non-Fatigue EMG signal KNN performance.

TABLE II.   PERFORMANCE SUMMARY TABLE OF BICNF KNN CLASSIFIER

		PATTERN	OUTLIERS
SCATTER PLOT	ECC	CONCENTRATED	3
	CON	CONCENTRATED	4
	ISO	CONCENTRATED	2
		TRUE POSITIVE	**FALSE NEGATIVE**
CONFUSION MATRIX	ECC	73%	27%
	CON	73%	27%
	ISO	40%	60%
	ACCURACY	62.2%	
		AUC	
ROC CURVE	ECC	0.86	
	CON	0.76	
	ISO	0.63	

TABLE III.   PERFORMANCE SUMMARY TABLE OF BRANF KNN CLASSIFIER

		PATTERN	OUTLIERS
SCATTER PLOT	ECC	CONCENTRATED	1
	CON	SPORADIC	-
	ISO	SPORADIC	-
		TRUE POSITIVE	**FALSE NEGATIVE**
CONFUSION MATRIX	ECC	33%	67%
	CON	73%	27%
	ISO	67%	33%
	ACCURACY	57.8%	
		AUC	
ROC CURVE	ECC	0.68	
	CON	0.81	
	ISO	0.80	

Referring to ROC Curve from Tables II and III, the AUC (Area Under Curve) of the BicNF Classifier for ECC, CON and ISO is 0.86, 0.76, and 0.63 respectively whereas the BraNF classifier has AUC of 0.68, 0.81 and 0.80 respectively. An AUC above 0.5 indicates the capability of separability. Both BicNF and BraNF had AUC above 0.5 for all EMG features. The closer the AUC is to 1.0, the better the performance of the classifier. This shows that the KNN classifier has a decent performance in classifying EMG features.

The accuracy of the BicNF classifier is 62.2% whereas for BraNF classifier is 57.8%. The KNN classifier can classify EMG signal features of Biceps Brachii better than brachioradialis for the non-fatigued condition.

Referring to the Confusion Matrix in Tables II and III, the BicNF classifier has an accuracy of 73% for ECC and 73% for CON whereas the BraNF classifier has an accuracy of 33% for ECC and 73% for CON. Compared to the BicNF classifier, the BraNF classifier was not able to classify Eccentric contraction features as well as the BicNF classifier. This is implied by the higher accuracy of the BicNF classifier compared to the BraNF classifier.

For the ISO classification, the BicNF classifier has an accuracy of 40% whereas the BraNF classifier has an accuracy of 67%. The KNN can classify the Isometric contraction feature better in brachioradialis compared to Biceps Brachii. This is also implied by the AUC as the ISO AUC of BraNF is 0.80, whereas the ISO AUC of BicNF is 0.63. This shows that the BicNF classifier has lower separability compared to the BraNF classifier in classifying Isometric contraction features.

Referring to Tables II and III, the performance of the KNN classifier was not affected by the pattern of the data scatter plot. This can be seen in the BicNF classifier when CON had four outliers, but accuracy remained high at 73.3% and also in the BraNF classifier when CON had most outliers but has the highest accuracy compared to ECC and ISO.

B. Fatigue EMG Signal KNN Performance

Tables IV and V show the summary of the result obtained from the Scatter Plot, Confusion Matrix, and ROC curve. These results are for the Fatigue EMG signal KNN performance.

Referring to ROC Curve from Tables IV and V, the AUC of the BicF Classifier for ECC, CON and ISO is 0.82, 0.89, and 0.75 respectively whereas the BraF classifier has AUC of 0.65, 0.61 and 0.82 respectively. An AUC above 0.5 indicates the capability of separability. Both BicNF and BraNF had AUC above 0.5 for all EMG features. This shows that the KNN classifier has a decent performance in classifying EMG features. The accuracy of the BicF classifier is 71.1% whereas for BraF classifier is 53.3%. The KNN classifier can classify EMG signal features of Biceps Brachii better than brachioradialis for fatigue condition. Referring to the Confusion Matrix in Tables IV and V, the BicF classifier has an accuracy of 73% for ECC and 80% for CON whereas the BraF classifier has an accuracy of 47% for ECC and 60% for CON. Compared to the BicF classifier, the BraF classifier was not able to classify Eccentric and Concentric contraction features as well as the BicF classifier. This is implied by the

higher accuracy of the BicF classifier compared to the BraF classifier.

TABLE IV. PERFORMANCE SUMMARY TABLE OF BICNF KNN CLASSIFIER

		PATTERN	OUTLIERS
SCATTER PLOT	ECC	SPORADIC	-
	CON	SPORADIC	2
	ISO	CONCENTRATED	5
CONFUSION MATRIX		TRUE POSITIVE	FALSE NEGATIVE
	ECC	73%	27%
	CON	80%	20%
	ISO	60%	40%
	ACCURACY	71.1%	
ROC CURVE		AUC	
	ECC	0.82	
	CON	0.89	
	ISO	0.75	

TABLE V. PERFORMANCE SUMMARY TABLE OF BRANF KNN CLASSIFIER

		PATTERN	OUTLIERS
SCATTER PLOT	ECC	SPORADIC	-
	CON	SPORADIC	5
	ISO	SPORADIC	3
CONFUSION MATRIX		TRUE POSITIVE	FALSE NEGATIVE
	ECC	47%	53%
	CON	60%	40%
	ISO	53%	47%
	ACCURACY	53.3%	
ROC CURVE		AUC	
	ECC	0.65	
	CON	0.61	
	ISO	0.82	

For the ISO classification, the BicF classifier has an accuracy of 60% whereas the BraNF classifier has an accuracy of 53%. The KNN can classify the Isometric contraction feature better in Biceps Brachii compared to brachioradialis. However, the ISO AUC of BraF is 0.82, whereas the ISO AUC of BicF is 0.75. This shows that despite having lower accuracy than the BraF classifier, the BicF classifier has a better separability compared to BraF. Referring to the scatter plot in Tables IV and V, the BicF classifier ISO has concentrated pattern but had the lowest accuracy at 60% when compared to ECC and CON. Since the data for ISO is concentrated, there should be less error in classifying as the distance between neighbours of the same data type is near. However, the ISO also has five outliers which may contribute to the lower accuracy. For the BraF classifier, CON had an irregular pattern as well as the most outliers at 5. However, the accuracy of CON is highest at 60%. This shows that the pattern of the data scatter does not affect the accuracy directly. This is supported by the AUC of the classifier. Despite having the highest accuracy, the CON AUC is lowest at 0.61. This shows that the BraF classifier has the lowest separability for CON features. Meaning it is least capable of classifying ISO features.

IV. CONCLUSION

In conclusion, the KNN classifier is not suitable for classifying time-domain features of EMG data due to inconsistent performance. The accuracy of the KNN classifier in classifying signals into eccentric, concentric, or isometric contraction during dynamic and isometric movement is subjective to the data used. Accuracy differs from one set of data to another. The accuracy of the KNN classifier in classifying time-domain features into eccentric, concentric or isometric contraction when the muscle is in Non-Fatigue condition is higher than in Fatigue condition. The lower accuracy for Fatigue condition may be due to the varying degree of muscle twitch when their muscle is fatigued for different individuals. Muscle twitch promotes much higher muscle action potential, which causes spikes in the EMG data. Even after applying filters to the raw data, there is still too much variation in the data from one person to another. The KNN classifier was not able to classify correctly due to this variation. Overall, the KNN classifier has higher accuracy in classifying Biceps Brachii EMG data compared to Brachioradialis EMG data. The accuracy of KNN in classifying dynamic and isometric movement is also inconsistent with varying degree of separability from one muscle group to another.

V. RECOMMENDATIONS

The number of volunteers that are involved in this study is six people. More volunteers would help to ensure the data obtained is not biased. The age group, height, and BMI of the volunteer is recommended to be about the same to ensure data collected is not biased. With more volunteers, the performance of the KNN classifier may become more consistent due to more training data. When doing the biceps curl exercise, it is best to maintain a straight posture while keeping elbow tugged tightly to the side of the body and keeping the arms perpendicular to the ground. Volunteers should breathe in a controlled manner while doing the exercise. The wire that connects the EKG-BTA and the surface electrodes should be fixed to any surface to ensure minimal wire sway when doing the exercise. The EKG-BTA must be kept further away from any electronics which may produce sound such as laptop fan. The main reason for the discrepancy in the result which causes inconsistent performance is due to the wrong method of comparison. A different individual would inevitably produce different EMG. Therefore, for future research, it is suggested to collect multiple sets of data from one individual and one test data from the same individual. The KNN is trained and tested with the data of the same individual. This is done for multiple individuals. The results of the test data from different individuals can then be compared with one another. This will remove the discrepancy resulted from the difference in physiology.The finding of this study may benefit as reference for those who are considering using the KNN classifier for the application of EMG signal classification.

REFERENCES

[1] M. R. Al-Mulla, F. Sepulveda, and M. Colley, "A review of non-invasive techniques to detect and predict localized muscle fatigue," Sensors, vol. 11, no. 4, pp. 3545–3594, 2011.

[2] A. Phinyomark, S. Thongpanja, H. Hu, P. Phukpattaranont, and C. Limsakul, "The Usefulness of Mean and Median Frequencies in Electromyography Analysis," Comput. Intell. Electromyogr. Anal. - A Perspect. Curr. Appl. Futur. Challenges, no. October 2012.

[3] J. Feher, "Skeletal Muscle Mechanics," Quant. Hum. Physiol., pp. 239–248, 2012.

[4] "Muscles of the Upper Arm - Biceps - Triceps - TeachMeAnatomy." [Online]. Available: https://teachmeanatomy.info/upper-limb/muscles/upper arm/#:~:targetText=The upper arm is located, posterior compartment (triceps brachii). [Accessed: 11-Dec-2019].

[5] K. Marri and R. Swaminathan, "Analysis of Biceps Brachii Muscles in Dynamic Contraction Using sEMG Signals and Multifractal DMA Algorithm," Int. J. Signal Process. Syst., vol. 4, no. 1, 2015.

[6] S. Boyas and A. Guével, "Neuromuscular fatigue in healthy muscle: Underlying factors and adaptation mechanisms," Ann. Phys. Rehabil. Med., vol. 54, no. 2, pp. 88–108, 2011.

[7] R. M. Enoka and J. Duchateau, "Muscle fatigue: What, why and how it influences muscle function," J. Physiol., vol. 586, no. 1, pp. 11–23, 2008.

[8] M. Sarillee, M. Hariharan, M. N. Anas, M. I. Omar, M. N. Aishah, and Q. W. Oung, "Non-invasive techniques to assess muscle fatigue using biosensors: A review," Proc. - 2014 5th IEEE Control Syst. Grad. Res. Colloquium, ICSGRC, 2014, no. October, pp. 187–192, 2014.

[9] Daud, W. M. B. W., Abas, N., and Tokhi, M. O. (2018). Effect of two adjacent muscles of flexor and extensor on finger pinch and handgrip force. In 2018 5th IEEE International Conference on Control, Decision and Information Technologies (CoDIT), pages 140–145

[10] W. Besio and A. Prasad, "Analysis of skin-electrode impedance using concentric ring electrode," Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc., pp. 6414–6417, 2006.

[11] P. Konrad, "The ABC of EMG," A Pract. Introd. to Kinesiol. Electromyogr., vol. 1, no. April, pp. 1–60, 2005.

[12] Y. Narayan, L. Mathew, and S. Chatterji, "SEMG signal classification with novel feature extraction using different machine learning approaches," J. Intell. Fuzzy Syst., vol. 35, no. 5, pp. 5099–5109, 2018.

[13] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, and G. Rau, "Development of recommendations for SEMG sensors and sensor placement procedures," Electromyogr. Kinesiol., vol. 10, no. 1, pp. 361–374, 2000.

[14] I. Himmelsbach, J. Lipinski, and M. Putzke, "Sampling, noise-reduction and amplitude estimation issues in surface electromyography," J. Electromyogr. Kinesiol., vol. 113, no. 11, pp. 933–942, 2002.

[15] Y. Blanc and D. Ugo, "Electrode Placement in Surface Electromyography(sEMG) 'Minimal Crosstalk Area,'" vol. 4, pp. 110–126, 2010.

[16] C. D. Kuthe, R. V. Uddanwadiker, and A. A. Ramteke, "Surface electromyography based method for computing muscle strength and fatigue of biceps brachii muscle and its clinical implementation," Informatics Med. Unlocked, vol. 12, no. March, pp. 34–43, 2018.

Road Object Detection using Yolov3 and Kitti Dataset

Ghaith Al-refai[1]

School of Electrical and
Computer Engineering
Oakland University
Rochester, Michigan 48309

Mohammed Al-refai[2]

School of Computer Science and Information Technology
Jordan University of Science and Technology
Irbid, Jordan 22110

Abstract—**Road objects (such as pedestrians and vehicles) detection is a very important step to enhance road safety and achieve autonomous driving. Many on-vehicle sensors, such as radars, lidars and ultrasonic sensors, are used to detect surrounding objects. However, cameras are widely used sensors for road objects detection for the rich information they provide and their inexpensive prices with compared to other sensors. Machine learning and computer vision algorithms are utilized to classify objects in the collected images and videos. There are many computer vision algorithms proposed for image and video object detection, e.g. logistic regression and SVM with feature extraction. However, Convolutional Neural Network (CNN) algorithms showed a high detection accuracy compared to other approaches. This research implements You Only Look Once (YOLO) algorithm that uses Draknet-53 CNN to detect four classes: pedestrians, vehicles, trucks and cyclists. The model is trained using Kitti images dataset which is collected from public roads using vehicle's front looking camera. The algorithm is tested, and detection results are presented.**

Keywords—*Pedestrian detection; computer vision; CNN; machine learning; artificial intelligence; vehicle safety*

I. Introduction

Due to the increase in road accidents, the necessity of improving vehicle safety has increased in the past few years. According to world health organization, more than 1.35 million people die every year because of vehicle accidents [1]. Vehicle safety features started with passive safety approaches such as the three points seatbelts [2]. After that, active safety features introduced in vehicles, such as airbags, Anti-lock Braking System (ABS). At the beginning of the 20th century, Advanced Driver Assistance Systems (ADAS) introduced in vehicles. ADAS utilizes on-vehicle sensors to detect surroundings and notify drivers to avoid accidents. ADAS can provide the assistance to drivers as visualization, warning and control. Back up camera is an example of a visualization assist, where a rear camera sends video information to the driver display. Lane Departure Warning (LDW) is an example of warning feature, where an alarm is sent to the driver when vehicle is leaving the ego lane without turning the blinker on. Automatic Emergency Braking (AEB) is an example of a control feature, where an automatic braking command is sent automatically by the ADAS module if an object is detected in the front of the vehicle. Ziebinski et al. provided a review for the recent ADAS features in vehicles [3].

In order to implement the ADAS features, road sensing to detect surrounding objects is required. Pedestrians, cars, trucks and cyclist are examples of common road elements that needs to be detected. Ultrasonic sensors are commonly used to detect objects within short range distance from the vehicle. Radars, Lidar and cameras are used for long distance detection. Machine learning algorithms are used to analyze and understand the collected information by the sensors and classify objects in a scene. The improvement achieved in machine learning and computer vision algorithms in the past few years, and the low cost of cameras compared to radars and Lidars, made cameras a widely used sensor in vehicles for object detection and tracking to implement ADAS features.

Many computer vision algorithms were developed for road object detection. Template matching is the basic approach for object detection. In such approach, a template is used to describe an object, then objects in captured images are compared to the template to check if it matches. Shen and Steng introduced an algorithm for vehicle detection using template matching [4]. James et al. [5] presented a two-stage template-based method to detect people in thermal images. A review for template matching algorithms for object detection was provided in [6].

Feature extraction and logistic regression classification algorithms are other approaches for objects classification. Feature extraction step describes objects by their unique features, such as edges, textures and contours. Then a trained classifier is used to classify the objects. Dalal and Triggs used Histogram Oriented Gradient (HOG) and Support Vector Machine (SVM) for human detection [7]. The algorithm showed a near perfect result in MIT pedestrian database. Tsai et al. [8] proposed a vehicle detection algorithm that uses three features, including corners, edge maps, and coefficients of wavelet transforms, then a cascade multichannel classifier is used to classify the features. Li et al. [9] proposed a method based on adaboost classifier for Haar-like features and linear Discriminant Analysis (LDA) to detect the traffic signs.

A detailed review for pedestrian detection system was done by Gerónimo et al [10]. Sivaraman et al. [11] provided an overview review of the past decade's literature in on-road vision-based vehicle detection.

Convolutional Neural Network (CNN) have recently shown outstanding results in objects detection and classification. The huge computations and the large memory requirement for the CNN made it difficult to implement for real time detection

applications. However, the hardware improvement of the processors, GPUs and memory, in addition to parallel processing, made it possible to implement real-time CNN algorithms with low cost.

In this work, Darknet-53 CNN with YOLO algorithm is used for road object detection. The Kitti dataset is adopted to train and test the algorithm and its dataset. The algorithm possibly detects four objects: cars, trucks, pedestrians and cyclists. This paper provides a brief review for related works. Section 3 presents the algorithm implementation and presents detection results. Finally, Section 4 summarizes the conclusion of this research work.

II. RELATED WORK

CNN is one of the most widely used machine learning technique in vision related application. Generally, CNN includes the following processes: convolution, pooling, activation functions and fully connected layers. The convolution works as a feature extractor, while pooling is used for down sampling. The activation functions add nonlinearity to the algorithm to handle complex features. The fully connected layers are used to classify the features. A training dataset is required to train the CNN. Back propagation technique is used to train the algorithm to reduce the classification error. Shridhar provided a guidance for deep learning algorithms and applications [12].

CNN architecture can vary by varying the number of layers, neurons, activation functions and the order of computation elements. The architecture varies based on the application and the available training data. There are many architectures have been introduced in the past few years. Alex et al. introduced Alexnet [13]. This neural network has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully connected layers with a final 1000-way softmax. VGG-16 network introduced with 138 million parameters and more layer depth than Alexnet [14]. Googlenet and Resnet are also examples of CNN architectures [15], [16]. Khan et al. [17] provided a survey of the recent architectures of deep convolutional neural networks.

The captured images by vehicle's front camera includes many objects. In autonomous driving and ADAS applications, Objects like vehicles and pedestrians are the objects of interest. Other objects are considered as background and shouldn't be detected, such as buildings and trees. There are many regional CNN techniques developed to classify the objects of interest in images.

The basic approach for selective search CNN is to scan the whole image by bounding boxes and pass it to the CNN to classify them. The disadvantage of this approach is the large number of bounding boxes to be classified, which makes it a challenge for real time implementation and increase the chances for false positives. Ross et al. [18] proposed a method that uses selective search to extract just 2000 regions from the image and they are called region proposals. Therefore, instead of trying to classify a huge number of regions, the CNN just work with 2000 regions. These 2000 region proposals are generated using the selective search algorithm. This algorithm is called regional-CNN. Fig. 1 shows the system overview of

R-CNN as proposed in [18]. 2000 candidates are still a huge number of region proposals for real time application.

Ross [19] proposed a newer version of R-CNN to improve the algorithm processing time, he called the algorithm Fast R-CNN. This approach is similar to the R-CNN algorithm. But, instead of feeding the region proposals to the CNN, the whole image is fed to the CNN to generate a convolutional feature map. From the convolutional feature map, it identifies the region of proposals and warp them into squares. Fast R-CNN improved algorithm runtime compared to R-CNN. Even though, the algorithm is still time consuming and slow.

Shaoqing Ren et al. [20] came up with an object detection algorithm that eliminates the selective search algorithm and lets the network learn the region proposals In this approach Regional Proposal Network (RPN) is used to generate region proposals, then, these proposals are fed to CNN for object classification. Fig. 2 shows the block diagram of the Faster RCNN block diagram. Mask R-CNN is an improved version of Faster R-CNN that reduces the processing time of the algorithm [21].

All the previous mentioned algorithms use regions to locate objects in images, the algorithm doesn't look to the complete image. You Only Look Once (YOLO) algorithm looks to the complete image one time by passing the whole image to the CNN. YOLO showed a massive improvement in reducing the algorithm computation and it is faster than the other regional CNN. The next section explains YOLO architecture and our detection system implementation.

III. YOLO FOR OBJECT DETECTION

Yolo algorithm was proposed for the first time in 2016 by Redmon et al. [22]. Then a second version was proposed by the same author with better, faster and stronger performance [23]. Redmon and Ali proposed the third and the latest version of Yolo, and they called Yolov3 [24]. In this research, Yolov3 was used to implement road object detection system.

Yolov3 network proposed a CNN with 53 layers, so the network is called Dearknet-53. Fig. 3 shows the architecture of Darknet-53. The convolutional layers are followed by batch normalization layer and Leaky ReLU activation. No form of pooling is used, and a convolutional layer with stride 2 is used to down sample the feature maps.

Yolo algorithm works as following:

- The input image is divided to SxS cells
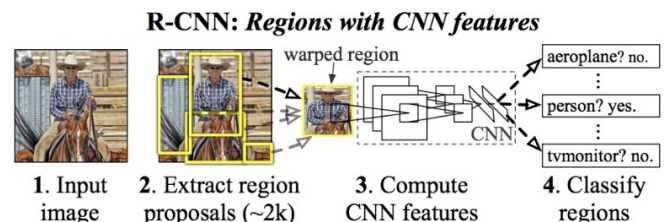
Fig. 1. R-CNN System Overview [18].
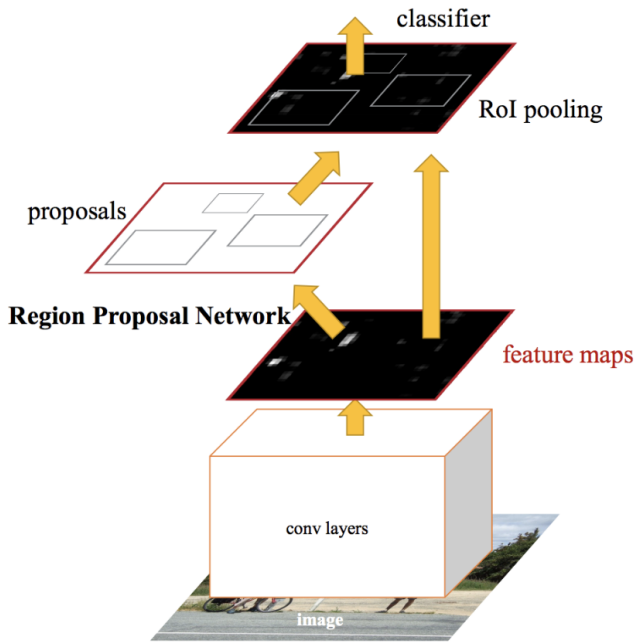
Fig. 2. Faster R-CNN System Block Diagram [20]

- B bounding boxes are generated from each cell

- Each bounding box is described by a vector, the

- vector includes Tx, Ty, Th, Tw, Pc, C

Where:
Tx, Ty: the top left corner of the bounding box coordinates
Th, Tw: the height and the width of the bounding box
Pc: Object score, between 0 and 1
C: class scores vector

In our implementation, the input image is resized to 416x416x3. And divided to 13x13 cells, each cell is 32x32 pixels. 9 bounding boxes are generated from each box. The scoring of each box takes a value from 0 to 1. It represents the probability that the bounding box includes an object. C is 4x1 vector, and it represents the prediction of a bounding box as one of the following classes: pedestrian, truck, car or cyclist. The sum of the prediction should add to 1.

Each cell produces a vector with size of B*(5+C). In our implementation the size of the vector is equal to 9*(5+4) = 81. And the output of the CNN per image equals to 13*13*81 = 13689.

For example, if a bounding is given by the following vector: [5, 6, 50, 70, 0.5, 0.05, 0.7, 0.05, 0.2], it means the top left corner of the bounding box is located at (5,6) with height of 50 pixels and width of 70 pixels. The scoring of the object is 0.5, which means the bounding box includes an object with probability of 0.5. The last four elements represent the object class, it shows that the object classified as pedestrian, truck, car and cyclist with probabilities of 0.05, 0.7, 0.05 and 0.2, respectively.

Fig. 3. Darknet-53 Network Architecture [24]

TABLE I. YOLOV3 IMPLEMENTATION PARAMETERS

Parameters	Value
Input image size	0.005
Input image size	416X416X3
Number of cells per image	13x13
Number of bounding boxes per cell	9
Classes	[Pedestrian, Truck, Car, Cyclist]
Classification threshold	0.5
Non-Maximum suppression overlapping threshold	0.5

The class confidence is given by the product of the object score and the maximum element of vector C. In the previous example, the class confidence equals 0.5*0.7 = 0.35. The classification of object is done based on a fixed threshold value. The non-maximum suppression algorithm is used to eliminate the overlapping detection boxes for the same object. Table I summarizes the parameters of Yolo implementation in our algorithm.

IV. Kitti Dataset and Algorithm Training

A labeled dataset is needed to train the Yolov3 algorithm. There are many datasets available online that can be used for training and testing such as COCO dataset [25], Pascal VOC dataset [26] and google open images V5 dataset [27]. However, Kitti dataset is collected by a vehicle equipped with dash camera and other sensors for autonomous driving testing and benchmarking [28]. This makes it a perfect fit for our detection system as we aim to detect road objects. Kitti dataset consists of 7481 training images with seven labeled classes: cars, van, tram, trucks, pedestrian, person sitting and cyclist. Fig. 4 shows some samples of Kitti dataset.

In our implementation, the dataset labels were reorganized to fit in our four classes. Van, tram and cars are considered as one class, it is called cars. Sitting person and pedestrian are merged, the class named as pedestrian. Trucks and cyclists taken from the dataset without any change.

Kitti dataset labeling format is incompatible with Yolov3. Labels were formatted to match Yolov3 algorithm. The label should be constructed as following: [P1, P2, P3, P4, Class], where
(P1, P2) the top left corner of the bounding box
(P3, P4) the bottom right corner of the bounding box
Class: the bounding box class, it takes value between 0 to 3. It means Pedestrian, Truck, Car, Cyclist, respectively.

The pre-trained weights of Yolov3 with COCO dataset were uploaded as initial weights for the model. The model was trained in two steps. The first step is the frozen training, where the initial layer weights were frozen, and the end layer weights were trained. The epoch is set to 25 and the batch to 32. The second step was the unfrozen training for fine tuning. The epoch is set to 25 and the batch to 4. Adam optimizer is used in the training model [29]. Fig. 5 shows the training loss with epoch value.

V. Detection Results

The algorithm is tested using 300 images from Kitti test dataset. The detection was analyzed by counting the true positives, true negatives and false positives. The scoring threshold for classification was set 0.5, and the non-maximum suppression for the detection boxes overlapping is set 0.5. Fig. 6 shows some samples of the detection in our implementation.

Two concerns can be made about Kitti test dataset. The first concern is the test images are not labeled, so detection results can't be automated and have to be done manually. The second concern is cars are the dominant object in the dataset and the dataset doesn't have enough samples for pedestrians, cyclists and trucks.

Table II summarizes the detection results for all the classes. Results show that 615 objects are classified correctly, while 111 objects were miss classified and 88 objects were false positives. It is noticed that false negative detection was high for pedestrians and cyclists. For better understanding of the results, the precision and recall were calculated for each class. Precision and recall are given in the following equations:

$$Precision = True positives/(True positives + False positives) \tag{1}$$

$$Recall = True positives/(True positives + False negatives) \tag{2}$$

Table III shows the precision and recall for each class. The precision is higher than 98% for all the classes except for cyclists, it is 88.23%. So, we can conclude that detection accuracy is very high for all the classes. There is a significant drop in the recall values for pedestrians and cyclists because of the high values for the false negatives. Due to the small
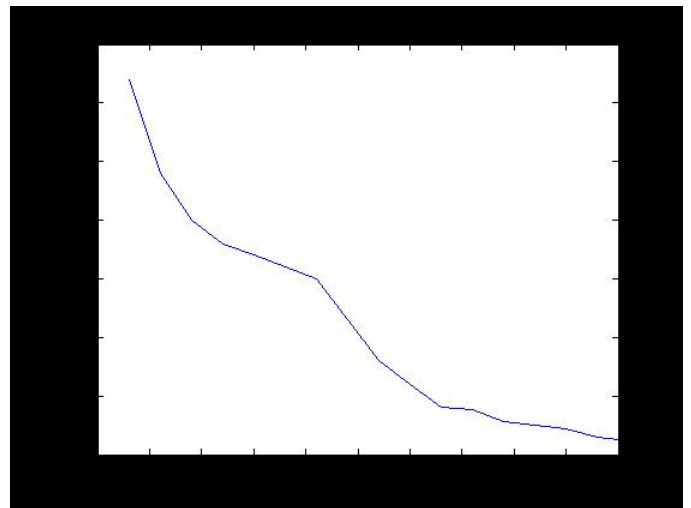
Fig. 5. Darknet-53 Training Loss with Epoch using Kitti Dataset

TABLE II. Detection results for Kitti dataset using Yolov3

Objects	True Positives	False Positives	False Negatives
Cars	559	6	72
Trucks	3	0	3
Pedestrians	38	0	32
Cyclists	15	2	4
Total	615	8	111

Fig. 4. Image Samples from Kitti Training Dataset
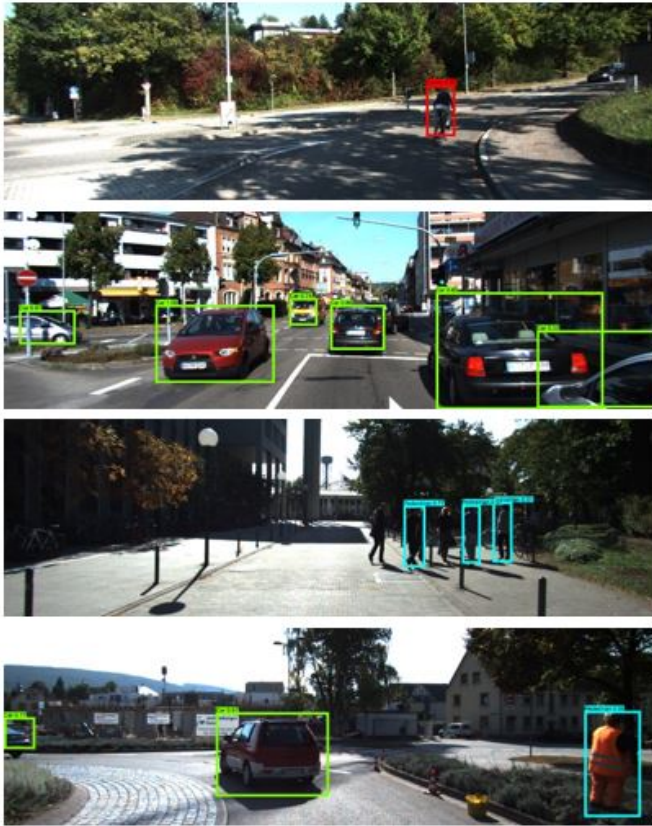
Fig. 6. Detection Results Image Samples

TABLE III. PRECISION AND RECALL VALUES FOR THE CLASSES

Objects	Precision	Recall
Cars	98.9%	88.59%
Trucks	100%	50%
Pedestrians	100%	54.28%
Cyclists	88.23%	78.9%
Total	98.7%	84.7%

samples of trucks available in the test dataset, we can't judge the algorithm performance in trucks detection.

The summary we can make from the above results is that the algorithm showed very excellent detection accuracy for cars. It also showed a high precision for pedestrians and cyclists but with low recall values due to the high number of false negatives. That means the algorithm has high miss detection rate for small objects like pedestrians and cyclists compared to larger objects like cars.

VI. CONCLUSION

Convolutional neural networks are a widely used algorithm for object detection and classifications. This research implemented Yolov3 algorithm which uses Darknet-53 CNN for road object detection. Cars, trucks, pedestrians and cyclists

are the algorithm detection classes. Kitti dataset is used for the algorithm training and testing.

The paper provided a quick review for the different approaches for road object detection and presented the recent CNN architectures and methods for object detection. The Darknet-53 architecture of Yolov3 is explained and the algorithm hyper-parameters were discussed. Algorithm training steps and parameters like epoch and batch were presented.

Detection results are displayed by counting true positives, false positives and false negatives per each class. Precision and recall also calculated for each class. The algorithm showed a very good detection accuracy for cars. Pedestrian and cyclist detection showed more false negatives than cars. The test dataset includes only six trucks; therefore, no solid conclusion can be made about truck detection. A conclusion we can make from the detection results is that the algorithm produces more false negatives in small objects detection.

A suggested solution to improve detection accuracy for pedestrians and cyclists is to reduce the cell size. In our implementation the cell size is 32x32 pixels. Smaller cells should improve detection for the small objects, in the other hand, it increases the processing time of the algorithm. Another solution to reduce the false negatives is to increase the epoch from 25 to a higher number in the training.

Kitti dataset have enough trained labeled images to achieve good performance. However, cars are the dominant object in the dataset, and the test images are not labeled. Which makes detection result analysis more difficult. There are other open source datasets that can be used in road object detection such as Waymo open dataset and Bekerly deep drive.

REFERENCES

[1] World Health Organization (WHO). "Global status report on road safety 2018 (2018)." Geneva (Switzerland): WHO (2019).

[2] Ivar, Bohlin Nils. "Safety belt." U.S. Patent 3,043,625, issued July 10, 1962.

[3] Ziebinski, Adam, Rafal Cupek, Damian Grzechca, and Lukas Chruszczyk. "Review of advanced driver assistance systems (ADAS)." In AIP Conference Proceedings, vol. 1906, no. 1, p. 120002. AIP Publishing LLC, 2017.

[4] Shen, X. L., and D. C. Tseng. "Vision detection of lanes and vehicles for advanced safety vehicles." National Central University, Chung-li (2007).

[5] Davis, James W., and Mark A. Keck. "A two-stage template approach to person detection in thermal imagery." In 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1, vol. 1, pp. 364-369. IEEE, 2005.

[6] Nath, Rajiv Kumar, and Swapan Kumar Deb. "On road vehicle/object detection and tracking using template." Indian Journal of Computer Science and Engineering 1, no. 2 (2010): 98-107.

[7] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, pp. 886-893. IEEE, 2005.

[8] Tsai, Luo-Wei, Jun-Wei Hsieh, and Kuo-Chin Fan. "Vehicle detection using normalized color and edge map." IEEE transactions on Image Processing 16, no. 3 (2007): 850-864.

[9] Li, Zhijiang, Chuan Dong, Ling Zheng, and Long Liu. "Traffic signs detection based on haar-like features and adaboost classifier." In IC-TIS 2013: Improving Multimodal Transportation Systems-Information, Safety, and Integration, pp. 1128-1135. 2013.

[10] Gerónimo, D. and López, A.M., 2014. Vision-based pedestrian protection systems for intelligent vehicles (pp. 1-114). New York, NY, USA: springer.

[11] Sivaraman, Sayanan, and Mohan M. Trivedi. "A review of recent developments in vision-based vehicle detection." In 2013 IEEE Intelligent Vehicles Symposium (IV), pp. 310-315. IEEE, 2013.

[12] Shridhar, A. "A beginner's guide to deep learning." (2017).

[13] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pp. 1097-1105. 2012.

[14] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[15] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.

[16] 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.

[17] Khan, Asifullah, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. "A survey of the recent architectures of deep convolutional neural networks." arXiv preprint arXiv:1901.06032 (2019).

[18] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587. 2014.

[19] Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

[20] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In Advances in neural information processing systems, pp. 91-99. 2015.

[21] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969. 2017.

[22] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.

[23] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271. 2017.

[24] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).

[25] 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.

[26] Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes (voc) challenge." International journal of computer vision 88, no. 2 (2010): 303-338.

[27] Open images dataset V5, www.storage.googleapis.com/ openimages/web/visualizer/index.html?set=train&type=detection&c=%2Fm%2F01j51

[28] Geiger, Andreas, Philip Lenz, Christoph Stiller, and Raquel Urtasun. "Vision meets robotics: The kitti dataset." The International Journal of Robotics Research 32, no. 11 (2013): 1231-1237.

[29] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

Predicting Breast Cancer via Supervised Machine Learning Methods on Class Imbalanced Data

Keerthana Rajendran[1], Manoj Jayabalan[2], Vinesh Thiruchelvam[3]

School of Computing, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia[1, 2, 3]
Faculty of Engineering and Technology, Liverpool John Moores University, Liverpool, UK[2]

Abstract—**A widespread global health concern among women is the incidence of the second most leading cause of fatality which is breast cancer. Predicting the occurrence of breast cancer based on the risk factors will pave the way to an early diagnosis and an efficient treatment in a quicker time. Although there are many predictive models developed for breast cancer in the past, most of these models are generated from highly imbalanced data. The imbalanced data is usually biased towards the majority class but in cancer diagnosis, it is crucial to diagnose the patients with cancer correctly which are oftentimes the minority class. This study attempts to apply three different class balancing techniques namely oversampling (Synthetic Minority Oversampling Technique (SMOTE)), undersampling (SpreadSubsample) and a hybrid method (SMOTE and SpreadSubsample) on the Breast Cancer Surveillance Consortium (BCSC) dataset before constructing the supervised learning methods. The algorithms employed in this study include Naïve Bayes, Bayesian Network, Random Forest and Decision Tree (C4.5). The balancing method which yields the best performance across all the four classifiers were tested using the validation data to determine the final predictive model. The performances of the classifiers were evaluated using a Receiver Operating Characteristic (ROC) curve, sensitivity, and specificity.**

Keywords—Breast cancer; class imbalance; diagnosis; bayesian network

I. INTRODUCTION

The World Health Organization reported in 2018 that there were 627,000 deaths worldwide due to breast cancer [1]. Breast cancer is the second most common cancer death among women, especially in developing countries [2]. This cancer type accounts for 25% of all cancers among women and affects 10% of women globally at some stage of their life [3]. This is a more common issue in developing countries where the mortality rate is greater due to the prohibitive cost incurred for extensive diagnostic tests and treatments required to treat breast cancer completely [4].

American Cancer Society statistics exhibited that there will be about 252,710 new patients with invasive breast cancer and 63,410 patients with in situ breast carcinoma that are expected to be diagnosed among US women in 2017 [5].

The clinicians must check the stage of breast cancer before conducting further assessments on the patients as this step is vital for starting the treatment process and to allow prognosis of the time of recurrence of cancer. However, the multitude of diagnoses carried out to assess the cancer stage require an extended period for the clinicians to obtain medical results.

This period of waiting can cause deterioration of cancer where it will be too late for the patients to acquire any complete treatments. Researchers have suggested the involvement of an intelligent decision support system that can identify the cancer types, which will benefit both the patients and clinicians in aspects of treatment options and expenditures incurred [6].

The involvement of data mining techniques in predicting breast cancer based on the patterns and relationships found among the breast cancer risk factors reduce diagnosis time by physicians and cost [7]. Thus, the survival rate for breast cancer can be increased immensely with diagnosis and treatment at an early stage [8].

To predict the susceptibility to breast cancer depends on breast cancer risk parameters [9]. The risk factors of breast cancer include non-preventable factors such as gender, age and family history of cancer, and preventable factors such as body mass index (BMI) and hormone replacement therapy. Other risk factors include menopause, delayed pregnancy, race, radiation therapy before age 30 and high bone density. These abovementioned risk factors are included as part of this study. Genetic risk factors and lifestyle habits (smoking and alcohol consumption) which are also causatives of breast cancer are not included in this study.

As the chances of survival differ largely by breast cancer stages, the earliest diagnosis will improve the rate of survival greatly. Women who were diagnosed at the early, non-invasive stage will have better chances of survival than those diagnosed at the later invasive stages. It is crucial for clinicians to diagnose women who have breast cancer accurately and prevent false positive results [8]. Therefore, the purpose of this study is to develop a predictive model on the breast cancer occurrence for women by applying class balancing techniques on breast cancer risk factors data to cater a fair decision support system for medical practitioners to diagnose the incidence of breast cancer accurately and enhance the survivability rate of patients.

This study on breast cancer classification using BCSC dataset yielded two contributions. Class balancing methods on the imbalanced BCSC dataset were introduced in this study as the problem of class imbalance in BCSC dataset was not addressed in existing studies. Thus, several balancing methods were proposed in this study and the hybrid balancing method achieved greater performance across the proposed classifiers. The breast cancer predictive model developed using Bayesian Network was rarely explored in previous breast cancer studies and in this study, this classifier proved to achieve the highest

accuracy when compared to other works done using BCSC dataset. Thus far, the Bayesian Network model can be used as an effective model to predict the breast cancer occurrence based on the risk factors available in the BCSC data.

II. RELATED WORKS

The risk prediction model employing logistic regression on the BSCS dataset, which consisted of 2.4 million records of screening mammograms and breast cancer associated risk factors [10]. The risk models were built on two folds with the menopause status and four risk factors for premenopausal women and ten risk factors for postmenopausal women in second fold were found to be significant in the respective models. Compared to Gail's model that was developed in the late 1980s, this model enhanced the prediction of high-risk women through the addition of two more attributes, which were breast density and hormone replacement therapy [11]. The study reported a ROC of 0.631 for premenopausal women and 0.624 for postmenopausal women.

Another study by [12] was done to identify the factors for disparities in breast cancer outcomes between racial and ethnic groups. The prospective cohort study done showed that African-American women had a higher relative risk of advanced breast tumor compared to white women as African-American women had a less frequent mammographic screening. In [13] developed a breast cancer prediction model that includes breast density as an important risk factor. The researchers used BCSC mammography data where the proportional hazards model was employed to predict the hazard ratios for each BI-RADS breast density category in a 5-year follow-up cohort study. The model was validated using 5-fold cross-validation. The results showed that the average c-statistic was 0.6576 where there was slight discrimination between women who develop breast cancer and those who do not. This risk prediction model can assess 5-year risk for invasive breast cancer depending upon breast density and calibrate with common races and ethnic groups in the United States.

One study employed k-NN algorithm to develop a statistical risk score using four factors such as breast density, age, breast procedure and a number of first degree relatives which were based on the domain expert advice [14]. The area under the ROC was reported as 0.642, which suggested a better model compared to Barlow's logistic regression models [10].

Another study focused on comparative modeling to determine the threshold relative risks at which the harm-benefit ratio of screening women at two different age groups [15]. The authors used four microsimulation models on the film and digital mammography data obtained from BCSC. The results showed that the harm-benefit ratios for women aged 40-49 years with a two-fold elevated risk of breast cancer were similar to that for average-risked women aged 50-74 years for biennial screening mammography. The threshold relative risks were reported to be higher for annual screening using digital mammography, but the harm-benefit ratios were greater for film mammography as they have reduced the false-positive rate.

In [16] used 117,136 diagnostic mammograms pooled from six mammography registries under BCSC to construct logistic regression model to determine the adjusted effect on sensitivity, false positive rates, and cancer detection rates. Patient profile and mammography results were used as determinants in the model. The authors postulated that diagnostic interpretive volume was a crucial factor in considering the thresholds for abnormal diagnostic mammograms. Another study focused on generating an approximation to the logistic regression score function using four different algorithms, namely, ApproxMLE, W.ApproxMLE, WGD and WSGD [17]. These algorithms were applied to BCSC and record linkage datasets. The results showed that ApproxMLE method had excellent performance in aspects of accuracy, time scalability and parallel efficiency. This algorithm had an area under ROC of 0.92 and 3.24 minutes as execution time.

A study by [18] used the BCSC dataset in proposing a method to estimate the rate of missing values due to incomplete data in latent class regression. Two models, one without adjustment of mammography history and the other with the adjustment, were developed. Three approaches, namely maximum likelihood (ML) using the Expectation Maximization (EM) algorithm, multiple imputations (MI), and the proposed method of two-stage MI were compared in terms of the regression coefficient for both the models. The results highlighted that the proposed two-stage MI is better than the EM algorithm and standard MI as it allows for further separation of missing information rates into two parts.

In [19] introduced an adaptive online learning framework which integrated supervised learning (SL) and reinforcement learning (RL) models for clinical breast cancer diagnosis. Three machine learning algorithms such as linear regression, logistic regression, and neural network were employed on BCSC and WBC datasets. This framework had the leverage of gaining high diagnosis accuracy in real-time and reducing the amount of diagnosis required for efficient treatment. Logistic regression was found to achieve optimal performance rapidly. Overall, the SL model was reported to attain accurate risk assessment of breast cancer from incremental features and sequential data while the RL model catered better decision-making of clinical measurements. One study applied association rule mining with feature selection on three breast cancer datasets including BCSC dataset [20]. Syntax and dimension reduction constraints were applied to prune the association rules generated from the apriori algorithm. This resulted in a reduction of the feature subsets by more than 50%. The models were then validated using SVM and the results showed that this approach yielded a classification accuracy of approximately 98% for the BCSC dataset.

Most of the research papers that have published on the predictive model for breast cancer have shown relatively high prediction accuracies [7], [8], [21]. However, a widespread problem in medical data is a class imbalance, which was failed to be addressed by any of these previous papers. In case of breast cancer, most of the previous works done on breast cancer have employed datasets with extremely uneven distribution of the class labels, such as non-cancerous (97%) and cancerous (3%) in the BCSC dataset or survival (91%)

and non-survival (9%) of cancer in the SEER dataset [22] or 66% benign and 36% malignant in the WBC dataset [8]. Thus, the results are likely to be biased towards the majority class, which is non-cancerous or survival or benign group, even if the prediction accuracies were high. Most of the cancer diagnosis need pivotal information on the accuracy and false positive rates in the prediction of the cancerous cases. Development of a prediction model using a class-balanced data will cater to a more affirmative decision-making process during a breast cancer diagnosis.

III. METHODOLOGY AND TECHNIQUES

The methodology deployed involves key processes such as the selection of target data, pre-processing the chosen data, transforming the data into a structured and comprehensible format, balancing the dataset, implementing supervised learning techniques and evaluating the machine learning performance using evaluation measures. These steps ultimately lead to knowledge extraction from the target dataset where new insights and ideas can be developed to assist in enhancing business operations or in this case, aid in early diagnosis and prediction of diseases such as breast cancer.

A. Dataset Selection

The data for this study was obtained from the BCSC Data Resource [23]. The dataset comprises of information on women with breast cancer in the age range of 35 years and above obtained from seven mammography registries of the BCSC-Carolina Mammography Registry, Colorado Mammography Project, Group Health Cooperative's Breast Cancer Surveillance Project, New Hampshire Mammography Network, New Mexico Mammography Project, San Francisco Mammography Registry, and the Vermont Breast Cancer Surveillance System in the United States from 1996 to 2002.

The dataset consists of 280,660 screening mammograms (known as "index mammogram") of the women. The data on the variables of interest were gathered via questionnaires given to women when they were present for their mammogram and through the radiologist who assessed the mammogram results at the screening facility. Besides this, cancer data and pathology registry were merged into the mammography data, thus adding on to the related variables of breast cancer [10]. The dataset was anonymized to preserve the confidentiality of the patients, the mammography registries and radiology facilities. Other identifiers such as the origin of the data, dates, patient identifiers and the index screening mammogram assessment results are not included in the data, refer Table I.

TABLE I. DATASET SUMMARY

No.	Variable Name	Coded values
1.	menopaus	Indicates the stage of menopause of each patient. 0 = Premenopausal; 1 = Postmenopausal or age is more than 55; 9 = Missing or unknown
2.	agegrp	Indicates the age group (in years) that the patient belongs to. Kindly note that the code value 9 in this variable represents an age category, instead of a missing value. There is no missing value found in this variable. 1 = 35-39; 2 = 40-44; 3 = 45-49; 4 = 50-54; 5 = 55-59; 6 = 60-64; 7 = 65-69; 8 = 70-74; 9 = 75-79; 10 = 80-84
3.	density	Indicates the patient's breast density based on the findings from Breast Imaging Reporting and Data System (BI-RADS) mammogram screening. 1 = Almost entirely fat; 2 = Scattered fibroglandular densities; 3 = Heterogeneously dense; 4 = Extremely dense; 9 = Unknown or different measurement system
4.	race	Indicates the ethnic background or race of the patient. 1 = White; 2 = Asian/Pacific Islander; 3 = Black; 4 = Native American; 5 = Other/mixed; 9 = Missing or unknown
5.	Hispanic	Indicates whether the patient has Hispanic background. 0 = No; 1 = Yes; 9 = Missing or unknown
6.	bmi	Indicates the Body Mass Index (BMI) of each patient under study. 1 = 10-24.99; 2 = 25-29.99; 3 = 30-34.99; 4 = 35 or more; 9 = Missing or unknown
7.	agefirst	Indicates the age of the patient when she had her first birth. 0 = Age less than 30; 1 = Age 30 or greater; 2 = Nulliparous (has not borne an offspring); 9 = Missing or unknown
8.	nrelbc	Indicates the number of first degree relatives of the patient with breast cancer. 0 = Zero; 1= One; 2 = Two or more; 9 = Missing or unknown
9.	brstproc	Indicates the presence of any previous breast procedure on the patient. 0 = No; 1 = Yes; 9 = Missing or unknown
10.	lastmamm	Indicates the outcome of the patient's last mammogram before the index mammogram. 0 = Negative; 1 = False positive; 9 = Missing or unknown
11.	surgmeno	Indicates whether the patient underwent surgical or natural menopause. 0 = Natural; 1 = Surgical; 9 = Missing or unknown or not menopausal (menopaus = 0 or menopaus = 9)
12.	hrt	Indicates whether the patient has undergone any current hormone replacement therapy. 0 = No; 1 = Yes; 9 = Missing or unknown or not menopausal (menopaus = 0 or menopaus = 9)
13.	invasive	Indicates the results for the diagnosis of invasive breast cancer of the patient within one year of the index screening mammogram. 0 = No; 1 = Yes
14.	cancer	Indicates the results for the diagnosis of invasive or ductal carcinoma in situ breast cancer of the patient within one year of the index screening mammogram. 0 = No; 1 = Yes

B. Data Pre-Processing and Transformation

In the stage of pre-processing, it is essential to eliminate any missing values, noise and other anomalies in the selected data. Any inconsistency in the chosen data, especially disease-related data may lead to unreliable results or misdiagnosis of test data, which could be fatal if the model is implemented in real-life situations [24]. One of the steps of pre-processing is the elimination of unrelated variables, as these variables are not required to meet the goal of the study. Besides that, missing values or anomalies occur due to lack of information and unprecise measurement values leading to inadequate accuracy and a greater percentage of error in the process of data evaluation. For handling missing values which is very common in cancer datasets, imputation have to perform before implementing the model [25]. The missing values for the nominal and numerical attributes in the dataset with the modes and means from the training data. Since all the variables were identified as the nominal (categorical) type, modes, which are values with the highest frequency, from the training data was used to impute the missing values.

For further processing, the data must be transformed into an appropriate format that is readable and compatible with the data mining techniques employed on the dataset [26], [27]. The transformation such as numerical to nominal conversion is done to cater to the requirements of distinct types of data mining techniques.

C. Class Balancing

The imbalance is a problem that is very commonly found in disease-related datasets, such as the breast cancer dataset used in this study, where the class with a greater number of instances is known as the majority class whereas the one with comparatively less number of instances is known as the minority class. In a scenario where the imbalanced dataset is used, the classifiers tend to favor the majority class, thus exhibiting very weak classification rates on the minority class. There is also a possibility that the classifiers predict all as the majority class and disregard the minority class. This is a very common scenario in medical datasets where the patient with the disease tends to be the minority class. Therefore, a good sampling technique is required for medical datasets. To solve the problem of class imbalance, various sampling techniques have been introduced which include undersampling, oversampling and a combination of both. Sampling strategies are introduced to overcome the class imbalance issue through the removal of some data from the majority class (undersampling) or the addition of some artificially synthesized or replicated data to the minority class (oversampling) [28]–[30].

To build a good prediction model from the training set, the data must be well-balanced. But, the class labels of the target variable, cancer in the breast cancer dataset used in this study are not balanced. This may result in a mediocre performance of the classifiers on the minority class label, which is the Yes label, especially when the data is extremely imbalanced with 97.2% of No and 2.8% of Yes. The key reason behind this is because the classifiers neglect the relative distribution of each class, but they tend to focus on optimizing the overall precision [28].

Oversampling methods multiply the number of members in the minority class in the training group. A benefit of oversampling is that there is no loss of information from the original training dataset as all the observations from the majority and minority classes are retained. The disadvantage of this technique is that it may take longer training time and result in over-fitting since there is a significant increase in the size of the training set. A well-known oversampling technique known as Synthetic Minority Oversampling Technique (SMOTE), is used to oversample the minority class by creating synthetic instances to replicate the minority classes and increase their number of instances in the training set [31]. These synthetic instances are produced by considering two key parameters which are the number of instances (n) and the nearest neighbors (k). As the new minority instances are generated by interpolating between several minority instances that lie together, the problem of overfitting in prevented [32].

On the other hand, undersampling also overcomes the class imbalance problem wherein this technique, the number of samples in the majority class is decreased to balance the class distribution between the minority and majority classes. As the size of the training dataset is reduced significantly, the training time taken is lesser and more efficient which serves as an advantage of this technique. A disadvantage of this method is that important information in the training data will be lost. As for undersampling, SpreadSubsample can be used to decrease the number of samples in the majority class from the original dataset so that the class distribution can be balanced with the minority class. The distribution spread can be set as 0, 1 or 10 for different distribution spread of the class values. A value of 1 spreads the class into a uniform distribution, where the class labels are balanced equally.

A combination of oversampling and undersampling in some cases would be a better option as it generates better-defined areas in the data space and avoids over-generalization [29].

The training dataset which will be used in further analysis has the problem of class imbalance with regards to the target variable, cancer. In the imbalanced dataset, the values are more biased towards the value No than Yes. The distribution of No and Yes is largely uneven in the dataset. The dataset which has 180,465 observations consists of 175,339 No values (97.2%) and 5,126 Yes values (2.8%). This major difference between the values in the class variable could lead to the results to favor towards the majority value which is No. This impacts the efficiency of the results negatively which will, in turn, reflect uncertainty in the choice of an ideal prediction model developed from a machine learning algorithm. The training dataset without any class balancing is kept as it is to perform classification tasks on the data.

As one of the methods of class balancing, SMOTE is applied on the training dataset to oversample the minority class label and a separate dataset is created with this oversampling method. The dataset was resampled where the minority class value (which is Yes) is oversampled to increase its number of instances. The default parameter set in WEKA were used which include the creation of 100% SMOTE instances and the nearest neighbor of 5. Upon execution of the

SMOTE filter, the minority class value (which is Yes) has increased to 10,252 instances while the majority remains the same. Now, the target variable, cancer in the class-balanced training set distribution consisted of 10,252 Yes values (5.5%) and 175,339 No values (94.5%). The percentage of the minority class value has doubled via this method of class balancing. The total number of instances in the dataset was observed to be 185,591 instances.

As the second method of class balancing, the technique of undersampling was applied on the training dataset and a training dataset with this method was created. This was done using the SpreadSubsample method where the class distribution spread is set as 1.0 to allow uniform distribution between the two class values (Yes and No). As a result, the majority class value was undersampled to match the minority class value. Upon applying the SpreadSubsample filter, the number of instances in the majority class value (which is No) has reduced to 5,126 which is the same as in the minority class value (which is Yes). Both the class values, Yes and No, have an equal percentage of distribution, which is 5,126 instances (50%) respectively, for the target variable, cancer in the training dataset. The total number of instances in the dataset was observed to be 10,252 instances.

Following this, both the oversampling and undersampling techniques were combined to resample the imbalanced dataset and a training dataset with this method was created. The oversampling method (SMOTE) was applied first followed by the undersampling method (SpreadSubsample) to resample the distribution of the class values in the target variable, cancer. This resulted in the minority class which is the Yes value to be oversampled first, then the majority class which is the No value to be undersampled. The same parameters applied in the previous two methods of class balancing, which are SMOTE and SpreadSubsample, were applied here to obtain a uniform result when comparing these three techniques of balancing. The number of instances in both class values were equal with 10,252 instances (50%) in each of the class value, No and Yes. The total number of instances in the dataset was observed to be 20,504 instances.

D. Data Mining Techniques

Data mining plays as a powerful tool in acquiring valuable information from a large volume of transformed data to aid in quicker decision making and discovery of knowledge. Data mining techniques enable the identification of novel and hidden patterns from the data, facilitate the data experts in uncovering relationships among the data and make statistically-proven and informed decision. Employment of data mining techniques in medical diagnosis such as breast cancer prediction is of utmost importance as it allows the clinicians to make a quick decision on the effective treatment method, early detection, and prediction of cancer and other diseases which in turn improves the survival rate of patients and reduces the cost of treatment. There are various data mining techniques such as classification, clustering, association and regression which are commonly used in medical diagnosis and disease prediction [33].

Classification is a supervised learning technique which is employed to automatically generate a model that can classify or group a class of items, thus the unknown class values of future objects can be predicted. In this two-step process, a training step and validation step are involved to classify new objects. In the first step, the training dataset is used to construct a model to elucidate the characteristics of a group of data classes or notions. As the data classes or notions are predefined where the class which the training sample falls into is given, this process is known as supervised learning. In the second step, the model is implemented to predict the classes of future data or objects [34].

It is a popular technique in studies on cancer prediction and early diagnosis. Some of the classification algorithms which have been employed in previous studies on breast cancer prediction include Naïve Bayes, Bayesian Network, decision tree and association-based classification. In classification, the data is partitioned into two groups which are the training set and testing set. In the training phase, a model is constructed by employing the classifier on the training data and the performance of this model is validated by using the model to predict or assign a class label to the test data which is unlabeled.

1) Bayesian network: Bayesian Network is represented in a graphical model to portray the probabilistic relationships among the variables under study. The Bayesian model assumes conditional independence over the various random variables and this assumption gives information on the probability distribution that is illustrated within the network [35].

Overall, this Bayesian Network is made up of a qualitative element (structural model) that caters a visual depiction of the interactions among the variables, and a quantitative element (a group of local probability distributions) which allows probabilistic inference and mathematically measures the significance of a variable or a group of variables on others. These qualitative and quantitative components establish a singular joint probability distribution on the variables for a problem [36]. From a Bayesian point of view, the classification problem can be described as the challenge of identifying the class with maximum probability given that there is a set of observed variable values. Such probability is viewed as the posterior probability of a class by considering the given set of data and is computed based on the foundation of Bayesian theorem.

This classifier requires a very large training dataset to significantly analyze all the likely combinations and eventually estimate the probability distribution from the training set. This task could be arduous which serves as a disadvantage of this data mining technique [26]. One of the greatest advantages of Bayesian Network is that it permits the compact and economical representation of the joint probability distribution by using conditional independence extensively. The Bayesian Network is preferred as past academic works have shown that this classifier exhibits a strong correlation among the attributes in the patient disease diagnosis. Other than that, the classifier is robust to unrelated variables, noise and confounding factors that are not part of the classification [31]. Bayesian Network has been broadly employed in many medical diagnoses based on previous literature studies,

especially for cancer prediction and recently the use of Bayesian Network classifiers in breast cancer prediction is trending. It is a well-known classifier in medical diagnosis in case of the non-deterministic relationship between the class variable and the attribute set. One of the learning algorithms applied to the Bayesian Network known as K2 has been utilized in breast cancer classification due to its rapid convergence ability.

The structure of the Bayesian Network from the data is learned using search algorithms. Among the several types of learning algorithms such as AD (All Dimensions) Trees, TAN (Tree Augmented Naïve Bayes) and others, K2 is a very well-known algorithm used in cancer classifications which employs a heuristic greedy search method [35]. The K2 algorithm searches the space across all the potential acyclic digraphs by producing many distinct graphs in a heuristic way and based on this, their ability to interpret the data is compared. The apriori ordering of the variables limits the search space by only permitting parents that precede the variables in the ordering. The algorithm begins its search by assuming that each node has an empty set of parents and iteratively adds the parent based on a given ordering that the addition increases the probability (score) of the final structure the most. The algorithm stops the addition of the variables to a parental set when further addition of a parent does not increase the probability and carries on to the next variable present in the ordering. The model building process involves iterative permutations of the ordering and the network that gives the highest probability is selected. Once the structure has been learned, the conditional probabilities of the Bayesian Network are estimated directly from the data using a Simple Estimator method.

2) Random forest: Random Forest is a tree-based method where it creates multiple classifiers and aggregates the outcomes using ensemble learning method to make the predictions. The approach used in such Ensembles of Classifiers is that there is a level of randomness to generate their tree-based components. This technique creates a collection of hundreds to thousands of unpruned classification and regression trees (CART) based on the random selection of records in the original training data. Although Random Forest is derived from the CART technique, it differs from CART based on the non-deterministic growth via a two-level randomization process. Each tree is grown using the bootstrap sample of the training data and explores across a randomly chosen subset of features (input variables) to determine the split at the node level during the tree growth. The random selection of the features reduces the correlation between the trees which enhances the prediction power and gives higher efficiency. The low variance of the forest ensemble is known as the bagging phenomenon [37]. The splitting criterion in the Random Forest technique is based on the Gini measure of impurity where the lowest impurity value is computed at each node for a set of variables.

One of the key features of Random Forest in classification is that it provides the measure of variable importance where it shows the degree of association between a particular feature

and the classification result. To test the trees developed from the bootstrap data, the out-of-bag samples can be used to provide the two by-products which are the unbiased test set error estimate and variable importance measure. Due to the many benefits offered by Random Forest, this method is very popular and preferred for classification tasks such as in breast cancer prediction studies. The advantages of this technique are given as the following:

- It can handle high dimensional data which contains missing values and variables which are continuous, binary and categorical.

- It is robust enough to overcome over-fitting of data, thus does not require pruning of the trees.

- It is a simple, efficient and comprehensible non-parametric method that can be employed on diverse types of datasets.

- It has greater prediction accuracy and better generalization.

3) Decision tree: A decision tree is a supervised technique which applies the reasoning approach to obtain solutions for a given problem. This data mining technique is very flexible and simple which makes it an attractive choice for applications in diverse fields, particularly because it exhibits advice-oriented visualization to make the prediction decision based on the observed outcomes. A decision tree is commonly applied in decision-making processes in the medical field for disease diagnosis such as cancer prediction. The tree-shaped structures in decision tree represent decision sets which are easy to interpret and understand for decision-makers to assess and choose the best course of action based on the risk and benefits for each possible outcome for distinct options [33]. The basic structure of a decision tree is composed of the following elements.

- A root node which does not have any incoming branch but consists of zero or more outgoing branches.

- Internal nodes where each node has one incoming branch and two or more outgoing branches.

- Leaf or terminal nodes, each of which comprises of one incoming branch but no outgoing branches.

Each node represents the attribute in the input attribute space, while each branch in the decision tree represents a condition value for the corresponding node. The non-terminal nodes have attribute test conditions to divide the records based on their distinguishing characteristics which are represented by the branches.

One of the popular classification types in the decision tree, especially in breast cancer prediction and diagnosis, is the C4.5 algorithm which is an extension of the ID3 algorithm [38]. C4.5 generates decision trees from a group of defined training data based on the information entropy concept. This approach utilizes the fact that each variable in the data is involved in decision-making by splitting the records into smaller subsets. Using the normalized information gain

(difference in entropy), C4.5 determines the selection process of an attribute to split the data. The attribute with the greatest normalized information gain is chosen as the decision node. The branch with zero entropy is taken as the leaf node in the decision tree. This algorithm runs recursively on smaller subsets which are non-leaf nodes with non-zero entropy. The splitting process stops when all the samples in a given subset or node fall under the same class. Then a lead node is generated so that the class can be chosen. But in case of lack of information gain from any of the attributes, the C4.5 generates a decision node using the class expected value from the nodes higher up in the tree.

The C4.5 algorithm has a few advantages such as it is easy and simple to construct in a comprehensible format, can be applied on data with discrete and continuous attributes, can handle attributes with missing values and differing costs in the training data and has greater precision due to pruning procedure. The disadvantages of C4.5 classifier include the expensive cost incurred and high computational time.

IV. Experimentation and Results

The experiment was carried out using the Waikato Environment for Knowledge Analysis (WEKA) [15]. Model validation using k-fold (10 folds) cross-validation was applied on the class-imbalanced training data as well as the three class-balanced training sets respectively. The generated classifier model in this study using the training dataset was validated using this 10-fold cross-validation method which is preferred in disease-related analysis, including breast cancer diagnosis and prediction.

The class balancing methods were applied to the training dataset which consists of 180,465 instances. These balancing methods were applied to overcome the issue of the class imbalance of the target variable, cancer. The balancing methods that were applied include SMOTE (for oversampling), SpreadSubsample (for undersampling) distribution spread was set as 1.0 and a combination of SMOTE and SpreadSubsample.

Each of the classifier performance was compared based on their sampling methods and the classifier with best overall performance was chosen as the best prediction model for the breast cancer dataset. The classifiers' performances were assessed based on several evaluation metrics which include the correctly classified instances percentage or the accuracy, ROC, PRC Area, FP Rate, specificity, precision, recall and F-measure. As the breast cancer dataset used in this study is medical data, there are certain evaluation measures which are of key importance in evaluating the prediction model developed from an algorithm. These measures are the accuracy, TP rate (or sensitivity or recall), FP rate, precision, ROC and PRC area.

For an ideal breast cancer prediction model, a greater TP rate indicates that cancer patients are predicted correctly to have cancer. A higher TN rate is also preferred but this measure does not carry as much importance as TP rate. A prediction model needs to detect the presence of a disease correctly and prevent any misdiagnosis. The misleading results due to FN rate and FP rate can be fatal to patients as

cancer is a lethal disease and the earlier the diagnosis, the better are the chances of survival. The lesser the FP rate and FN rate, and the higher the TP rate and TN rate, the better is the performance of the classification model. This is some general criteria for a disease prediction model, but this may vary depending on the dataset and the type of classifiers.

Based on the analysis, it was found that across all the four classifiers there were two sampling methods which showed better measures to evaluate the performance of the classifiers. These methods are without balancing, where the original training set is used, and applying a combination of SMOTE and SpreadSubsample on the training set.

It was observed that within each classifier, without balancing and combination of SMOTE and SpreadSubsample methods show better evaluation for certain measures. In Bayesian Network, Random Forest and Decision Tree C4.5 models, the accuracy, sensitivity, and precision values are greater in without balancing method, but the ROC is higher in the combination method. As a lower FP rate is better for disease diagnosis, the combination (hybrid) method shows a lower FP rate for all the four abovementioned classifiers.

Overall, it can be concluded that although all these four classifiers do perform well without class balancing method, the results produced are likely to be biased as the distribution of the class value is imbalanced. The accuracy, sensitivity, and specificity will be biased towards the majority class value, which in this case is the No cancer class value. Thus, the combination of SMOTE and SpreadSubsample method is a better model that can be used to validate these four classifiers using a validation (test) dataset. This hybrid balancing method has an even spread of the Yes cancer and No cancer class value which will aid in the development of a fair predictive model.

Further performance measures used to evaluate the classifiers were analyzed to determine the optimum classification model for the prediction of breast cancer using the BCSC dataset. The validation dataset which was used to validate the classifiers generated on the training set resulted in the construction of classifiers with similar values of evaluation metrics and there is no enormous difference between the classifiers from the training and validation set. This shows that all the classifiers have performed well upon the evaluation using the test set. But, to determine the best classifier or the most robust one among the four proposed classifiers, some of the common evaluation metrics found in medical diagnosis were compared between these classifiers.

Table II shows Bayesian Network classifiers have the highest accuracy with 99.1%. Random Forest yielded an accuracy of 94.8% which is the lowest. For the Yes class label, Bayesian Network have the lowest FP rate where it was shown that there were 0% of FP that was predicted. It is important to obtain an FP rate as low as possible to avoid the mistake of diagnosing healthy patients as having breast cancer. Bayesian Network classifier portrayed the greatest precision values for both the class labels and weighted average. The average sensitivity for Bayesian Network is given as 99.1% and the ROC is given as 93.7%, where these two measures are the highest across all the classifiers. Overall,

the results show that Bayesian Network can be adopted as the predictive model for this breast cancer study using BCSC data.

The rationale behind selecting Bayesian Network as the best classification model for this study is because this data mining technique have been commonly employed in the diagnosis of cancer and thus, has an evident record of working well as a prediction model for cancer studies. Further, the Bayesian network model is more comprehensible for the human brain as the model can be easily visualized using a graph.

To further prove that the Bayesian Network model yields a better prediction compared to other models employed in previous literature on the same BCSC dataset, the Table III the comparison of the evaluation measures between the models in the previous literature and this study. Bayesian Network model has yielded the highest accuracy and ROC compared to the other models. Thus, the Bayesian Network acts as a better predictive model in the classification of breast cancer occurrence based on the related risk factors. Besides that one major difference with the previous studies were the use of class balancing techniques. None of these studies addressed the class imbalance issue on the BCSC and employed several balancing techniques, which were done in this study. The approach of hybrid balancing technique with Bayesian Network produced a better prediction model, as was shown in this study.

TABLE II. PERFORMANCE EVALUATION METRICS

Classifier	Class label	Performance evaluation metrics				
		Accuracy	FP Rate	Precision	Sensitivity or Recall	ROC
Naïve Bayes	No		0.219	0.991	1.000	0.937
	Yes	0.991	0.000	1.000	0.781	0.937
	Weighted average		0.210	0.991	0.991	0.937
Bayesian Network	No		0.219	0.991	1.000	0.937
	Yes	0.991	0.000	1.000	0.781	0.937
	Weighted average		0.210	0.991	0.991	0.937
Random Forest	No		0.197	0.991	0.955	0.913
	Yes	0.948	0.045	0.434	0.803	0.913
	Weighted average		0.191	0.968	0.948	0.913
Decision Tree C4.5	No		0.214	0.991	0.993	0.914
	Yes	0.984	0.007	0.832	0.786	0.914
	Weighted average		0.206	0.984	0.984	0.914

TABLE III. COMPARISON WITH EXISTING STUDIES

Previous literature	Predictive model	Evaluation measure	Scope of study
[10]	Logistic regression	ROC = 0.631 (premenopausal); 0.624 (postmenopausal)	Risk prediction model for premenopausal and postmenopausal women
[13]	Proportional hazard models	c-statistic = 0.6576	5-year risk prediction model for invasive breast cancer based on breast density
[14]	k-NN	ROC = 0.642	Statistical risk score using four risk factors
[17]	ApproxMLE algorithm	ROC = 0.92	Approximation to logistic regression score function
[20]	Association rule mining with SVM	Accuracy = 98%	An association rule model with feature selection on the dataset
This study	Bayesian Network	Accuracy = 99.1% ROC = 0.937	Predictive model for breast cancer occurrence depending on risk factors

V. CONCLUSIONS

This study was conducted using the BCSC dataset which consisted of 280,660 screening mammography results and demographic profiles of breast cancer patients who are women aged 35 years and above. The issue of class imbalance in the training dataset was solved using three-class balancing techniques, namely, SMOTE, SpreadSubsample and hybrid of SMOTE and SpreadSubsample. These methods were used to construct the Bayesian Network, Random Forest and Decision Tree C4.5 classification models. When the sampling techniques were compared across each classifier using the performance evaluation metrics, the results showed that the classifiers generated using the hybrid balancing method had

the best performance in terms of false positive rate and area under the ROC. Thus, the best-suited class balancing method for the BCSC dataset was determined to be a hybrid method which was statistically proven to perform well over other sampling techniques.

The results showed that the Bayesian Network generated from the class balanced BCSC data using the hybrid method had greater overall performance in terms of ROC (0.937), sensitivity (78.1%), and False Positive rate (0%) or specificity (100%). This study proves that the Bayesian Network model can serve as a better decision support system for physicians, and as means for early diagnosis and treatment for patients by predicting the occurrence of breast cancer based on the risk factors.

In conclusion, this study proved that the hybrid balancing method with Bayesian Network algorithm achieved the greatest efficiency in predicting the breast cancer occurrence based on the risk factors. With this approach, clinicians can make fair and statistically-proven decisions on diagnosis and treatment options, while breast cancer patients can gain a better understanding of the disease and its risk factors.

Further works can involve feature selection on the BCSC dataset and segmentation of the variables with similar characteristics. A predictive model constructed from feature selection and similar variables may produce a generalized model with a minimum number of risk factors to diagnose. As it is not guaranteed that conclusions from this study could be generalized to other mammography datasets with different properties, it would also be interesting to apply this methodology on other data with features such as shape, location, tumor size or radiation intensity.

REFERENCES

[1] World Health Organisation, "Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018," Int. Agency Res. cancer, no. September, pp. 13–15, 2018.

[2] E. Kharazmi, A. Försti, K. Sundquist, and K. Hemminki, "Survival in familial and non-familial breast cancer by age and stage at diagnosis," Eur. J. Cancer, vol. 52, pp. 10–18, Jan. 2016, doi: 10.1016/J.EJCA.2015.09.015.

[3] S. Kharya, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease," Int. J. Comput. Sci. Eng. Inf. Technol., vol. 2, no. 2, pp. 55–66, 2012, doi: 10.5121/ijcseit.2012.2206.

[4] D. Carvalho, P. R. Pinheiro, and M. C. D. Pinheiro, "A Hybrid Model to Support the Early Diagnosis of Breast Cancer," Procedia Comput. Sci., vol. 91, no. Itqm, pp. 927–934, 2016, doi: 10.1016/j.procs.2016.07.112.

[5] C. E. DeSantis, J. Ma, A. Goding Sauer, L. A. Newman, and A. Jemal, "Breast cancer statistics, 2017, racial disparity in mortality by state," CA. Cancer J. Clin., vol. 67, no. 6, pp. 439–448, 2017, doi: 10.3322/caac.21412.

[6] R. J. Oskouei, N. M. Kor, and S. A. Maleki, "Data mining and medical world: Breast cancers' diagnosis, treatment, prognosis and challenges," Am. J. Cancer Res., vol. 7, no. 3, pp. 610–627, 2017.

[7] J. Majali, R. Niranjan, V. Phatak, and O. Tadakhe, "Data Mining Techniques For Diagnosis And Prognosis Of Cancer," Ijarcce, vol. 4, no. 3, pp. 613–615, 2015, doi: 10.17148/IJARCCE.2015.43147.

[8] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, "Diagnosis of Breast Cancer using Decision Tree Data Mining Technique," Int. J. Comput. Appl., vol. 98, no. 10, pp. 16–24, Jul. 2014, doi: 10.5120/17219-7456.

[9] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, "SVM and SVM ensembles in breast cancer prediction," PLoS One, vol. 12, no. 1, pp. 1–14, 2017, doi: 10.1371/journal.pone.0161501.

[10] W. E. Barlow et al., "Prospective breast cancer risk prediction model for women undergoing screening mammography," J. Natl. Cancer Inst., vol. 98, no. 17, pp. 1204–1214, 2006, doi: 10.1093/jnci/djj331.

[11] M. H. Gail et al., "Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually," JNCI J. Natl. Cancer Inst., vol. 81, no. 24, pp. 1879–1886, Dec. 1989, doi: 10.1093/jnci/81.24.1879.

[12] R. Smith-Bindman et al., "Does Utilization of Screening Mammography Explain Racial and Ethnic Differences in Breast Cancer?Mammography Use and Breast Cancer Rates in Ethnic Minorities," Ann. Intern. Med., vol. 144, no. 8, pp. 541–553, Apr. 2006, doi: 10.7326/0003-4819-144-8-200604180-00004.

[13] J. A. Tice, S. R. Cummings, R. Smith-Bindman, L. Ichikawa, W. E. Barlow, and K. Kerlikowske, "Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model," Ann. Intern. Med., vol. 148, no. 5, pp. 337–347, 2008, doi: 10.7326/0003-4819-148-5-200803040-00004.

[14] E. Gauthier, L. Brisson, P. Lenka, and S. Ragusa, "Breast cancer risk score: a data mining approach to improve readability.," Int. Conf. Data Min., pp. 15–21, 2011.

[15] N. T. van Ravesteyn et al., "Tipping the Balance of Benefits and Harms to Favor Screening Mammography Starting at Age 40 Years: A Comparative Modeling Study of Risk," Ann. Intern. Med., vol. 156, no. 9, pp. 609–617, May 2012, doi: 10.7326/0003-4819-156-9-201205010-00002.

[16] S. Haneuse et al., "Mammographic Interpretive Volume and Diagnostic Mammogram Interpretation Performance in Community Practice," Cancer, vol. 262, no. 1, pp. 69–79, 2012, doi: 10.1148/radiol.11111026/-/DC1.

[17] C. Ngufor and J. Wojtusiak, "Learning from large-scale distributed health data: an approximate logistic regression approach," Proc. ICML 13 Role Mach. Learn. Transform. Healthc., vol. 28, 2013.

[18] L. K. Koegel, T. W. Vernon, R. L. Koegel, B. L. Koegel, and A. W. Paullin, "Improving Social Engagement and Initiations Between Children With Autism Spectrum Disorder and Their Peers in Inclusive Settings," J. Posit. Behav. Interv., vol. 14, no. 4, pp. 220–227, Oct. 2012, doi: 10.1177/1098300712437042.

[19] T. Chu, J. Wang, and J. Chen, "An adaptive online learning framework for practical breast cancer diagnosis," 2016, p. 978524, doi: 10.1117/12.2216550.

[20] M. U. Salma and others, "Reducing the Feature Space Using Constraint-Governed Association Rule Mining," J. Intell. Syst., vol. 26, no. 1, pp. 139–152, 2017.

[21] A. Hazra, S. K. Mandal, and A. Gupta, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms," Int. J. Comput. Appl., vol. 145, no. 2, pp. 975–8887, 2016.

[22] K.-M. Wang, B. Makond, W.-L. Wu, K.-J. Wang, and Y. S. Lin, "Optimal Data Mining Method for Predicting Breast Cancer Survivability," Int. J. Innov. Manag. Inf. Prod., vol. 4, no. 2, pp. 28–33, 2013.

[23] "Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C)," http://www.bcsc-research.org/.

[24] N. Rathore, S. Agarwal, and others, "Predicting the survivability of breast cancer patients using ensemble approach," pp. 459–464, doi: 10.1109/ICICICT.2014.6781326.

[25] H. L. Afshar, M. Ahmadi, M. Roudbari, and F. Sadoughi, "Prediction of breast cancer survival through knowledge discovery in databases," Glob. J. Health Sci., vol. 7, no. 4, p. 392, 2015.

[26] V. Krishnaiah, D. G. Narsimha, and D. N. S. Chandra, "Diagnosis of lung cancer prediction system using data mining classification techniques," Int. J. Comput. Sci. Inf. Technol., vol. 4, no. 1, pp. 39–45, 2013.

[27] H. Sawhney and H. Kaur, "Implementation and Applications of Data Minin in Medical Decision Making Predictions," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 7, 2017.

[28] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," Int. J. Mach. Learn. Comput., vol. 3, no. 2, p. 224, 2013.

[29] R. Longadge, S. S. Dongre, and L. Malik, "Class imbalance problem in data mining: review," Int. J. Comput. Sci. Netw., vol. 2, no. 1, pp. 83–87, 2013, doi: 10.1109/SIU.2013.6531574.

[30] R. Kothandan, "Handling class imbalance problem in miRNA dataset associated with cancer," Bioinformation, vol. 11, no. 1, p. 6, 2015.

[31] R. R. Rao and K. Makkithaya, "Learning from a Class Imbalanced Public Health Dataset: a Cost-based Comparison of Classifier Performance," Int. J. Electr. Comput. Eng., vol. 7, no. 4, p. 2215, 2017.

[32] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 20–29, 2004.

[33] T. M. Fahrudin, I. Syarif, and A. R. Barakbah, "Data Mining Approach for Breast Cancer Patient Recovery," Emit. Int. J. Eng. Technol., vol. 5, no. 1, pp. 36–71, 2017.

[34] S. Kharya and S. Soni, "Weighted naive bayes classifier: A predictive model for breast cancer detection," Int. J. Comput. Appl., vol. 133, no. 9, pp. 32–37, 2016.

[35] H. You and G. Rumbe, "Comparative study of classification techniques on breast cancer FNA biopsy data," IJIMAI, vol. 1, no. 3, pp. 6–13, 2010.

[36] C.-R. Nicandro et al., "Evaluation of the diagnostic power of thermography in breast cancer using bayesian network classifiers," Comput. Math. Methods Med., vol. 2013, 2013.

[37] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," J. Biomed. Sci. Eng., vol. 06, no. 05, pp. 551–560, 2013, doi: 10.4236/jbise.2013.65070.

[38] D. Patel, B. Tanwala, and P. Patel, "Breast Cancer Using Data Mining Techniques," Int. J. Comput. Sci. Eng., vol. 6, no. 7, pp. 1531–1536, 2018, doi: 10.26438/ijcse/v6i7.15311536.

A Novel Fuzzy Clustering Approach for Gene Classification

Meskat Jahan[1], Mahmudul Hasan[2]

Computer Science and Engineering, Comilla University, Cumilla, Bangladesh[1, 2]
Information and Computer Sciences, Saitama University, Japan[2]

Abstract—**Automatic cluster detection is crucial for real-time gene expression data where the quantity of missing values and noise ratio is relatively high. In this paper, algorithms of dynamical determination of the number of cluster and clustering have been proposed without any pre and post clustering assumptions. Proposed fuzzy Meskat-Hasan (MH) clustering provides solutions for sophisticated datasets. MH clustering extracts the hidden information of the unknown datasets. Based on the findings, it determines the number of clusters and performs seed based clustering dynamically. MH Extended K-Means cluster algorithm which is a nonparametric extension of the traditional K-Means algorithm and provides the solution for automatic cluster detection including runtime cluster selection. To ensure the accuracy and optimum partitioning, seven validation techniques were used for cluster evaluation. Four well known datasets were used for validation purposes. In the end, MH clustering and MH Extended K-Means clustering algorithms were found as a triumph over traditional algorithms.**

Keywords—*Meskat-Hasan clustering (MH clustering); MH Extended K-Means clustering; K-Means; fuzzy clustering*

I. INTRODUCTION

Clustering divides the dataset based on data's attributes or characteristics. The fundamental purpose of clustering is to categorize the data based on their distinguishable attributes. For partitioning clustering, there are few established soft and hard versions of algorithms. Popular versions of hardcore clustering are K-Means, K-Medians, K-Modes, Forgy's algorithm and soft clustering are Fuzzy C-Means, Fuzzy K-Means (Fuzzy clustering). The variances of the Fuzzy C-Means, such as Gath-Geva (GG) clustering and Gustafson-Kessel (GK) clustering algorithms are used. Extended versions of the fuzzy clustering algorithm like E-FCM, Extended GK cluster exist. There are some other versions of the algorithm like Fuzzy K-NN algorithm and Fuzzy Local Information C-Means clustering algorithm. GrFPCM select features in the preprocessing step while FPCM and Granular Computing used for outlier detection and features selection [1]. To defeat high – dimensionality the problem, an ant-based algorithm used in the bioinformatics domain which enhanced with the use of FCM and heaps merging heuristic [2]. Gene ontology annotations based GO-FRC algorithm used biological data for gene clustering and this method may assign one gene into multiple clusters [3]. PSO clustering method established on fuzzy point symmetry used for gene expression classification [4]. FWCMR merge the sub clusters to form a final cluster which is implemented on the parallel and distributed environment [5]. WGFCM used entropy based weight vector calculation to

appropriately measure the distance [6]. Immune system behavior based MCSOA used a new fast convergence mechanism for optimum solutions where the number of clusters varies in a certain range [7]. Dynamic Time Wrapping distance technique is useful in shaped based clustering while grouping time series GE data [8]. Fuzzy decision tree algorithm outperforms over classical decision tree algorithm in analyzing cancer GE data [9]. Also, there are techniques for determining the number of clusters like FLAME clustering. These algorithms help to find the behavior of the dataset to reveal the underlying hidden pattern by grouping similar categories of data based on characteristics and most of them need a good initial guess of the number of clusters to perform clustering. Predicting the accurate number of cluster is challenging task. Lots of algorithms are developed but depend on some predefined or prior knowledge. For example, K-NN, K-Means, FCM, etc. are all need robust initial guess.

It has been founded that previously this scenario was solved by applying some post cluster analysis to predict and selecting the number of clusters that require time and cost. Therefore, a method to determine the number of clusters dynamically is required. However, we develop two new algorithms, named Meskat-Hasan clustering (MH clustering) algorithm and MH-Extended K-Means clustering algorithm are proposed for automatic cluster detection, dynamic cluster selection and partitions. Moreover, post cluster enhancement is not required for them. So, they minimize time and cost complexity. Performance of these methods was validated using seven validation criteria Separation Index(S), Partition Coefficient (PC), Dunn's Index(DI), Alternative Dunn Index(ADI), Classification Entropy(CE), Partition Index(SC), Xie and Beni's Index(XB) and comparing results with existing clustering techniques like Fuzzy C-Means, Gath-Geva clustering (GG), Gustafson- Kessel algorithm (GK), K-Means and K- Median. Four different datasets (Wisconsin Breast Cancer, Leukemia, Irises and Motor Cycle [10]) were used to evaluate the performance of the algorithms.

Finally, the proposed algorithms are performed better than other existing literature.

II. LITERATURE REVIEWS

Patrik D'haeseleer [11], described the working principles of gene expression clustering and suggested to use more than one clustering algorithms. According to Jain and Dubes [12], there is no single criterion to define a good clustering algorithm. Clusters are of arbitrary size and shapes in a multidimensional pattern space and clustering quality may be evaluated based on

internal criteria or external criteria. James C. Bezdek, Robert Ehrlich and William Full [13], proposed a program for Fuzzy C-Means (FCM) clustering algorithms to generate prototypes and fuzzy partitions for the numerical datasets. This fuzzy partition is useful for suggesting definite substructure for the raw datasets. Representing the similarity of a point is shared among the nearest clusters with the help of membership function whose values ranges from one to zero, is the idea given by Zadeh [14]. In FMLE algorithm good initial seed points are required because of the exponential distance helps to converge the algorithm to a local optimum rather in a narrow region. Except for this limitation, the FMLE algorithm is better than Gustafson and Kessel algorithm [15]. D. E. Gustafson and W. C. Kessel [15], developed a fuzzy clustering algorithm using fuzzy covariance matrix to prove the argument that in fuzzy clustering, fuzzy covariance has a natural approach. An expression of the interpretation of the membership functions was proposed by Ruspini [16]. This relationship denotes the similarity between samples where a fuzziness parameter is used, whose increasing value trend to indicate the more fuzziness of the clustering process. Jiye Liang, Xingwang Zhao, Deyu Li, Fuyuan Cao and Chuangyin Dang [17], proposed a clustering algorithm for special datasets like mixed datasets containing both numeric and categorical attributes. They presented a mechanism to characterize the data within the cluster and between cluster entropies and to detect the worst cluster in that particular dataset. Kaiser [18] proposed eigenvalues greater than one rule, which is now a commonly used criterion for finding the number of factors. It strongly states that the number of reliable factors is equal to the number of eigenvalues greater than one. As negative eigenvalues have negative reliability, the respective composite score should be reliable. In the internal consistency, it must have some positive reliability. Norman [19] concluded this by suggesting that more reliable components there will be that those are indicated in the eigenvalues greater than unity rules. `The convergence rate of evolutionary clustering methods is high enough than partition clustering methods [20]. This [21] comparative dissection paves the way of choosing the desirable clustering algorithm for some particular dataset.

III. PROPOSED ALGORITHMS

Clustering is an unsupervised technique for categorizing the data elements. Previously, it was not possible to predict the accurate number of a cluster without conducting pre cluster analysis. Proposed techniques have been developed based on the principal component analysis. Then, these techniques were applied to the two types of clustering (Fuzzy and hard-core) algorithms. Finally, validation of the approaches was done and in the next sections, proposed two algorithms based on fuzzy clustering and hard clustering algorithms are shown.

A. Meskat-Hasan Clustering (MH Clustering) Algorithm

MH clustering is a fuzzy approach for data clustering. It is an integrated package of all the tasks to perform clustering including the determination of the number of cluster and clustering. Most of the established clustering algorithm has some kinds of dependency or need some data related prediction for knowing the behavior or characteristics of the dataset. To overcome this, Meskat-Hasan clustering (MH clustering)

dynamically determine the clusters number. MH clustering has the following four steps:

Step-1: Normalize dataset.

Step-2: Extract underlying structure.

Step-3: Run time cluster number determination.

Step-4: Clustering using fuzzy c-means algorithm.

The generalized version of the algorithm is given below:

___

Meskat-Hasan (MH) clustering algorithm pseudocode
___

1. Let, $X = \{x_1, x_2, x_3 \ldots\ldots\ldots x_n\}$ be the dataset
2. Scale the data set,
$$X_{i= \{1....n\}} = \frac{Xi - Xmin}{X\max - Xmin}$$
3. Apply PCA and get variance.
4. Determine num_elements = count(variance)
5. Initialize i=1 and cumulative summation, cum_sum=0.
6. Initialize k=0 and sum=0;
7. For i=1 to num_elements{
$$cum\_sum = cum\_sum + \frac{variance\ (i)}{sum(variance)} * 100;$$
 }
8. Determine σ based on cum_sum.
9. Do {
 Calculate cumulative summation of variance;
 k=k+1;
 }
     Until (cum_sum≤ σ)
 End Do
10. Assign the number of clusters, c = k;
11. Determine the fuzzy membership [3],
$$\mu_{ij}^{(l)} = 1 / \sum_{p=1}^{c}(D_{ij} - D_{ip})^{(2/m-1)};$$
12. For j=1 to c, Calculate the fuzzy centers [3],

$$V_{ij} = (\sum_{i=1}^{n}(1 * \mu_{ij})^{m} x_i) / (\sum_{i=1}^{n}(1 * \mu_{ij}^{m});$$

___
13. Iterate step 11 and line 12 until $\| U^{(l)} - U^{(l-1)} \| < €$
___

where $€ = 1 \times 10^{-6}$ and m=1 to ∞ is the fuzziness parameter.

In steps 1 to 10 determine the desired number of clusters and steps 11 to 13 clustering is performed.

B. MH Extended K-Means Clustering Algorithm

The MH Extended K-Means clustering algorithm is an extension of the K-Means algorithm. In the MH Extended K-Means algorithm, we will apply the techniques of determining the number of the clusters dynamically along with the original k-means algorithm. That means, to implement it in a hardcore clustering process. Main steps of this algorithm are as follows:

Step-1: Scale dataset

Step-2: Extract underlying structure

Step-3: Approaches to determine the cluster number dynamically

Step-4: Perform clustering by K-Means clustering procedure

The generalized version of the algorithm is given below:

MH Extended K-Means Clustering Algorithm Pseudocode

---

1. Let, X = {x$_1$, x$_2$, x$_3$………… x$_n$} be the dataset.

2. Scaling of the dataset,

$$X_{i= \{1....n\}} = \frac{Xi - Xmin}{Xmax - Xmin} ;$$

3. Apply PCA and get variance.

4. Count num_elements= count(variance)

5. Initialize i=1 and cum_sum=0

6. Set k=0 and sum=0;

7. For i=1 to num_elements{

$$cum\_sum = cum\_sum + \frac{variance\ (i)}{sum(variance)} * \mathbf{100};$$

}

8. Determine σ based on cum_sum.

9. Do {Calculate cumulative summation of variance;

k=k+1;

}

Until (cum_sum≤ σ)

End Do

10. Assign clusters, c = k;

11. Initially, select c centres randomly.

12. Compute minimum distances between data elements and cluster centres.

13. Update c using mean values of elements

14. Repeat steps 12 to 13 until no data elements won't be reassigned.

MH Extended K-Means is designed to determine the number of clusters for the well-separated dataset. It is a combined method of dynamically determining the number of the cluster with the traditional K-means algorithm. K-means works for well-separated dataset but, it needs a strong initial assumption to the number of clusters. To defeat limitation, MH Extended K-Means is designed. It accurately determines the number of the cluster for the well-separated dataset.

C. Predicting the Number of Cluster-based on σ Value

Non-parametric value σ is the limiting criterion indicating the percentage of variation of the dataset. σ value is determined from the principal component analysis technique. The cumulative summation of main components variances of the dataset is the limiting value of σ. If the value of σ is high, then the number of the cluster becomes less and vice versa. For example, in leukaemia dataset, the summation of three main principal components is 66.3% indicating 66.3% variation of the total dataset. So for leukaemia dataset, we set σ = 66.3 and finally it concludes to set 2 as the number of clusters. For Wisconsin Breast Cancer (WBC) dataset the outcome of MH clustering algorithm is three number of cluster and σ value is 86.7 and so three cluster hold at least 86.7% data. For Leukemia dataset number of cluster is two and σ value is 66.3, so two clusters have 66.3% data. For Irises dataset the number of clusters is two and σ value is 92.46, therefore, two clusters keep 92%. For Motor Cycle dataset the number of clusters is three and σ value is 85.3448 thus three clusters contain 85%. In MH Extended K-Means clustering algorithm number of cluster for WBC, Leukemia, Irises and Motor Cycle dataset is three, two, two and three respectively. MH Extended K-Means clustering algorithm brings 86%, 85%, 92% and 85% underlying data into consideration for dynamically determination of cluster number.

IV. A COMPARATIVE STUDY AMONG CLUSTERING ALGORITHMS

Performance of algorithms is compared based on average execution time vs. no. of clusters. Considering a certain number of clusters, execution times of the corresponding algorithm is obtained and comparative study is as below:

A. Time Comparison among Clustering Algorithms

Fig. 1, the proposed MH clustering algorithm takes the lowest time for clustering and MH Extended K-means clustering takes highest times. WBC dataset is not well separated and for this nature of datasets proposed fuzzy algorithm takes less time comparatively. Fig. 2, MH clustering algorithm takes the lowest time to perform clustering compared to other fuzzy clustering algorithms. Hard-core clustering like K-Means, K- Mediods takes comparatively longer time than MH Extended K-Means clustering algorithm. Leukemia dataset is not well separated and for these nature of dataset proposed fuzzy algorithm takes comparatively much less time. Fig. 3. Meskat-Hasan (MH) clustering algorithm takes the lowest time. Others fuzzy clustering algorithm like FCM, GG and GK takes almost the same times to perform executions. Besides, hard-core clustering likes K-Means, K-Mediods takes comparatively more times than MH Extended K-means clustering. In Fig. 4. Meskat-Hasan (MH) clustering algorithm takes the lowest time and MH Extended K-Means clustering takes highest times to execute.
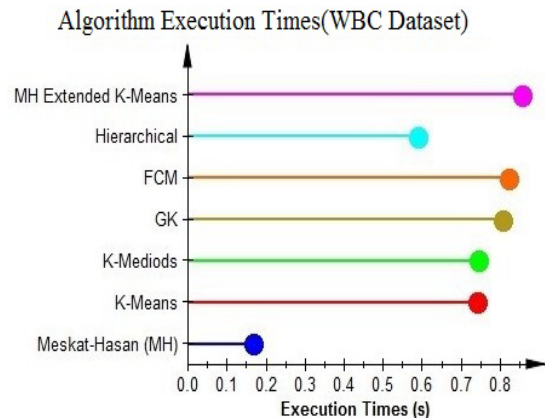
Fig. 1. Algorithms Performance Comparison on WBC Dataset.
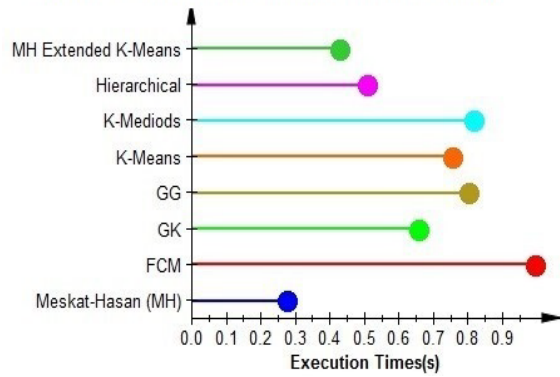
Fig. 2.   Algorithms Performance Comparison on Leukemia Dataset.
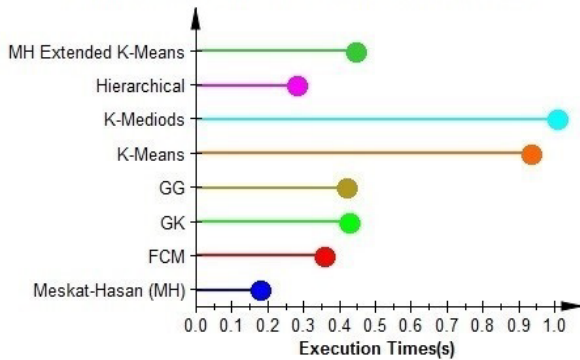
Fig. 3.   Algorithms Performance Comparison on Irises Dataset.
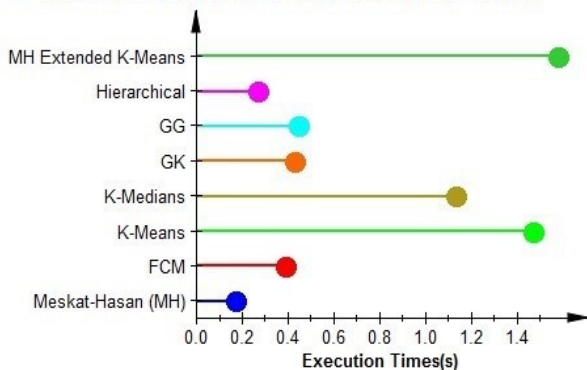
Fig. 4.   Algorithms Performance Comparison on Motor Cycle Dataset.

B. MH Over MH Extended K-Means Clustering Algorithm

MH uses the fuzzy approach in cluster implementation whereas MH Extended K-Means use hard clustering approach. MH Extended K-Means perform better when the input dataset is well separated and MH performs better for a non-linearly separable dataset.

V.   CLUSTER VALIDATION

To evaluate cluster performance analysis validation played an important role. For validating partition Separation index (S) takes the minimum separation distance where the smallest values of S indicate a valid optimal partition. To measure the

amount of overlapping between cluster partition coefficient (PC) indexing is used and its high value provide cluster accuracy. Dunn Index (DI) is internal evaluation criteria which identify the well separate cluster. Higher DI and ADI values indicate better clustering results. CE measures the cluster partitions fuzziness. Low values of CE and SC reflect good performance. XB index recognizes whole cluster compactness and smallest value site the optimum number of cluster. Validation result of the MH clustering is organized in Table I.

Table II reveals the performance of the MH Extended K-Means algorithm over the four datasets. As PC indicates the amount of overlapping between cluster regions, so it provides constant values in MH Extended K-Means which is a hard clustering algorithm.

TABLE I.   MH CLUSTERING ALGORITHM VALIDATION

Dataset	Cluster Number	PC	CE	SC	S	XB	DI	ADI
WBC	3	0.8	0.3	0.3	0.0	1.9	0.1	0.0
Leukemia	2	0.9	0.1	0.5	0.01	4.1	0.4	0.10
Irises	2	0.9	0.2	0.8	0.01	6.3	0.1	0.01
Motor Cycle	3	0.7	0.5	1.9	0.02	5.4	0.01	0.01

TABLE II.   MH EXTENDED K-MEANS CLUSTERING ALGORITHM VALIDATION

Dataset	Cluster Number	PC	CE	SC	S	XB	DI	ADI
WBC	3	1	-	0.31	0.001	3.2	0.17	0.004
Leukemia	2	1	-	0.56	0.007	4.7	0.38	0.03
Irises	2	1	-	0.62	0.004	41	0.1	0.002
Motor Cycle	3	1	-	1.37	0.012	16	0.01	2e-02

A. Analyzing Total Outcomes

All outcomes of both proposed and existing algorithms are compared based on the evaluation criteria for each dataset. Table III is the outcomes of a comparative view of the performance of the algorithm of WBC dataset. The values of PC & CE are respectively constant for hard clustering. The lowest values of SC, S & XB are provided by the MH clustering algorithm. Highest values of the DI and the lowest value of ADI are provided by the MH Extended K-Means clustering algorithms. Hence, MH clustering algorithm gives better results. Table IV shows the algorithms performance of Leukemia dataset. Based on values of PC, CE, SC, S, XB, DI and ADI it can be concluded that for Leukemia dataset GK and MH clustering algorithm gives better results. Table V organized the validation result of Irises dataset. The values of PC & CE are respectively constant and not working for hard clustering. The lowest value of SC & S is provided by MH Extended K-Means clustering algorithm. The low value of XB provides by GG. But, the low value of the DI provided by the GG and GK and the lowest value of ADI are provided by the K-Mediods algorithm. Table VI performs comparative studies of Motor Cycle dataset. MH clustering algorithm takes lowest times. The values of PC & CE are not working for hard

clustering. The lowest value of SC, S & ADI is provided by the MH Extended K-Means clustering algorithm. By considering above stated measurement MH Extended K-Means clustering algorithm gives a better result for the motorcycle dataset.

TABLE III.   ALGORITHMS PERFORMANCE COMPARISON ON WBC DATASET

Algorithms	SC	PC	S	CE	DI	XB	ADI
MH	0.3	0.8	6E-4	0.3	0.1	1.9	0
FCM	0.7	0.8	1E-3	0.3	0.1	2.6	0
GK	0.6	0.9	1E-3	0.2	0.1	2.2	0
K-Means	0.5	1.0	9E-4	NA	0.1	2.7	0
K-Mediods	0.5	1.0	1E-3	NA	0.1	Inf	0
MH Extended K-Means	0.3	1.0	7E-4	NA	0.2	3.2	0

TABLE IV.   ALGORITHMS PERFORMANCE COMPARISON ON LEUKEMIA DATASET

Algorithms	SC	PC	S	CE	DI	XB	ADI
MH	0.5	0.9	7E-3	0.1	0.4	4.1	0.11
FCM	0.6	0.9	9E-3	0.2	0.4	5.2	0.03
GK	0.5	0.9	7E-3	0.1	0.4	4.1	0.11
GG	1.7	1.0	2E-2	0.0	0.4	2.7	0.05
K-Means	0.5	1.0	7E-3	NA	0.4	4.7	0.03
K-Mediods	0.6	1.0	1E-2	NA	0.4	Inf	0.02
MH Extended K-Means	0.6	1.0	8E-2	NA	0.4	4.7	0.03

TABLE V.   ALGORITHMS PERFORMANCE COMPARISON ON IRISES DATASET

Algorithms	SC	PC	S	CE	DI	XB	ADI
MH	0.8	0.9	0.0	0.2	0.1	6.3	0.01
FCM	0.8	0.9	0.0	0.2	0.1	6.3	0.01
GK	0.6	0.9	0.0	0.2	0.2	16.6	0.01
GG	1.0	1.0	0.0	4E-4	0.2	1.9	0.0
K-Means	0.6	1	0.0	NA	0.1	41.2	0.0
K- Mediods	0.6	1	0.0	NA	0.1	Inf	0.0
MH Extended K-Means	0.6	1	0.0	NA	0.1	41.2	0.0

TABLE VI.   ALGORITHMS PERFORMANCE COMPARISON ON MOTOR CYCLE DATASET

Algorithms	SC	PC	S	CE	DI	XB	ADI
MH	1.9	0.7	0.02	0.5	0.01	5.3	0.01
FCM	1.9	0.6	0.02	0.5	0.0	5.4	0.01
GK	2.0	0.6	0.02	0.6	0.0	3.7	0.02
GG	4.9	0.9	0.05	0.1	0.03	1.9	0.06
K-Means	1.4	1	0.01	NA	0.01	14	0.02
K-Mediods	1.5	1	0.01	NA	0.02	Inf	0.0
MH Extended K-Means	1.4	1	0.01	NA	0.01	16	0.0

VI. RESULT AND DISCUSSION

MH clustering algorithm provides solutions for automatic clusters number detection, run time cluster selection, and performs fuzzy clustering accordingly. It appropriately determines clusters number and produces proper partitioning. MH clustering algorithm works well for the non-linear dataset. MH Extended K-Means algorithm performs hard clustering on the basis of dynamic cluster number determination. It works well for a clearly separable dataset. By analyzing all results, both MH and MH Extended K-Means clustering algorithms select the precise cluster number and produce optimum partitioning and perform clustering accordingly. MH clustering algorithm takes comparatively less time to execute than other algorithms. Based on validation techniques, MH clustering algorithm performance is quite better than the other established literature. MH clustering algorithm meets the objective of automatic cluster number detection without using post cluster analysis and performs clustering accurately and satisfy the time complexity. Though MH Extended K-Means clustering algorithm takes more time than the MH clustering algorithm but performs better than other hard clustering algorithms. Evaluating validation results MH and MH Extended K-Means clustering algorithm are acceptable.

VII. CONCLUSION

The idea of MH clustering algorithm and MH Extended K-Means clustering algorithm comes to solve the problem of the exact number of cluster detection automatically and perform clustering accordingly. MH and MH Extended K-Means were applied on both linear and non–linearly partitioned dataset. Performance of the proposed algorithms was compared with other selected algorithms by validating the cluster and performance evaluation. The comparison was done based on execution time and validation indexes and it provides an effective way of selecting an efficient clustering algorithm for the particular dataset. For linearly separable dataset performance of MH Extended K-Means clustering algorithm and non-linearly separable dataset, MH clustering algorithm is better. Both MH and MH Extended K-Means clustering algorithm meet the desired needs of dynamically determining the number of clusters accurately and provides better and efficient results than the selected clustering algorithms.

Here we work on gene expression datasets and algorithms are tested in standalone systems. In the future, we want to work with real-time microarray gene expression dataset and implement in parallel and distributed system and will upgrades the algorithms accordingly. So that classification of microarray gene expression data can take computational benefit from cloud infrastructure.

REFERENCES

[1] Truong, H.Q., Ngo, L.T. and Pedrycz, W., 2017. Granular fuzzy possibilistic C-means clustering approach to DNA microarray problem. Knowledge-Based Systems, 133, pp.53-65.

[2] Bulut, H., Onan, A. and Korukoğlu, S., 2020. An improved ant-based algorithm based on heaps merging and fuzzy c-means for clustering cancer gene expression data. Sādhanā, 45(1), pp.1-17.

[3] Paul, A.K. and Shill, P.C., 2018. Incorporating gene ontology into fuzzy relational clustering of microarray gene expression data. Biosystems, 163, pp.1-10.

[4] Das, R. and Saha, S., 2015, November. Gene expression classification using a fuzzy point symmetry based PSO clustering technique. In 2015 Second International Conference on Soft Computing and Machine Intelligence (ISCMI) (pp. 69-73). IEEE.

[5] Hosseini, B. and Kiani, K., 2018. FWCMR: A scalable and robust fuzzy weighted clustering based on MapReduce with application to microarray gene expression. Expert Systems with Applications, 91, pp.198-210.

[6] Jiang, Z., Li, T., Min, W., Qi, Z. and Rao, Y., 2017. Fuzzy c-means clustering based on weights and gene expression programming. Pattern Recognition Letters, 90, pp.1-7.

[7] Zareizadeh, Z., Helfroush, M.S., Rahideh, A. and Kazemi, K., 2018. A robust gene clustering algorithm based on clonal selection in multiobjective optimization framework. Expert Systems with Applications, 113, pp.301-314.

[8] Izakian, H., Pedrycz, W. and Jamal, I., 2015. Fuzzy clustering of time series data using dynamic time warping distance. Engineering Applications of Artificial Intelligence, 39, pp.235-244.

[9] Ludwig, S.A., Picek, S. and Jakobovic, D., 2018. Classification of cancer data: analyzing gene expression data using a fuzzy decision tree algorithm. In Operations Research Applications in Health Care Management (pp. 327-347). Springer, Cham.

[10] Asuncion, A., & Newman, D.: UCI machine learning repository. (2007).

[11] D'haeseleer, P.: How does gene expression clustering work? Nature biotechnology, 23(12), 1499-1501. (2005).

[12] Jain, A. K., & Dubes, R. C.: Algorithms for clustering data. Prentice-Hall, Inc... (1988).

[13] Bezdek, J. C., Ehrlich, R., & Full, W.: FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203. (1984).

[14] Lee, E. T. L., A. ZADEH: Note on Fuzzy Languages. Inform. Sciences, 1(4). (1969).

[15] Gustafson, D. E., & Kessel, W. C.: Fuzzy clustering with a fuzzy covariance matrix. In 1978 IEEE conference on decision and control including the 17th symposium on adaptive processes (pp. 761-766). IEEE. (1979, January).

[16] Ruspini, E. H.: A new approach to clustering. Information and control, 15(1), 22-32. (1969).

[17] Liang, J., Zhao, X., Li, D., Cao, F., & Dang, C.: Determining the number of clusters using information entropy for mixed data. Pattern Recognition, 45(6), 2251-2265. (2012).

[18] Kaiser, H. F.: The application of electronic computers to factor analysis. Educational and psychological measurement, 20(1), 141-151. (1960).

[19] Cliff, N.: The eigenvalues-greater-than-one rule and the reliability of components. Psychological bulletin, 103(2), 276. (1988).

[20] Patibandla, R. L., & Veeranjaneyulu, N.: Performance Analysis of Partition and Evolutionary Clustering Methods on Various Cluster Validation Criteria. Arabian Journal for Science and Engineering, 43(8), 4379-4390. (2018).

[21] Jahan, M., & Hasan, M.: Performance Analysis and Benchmarking of Clustering Algorithms with gene datasets. In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) IEEE. (pp. 1-5). (2019, May).

Local Binary Pattern Method (LBP) and Principal Component Analysis (PCA) for Periocular Recognition

Sereen Alkhazali[1], Mohammad El-Bashir[2]

Department of Computer Science
Faculty of Computer Science and Information Technology
Al-al Bayt University - Jordan

Abstract—**Identification of identity through eye is gaining more and more importance. Commonly, the researchers approach the eye from any of three parts, the iris, the circumference around the eye, and the iris and its circumference. This study follows a holistic approach to identity identification by using the iris and whole periocular area and proposes a periocular recognition system (PRS) that has been developed using the Local Binary Pattern (LBP) technique combined with Principal Component Analysis (PCA) at the feature extraction stage and the k-nearest neighbors (k-NN) algorithm as a classifier at the classification stage. This system achieves identity recognition through three steps: pre-processing, feature extraction, and classification. Pre-processing is applied to the images so as to convert them to grayscale. In the feature extraction step, the LBP method is applied to extract the texture feature from the images and use it in PCA to reduce data dimensionality and obtain the relevant data so that only the important features are extracted. These two steps are applied both in the training phase and the testing phase of image processing. On the other hand, the testing data sets are processed using the k-NN classifier. The proposed PRS was tested on data drawn from the PolyU database using more than one basis of system experience. Specifically, the system performance was tested once on all 209 subjects present in the database and once on 140 subjects. This database also contains images taken in the visible (VIS) and near-infrared (NIR) regions of the electromagnetic radiation (EMR) spectrum. So, the system was tested on images taken in both regions separately for matching. As well, the proposed PRS benefited from the availability of images for the right and left perioculars. Performance was, therefore, tested on images of each side of the periocular area (the left and right sides) separately, as well as for the combination of the two sides. The identity recognition rates characteristic of the proposed PRS were most often higher than the recognition rates produced by systems reported in the literature. The highest recognition accuracy obtained from the proposed system, which is 98.21%, was associated with the 140-subject data sub-set.**

Keywords—Periocular recognition; Local Binary Pattern (LBP); Principal Component Analysis (PCA); k-Nearest Neighbors (k-NN)

I. INTRODUCTION

Nowadays, security of systems is more important than the systems themselves and authentication serves as the first line of defense against any unauthorized persons. Most frequently, authentication can be achieved by one of three ways: password, biometry, and a characteristic identifier of the user such as a card. Biometry provides secure means of authentication and identification owing to that it is difficult to steal, or even replicate, biometric information. Furthermore, many options are available for biometric identification such as fingerprint, DNA, sound pattern, signature, iris patterns, periocular print, and retina patterns. As far as face recognition is concerned, the periocular print is the most accurate biometric property that can be used for face identification [1].

Periocular authentication is an automated method of biometric identification that applies mathematical pattern recognition techniques on the video images of one or both of the eyes of a person, whose complex patterns are unique, stable, and can be seen from some distance. According to Kumari and Seeja [2], authentication based on the periocular region builds on features taken from both the face and iris. In this context, the periocular biometric expresses the facial region right close to the eye [2], and the periocular region, which is the area around the eye, encircles the eyebrows, eyelashes, eyelids, and the adjacent skin area [3]. But nlike acquisition of many other ocular biometrics, acquisition of a periocular biometric does not demand a high user cooperation and a close capture distance [4]. However, even though the periocular area is regarded as highly-discriminating part of the human face, its utility as independent soft biometric or modality is still under scrutiny. Hence, it is a research goal to establish performance metric for features of the periocular area so that their likely use in combination with the face or iris can be assessed.

The periocular recognition system (PRS) is a system designed for identification of people through the iris and the region surrounding the eyes, which is a region that provides wide space and feature richness for discriminating people and identifying the authorized person. Within this context, the major objective of this study was to develop a PRS that can effectively recognize people through their eyes. After acquisition of cross-spectral iris data (e.g., from the PolyU Cross-Spectral Iris Database), the proposed PRS processes the images in three steps: (i) converting the original image to the grayscale; (ii) extracting the features by using the Local Binary Pattern (LBP) method and Principal Component Analysis (PCA); and (iii) classifying the extracted features using the k-nearest neighbors (k-NN) algorithm.

II. Literature Review

Many classifiers have so far been employed for recognition of the periocular iris. They include varied Artificial Intelligence (AI) systems like the SVM, Histogram of Oriented Gradients (HOGs), and the Feed-Forward Back Propagation Neural Network (FBPNN), amongst others.

Woodard et al. [5] illustrated application of the fusion methods on iris and periocular images in the case of non-ideal images of the eye, namely, images that are distinguished with poor contrast, occluded irises, illumination artifacts, and motion and spatial blur. Outcomes of their analysis indicated that using the Multi-Biometric Grand Challenge (MBGC) database with score level fusion can enhance recognition performance in such images. Periocular texture is extracted from small, fixed region of the skin that surrounds the eye. Experiments on the images extracted from the near infra-red (NIR) face videos of the MBGC dataset using the fusion techniques demonstrated that valuable information is contained in the periocular area and that it can be integrated with the iris texture data to enhance the overall identification accuracy in atypical conditions.

Periocular biometrics include valuable information for the iris and face recognition systems. Fasca et al. [3] developed a PRS that employed the HOGs and LBP method for extraction of features from periocular images. The LBP is a kind of features that is employed for classification in capturing vision. It is a powerful feature for the texture. In other respects, the HOGs were employed for the purpose of object detection using gradient features. For effective recognition and classification of authorized personnel, these researchers [3] used the back propagation neural network (BPNN) classifier. Eventually, the PRS developed by these researchers had a recognition accuracy of 91%.

The first algorithm based on the Artificial Neural Network (ANN) for characterizing and classifying the image variations caused by different spectral ranges was proposed by Sharma et al. [6]. The algorithm employed the ANN for distinguishing the variabilities generated by two spectra; the VIS and NIR regions of the EMR spectrum. First, the spectra were trained individually and, then, joined such that, by the use of cross-spectral training data, the proposed algorithm can learn the cross-spectral variations. For evaluation of performance, a cross-spectral periocular database was prepared that consists of NIR and VIS night spectra images. The database was the IIITD multispectral periocular (IMP) database that includes 1,240 iris images of 62 individuals and which were captured within the VIS spectrum, night vision, and NIR iris cameras. The evaluation results pointed out that this proposed approach produced much higher levels of accuracy than four methods (Pyramid of Histograms of Oriented Gradients (PHOG), HOG, FPLBP, and LBP). The concomitant recognition rates ranged from 76.97% for the visible spectrum to 92.5% (NIR) in the "same-spectrum" analysis and from 48.21% for the night vision NIR to 71.93% for the visible night vision in cross-spectral analysis. In most experiments, combined left and right (L+R) periocular recognition had higher accuracy than single left or right (L/R) recognition.

Nie et al. [7] clarified that accurate automatic biometric identification of identity by using periocular imaging had broad spectrum of applications, extending from human surveillance to enhancement of performance for the IRSs, particularly under less-constrained imaging environments. The Restricted Boltzmann Machine (RBM) is a generative stochastic ANN that has the ability to learn the probability distribution of its input data. The Convolutional Restricted Boltzman Machines (CRBM) aim at accommodating large image sizes and markedly reducing the computational burden. However, the methods of unsupervised learning of features had not been investigated in the biometrics domain, except in face recognition. These researchers [7] explored effectiveness of the CRBM in periocular recognition. They implemented experiments on a publicly-available database of periocular images of a large sample of subjects (300 test subjects), namely, the Ubripr database, and exploited key features simultaneously to improve the matching accuracy. The results of experiments confirmed effectiveness of RBM feature learning in automated periocular recognition with large numbers of subjects. In addition, their results suggest that supervised metric learning can result in better identity recognition than the Euclidean distance.

Behera et al. [8] underscored that periocular recognition had been active domain of research in the last few years. They suggested an identity recognition method that is based on illumination normalization of NIR and VIS periocular images. Their proposed method entailed normalization of the images by using the difference of Gaussian (DoG) filtering, followed by calculation of descriptor that captured the structural details present in the illumination-normalized images by using the HOG. Lastly, those feature vectors that correspond to the query and the enrolled image were compared based on the cosine similarity metric so as to produce a matching score. Performance of this algorithm was, then, assessed on three publicly-available benchmark databases of cross-spectral periocular images.

Ramaiah et al. [9] developed an identity recognition method based on Markov random fields (MRFs) and three-patch LBP (TPLBP) descriptor. They used exact pixel correspondences during image acquisition and synthesized the NIR image pixels from the VIS periocular images. Two versions of LBP descriptors, that is, the TPLBP and four-patch LBP (FLBP) were then extracted from the images. The method was then evaluated on the IMP and PolyU datasets using both descriptors. The evaluation results pointed out that the TPLBP descriptor outperformed the FLBP operator as it had Guanine Accept Rate (GAR) values of 18.35% and 73.20% at 1.0% FAR when tested on the IMP and PolyU databases, respectively.

III. Methodology

The periocular recognition system (PIRS) progresses in four sequential steps as shown in Fig. 1. The first step is acquiring image by camera. The second step is preprocessing, whereby the original image is converted to a grayscale image. The third step is extracting features, whereby both the LBP method and PCA are employed in order to extract the most important features for recognition. The last step is classifying

the extracted features, in which a comparison is held between a new biometric feature and one model or all models in the database. This step outputs a similarity score for every comparison by using the k-NN classifier. Then, performance analysis is conducted so as to determine whether or not the probe and the model that are obtained from the database match on the basis of whether the similarity score is higher or lower than a pre-set match threshold.

Sometimes, methods of transforming the raw feature vectors, i.e., feature extraction methods, are employed before application of the pattern-matching algorithm. The algorithms of feature extraction try to transform a large-dimensionality feature space to a smaller-dimensionality space that is easier to process and which encodes less redundancy by use of mathematical methods like the LBP method and PCA.

A. Database

The PolyU Cross-Spectral Iris dataset is a distinctive, bi-spectral dataset of iris images that was developed for cross-spectral iris recognition studies and made publicly available. This dataset is a compilation of face images taken under the condition of simultaneous bi-spectral imaging, from both the left and right eyes. A sample of such images is given in Fig. 2. Overall, this dataset is made up of a total of 12,540 iris images ($209 \times 2 \times 2 \times 15$) that were obtained from 209 subjects in 15 instances, each. Every one of these iris images has the dimensions of nearly 640×480 pixels. Moreover, the pixel correspondences of these iris images lie both in the VIS and NIR regions of the EMR spectrum [10].
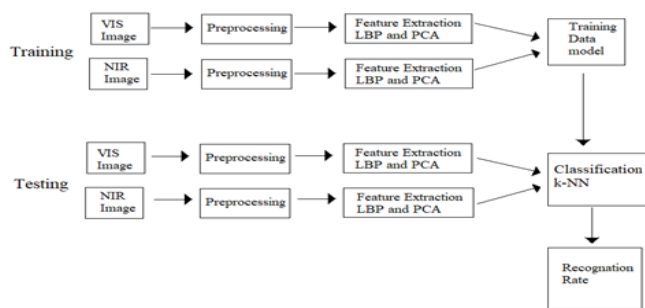
Fig. 1.   General Structure of the Periocular Recognition System (PRS).
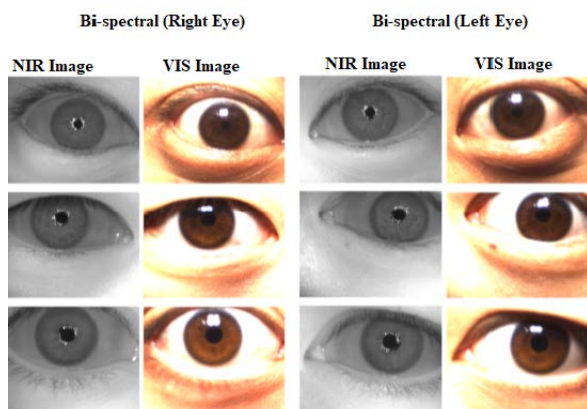
Fig. 2.   Sample Iris Images Simultaneously Shot in the NIR and VIS Regions of the EMR Spectrum with Full Pixel Correspondences (Source: [10]).

It should be highlighted that the researchers encountered several difficulties in the present study while working with this database, the most critical of which was that the related previous studies did not specify which particular data of this dataset they used. This prompted the researchers to expand the current inquiry in an effort to grow able to make reasonable comparisons with relating previous studies.

B. Periocular Recognition System (PRS)

This study develops a PRS that is based on a combined approach to feature extraction and the k-NN classifier. This proposed system achieves periocular recognition in three steps as explained in the subsequent sub-sections.

1) Preprocessing

The proposed PRS handles the inherent variations in illumination between the NIR and VIS images in the pre-processing phase by converting the images to the gray scale. Variations in illumination impact matching of the VIS periocular images negatively. Every image will be transformed into binary values and stored in a matrix for later evaluation.

After converting the periocular image to a grayscale image in the pre-processing step, it is analyzed for each person and features are, then, extracted using the combined LBP-PCA recognition system and stored in a log file in the database for verification and authentication. Upon receipt of an image for authentication, features are analyzed and extracted by using the combined system and a new record for comparison is created in the next (classification) step. A description of periocular image feature extraction is given in the subsequent sections.

The LBP method is very useful because it is a robust method for illumination of change. It is a sort of visual descriptor that is employed for classification in computer vision and which proved to be powerful tool for texture classification. It improves the detection performance and is simple and quick to compute. However, certain binary patterns exist more commonly in some texture images than others. In its simplest form, the LBP feature vector is created in the following manner: divide the periocular image into cells (e.g., 3x3 pixels for each cell), then encode the center pixel by a series of bits for each center pixel in a cell of eight neighbors. The neighboring pixels are given a binary value of 1 if they are equal to, or higher than, the central pixel and 0 if otherwise. Eight-bit binary number is then formed by concatenating the bit. Accordingly, its decimal equivalent becomes the label for the center pixel, which is denoted by a circle as shown in Fig. 3.
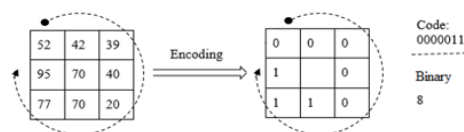
Fig. 3.   Example of LBP Encoding.
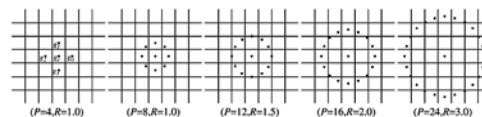
Fig. 4.   Symmetric Circular Neighbor Sets for Different P and R Values (Source: [11]).

This study employed the LBP features in an effort to analyze texture of the entire image for eye recognition. The fundamental LBP method takes into account small circularity neighborhood and a rotation invariant texture which has symmetric points (P) that are pixels spaced equally on circle of a radius (R) for the whole eye image by use of a global threshold. Fig. 4 shows some sizes and shapes of eye images produced by the LBP method.

2) Principal Component Analysis (PCA)

Principal Components Analysis (PCA) is a standard, and practical, statistical tool for data analysis that has established applications in various areas like neuroscience, face recognition, and image compression [12].

In this analysis, the first component that is extracted (Component 1) explains the maximum amount of the total variance in the observed variables. Under the ideal conditions, this implies that the first component correlates the highest with some observed variables. Though, however, it may correlate with many. The second extracted component (Component 2) has two important characteristics. First, it accounts for the maximum amount of the variance in the data that was not explained by the first component. Again, under the ideal conditions, this suggests that the second component correlates with some of the observed variables which do not show high correlations with Component 1. The second characteristic of Component 2 is that it does not correlate with the first component. That is, if one is to assess correlation between the first two components, then the correlation coefficient is zero. The rest extracted components exhibit the same two characteristics; every component explains the maximum amount of the variance in the observed variables that is not explained by the previous components and is not correlated with any previous component. Principal component analysis progresses in this way, with every new component explaining progressively lower and lower proportion of variance. This is why usually only the first few components are kept and interpreted. Once the analysis is complete, the extracted components manifest varying degrees of correlation with the observed variables, but are entirely uncorrelated with one the other [12].

The main steps in PCA are the following:

Step 1: Getting relevant data

In this study, a 2x2 data sub-set is drawn from pixel map of the image.

Step 2: Subtracting the mean

In PCA, one should subtract the mean from every data dimension. The subtracted mean is the average across every dimension. Hence, all the x values have the mean of the x values of all data points deducted from them. Likewise, all the y values have the mean of all the data points deducted from them. This generates a dataset with a mean of zero.

Step 3: Computing the covariance matrix

The covariance matrix is mainly created for the following two purposes:

i) Assessing the extent to which every dimension varies from the mean.

ii) Estimating the covariance between two dimensions to determine whether or not there are relations among them.

It should be underlined that covariance between any dimension and itself is, in fact, the variance.

In this regard, the following equation Eq. (1) was employed to compute the covariance of the two dimensions:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - X)(Y_i - Y)}{(n-1)} \tag{1}$$

The diagonal of the covariance matrix represents variances of x, y, and z, e.g., cov(x,y) = cov(y,x). Thus, the matrix is symmetrical about the diagonal element. The resultant covariance may be categorized into three types:

i) Positive covariance when the two dimensions decrease or increase simultaneously.

ii) Negative covariance when one dimension decreases and the other increases,

iii) Zero covariance when the two dimensions are independent.

The covariance is calculated to define the relations among dimensions in high-dimensional data (usually higher than three) where visualization is difficult.

3) Classification

Classification is a process which recognizes a periocular image by finding a match of its features with one of the periocular images in the dataset. The PRS proposed in this study went through two phases: a testing phase and a training phase. The classifier was used only in the testing phase to identify a periocular image via its features. The features of the periocular image were extracted in the testing phase by the combined LPB-PCA system according to the same steps followed in the training phase. Then, the features of the training model were matched so as to identify the unknown image.

The k-NN algorithm is a simple algorithm that stores the available cases and classifies the new ones based on some similarity measure. It is one of the predictive models and it does not need to learn complicated mathematical equations; it just needs to (i) have a way for calculating the distance between data and (ii) perform a hypothetical investigation to ensure that the data close together are similar and far away from others [13].

Most of the techniques used in the predictive model look at the dataset as a whole in order to characterize the data patterns, but the nearest neighbors conceal a lot of information because they are predicted for each new point depending only on the number of points near it.

The output of the k-NN algorithm is a label of the closest owner of the testing periocular image. The particular object is classified by means of plurality vote of its nearest neighbors, with the object chosen by sorting all objects in descending order of votes and selected from its k nearest neighbors, where

k is a positive, typically, small, integer. This study adopted the k value of 5.

The main steps of the k-NN algorithm are the following:

- Determine the value of the variable k, which is the number of neighbors.

- Compute distances between the new image and the images in the dataset.

The distances between the testing and training images (p and q) in the dataset are calculated using the equation (2):

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_{i})^2} \qquad (2)$$

- Arrange the images to get the neighbors based on the lowest distance calculated in the previous step, and take the number of the neighboring ks.

- Define a label for the neighbors.

The label, namely, the vast majority of neighboring images, is the expected label by Eq. (3)

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_{i})^2} \qquad (3)$$

Thereafter, the recognition rate (accuracy, in effect) is calculated by using the following equation Eq. (4):

$$RR = \frac{correctly\ detected\ periocular\ images}{Total\ no.of\ periocular\ image.} * \mathbf{100} \qquad (4)$$

4) Implementation

Practically, the proposed PRS was run on PC to examine its performance that has Intel (R) Core (TM) i5-8265U processor with CPU @ 2.20 GHz, fitted with 4 GB RAM, and operating on the Windows 8 operating system. The algorithms were written and executed in the Matlab 2018a programming language. However, some calculations were performed using Microsoft Excel 2010.

IV. RESULTS AND DISCUSSION

This study develops and proposes a PRS based on feature extraction with LBP and PCA. The features are fed into the k-NN algorithm for classification. The proposed system was tested on the PolyU Cross-Spectral Iris dataset and its performance was compared with levels of performance of contemporary PRSs based on the recognition rate. This section presents the experimental results of this system in all cases and compares them with results of previous related works.

A. Training and Testing

To evaluate the identity recognition efficiency of the proposed PRS, this system was tested on the PolyU Cross-Spectral Iris dataset of Hong Kong Polytechnic University. This dataset of periocular and iris images was built by drawing data from 209 subjects. Every single subject provided two sets of data. The first set comprises 15 images taken from the left side of the face whereas the second set contains 15 images taken from the right side of the face. In the present study, experiments were performed on iris images pertaining to 140 subjects. The first 10 images in each of the left and right data

sets were used in training while the last five images in each of these two sets were employed in the testing. As well, the proposed PRS was tested using images taken in the visible light, both in the VIS and NIR regions of the EMR spectrum. On the other hand, training was conducted on images using near-inference (VIS-VIS and NIR-NIR) analysis and cross-spectral (VIS-NIR) analysis.

B. Experimental Result

This section presents the results of testing the proposed PRS. To determine the conditions conducive to optimum classification, differing values of input parameters for feature extraction were examined. Then, the recognition results were compared with results of other systems based on the recognition rate. The analysis results are presented in two sub-sections, one each for:

i) The same spectral region.

ii) Cross-spectral regions.

1) Results Related to Images in the same Spectral Region

The PolyU dataset provides multispectral periocular images shot in the VIS and NIR regions of the EMR spectrum. In this sub-section, the researchers present the results of application of the proposed PRS to the same spectral regions in the training and testing steps, that is, the VIS-VIS and NIR-NIR regions.

The researchers chose to increase the parameters of the LBP (P, R) in order to get higher recognition accuracy than 95%. The parameter P is a positive integer whose typical values range from 8 to 24 and the parameter R is a positive integer, too, whose typical values range from 1 to 5. The levels of performance of the suggested system in recognizing the 140 subjects taken from the PolyU dataset and shot in the VIS region are summarized by Table I and Fig. 5.

The evaluation results uncover that the recognition accuracy is higher for the radius value of 5 than the radius value of 3 and that the neighborhood value affects the recognition accuracy only slightly. The best recognition accuracy is associated with the k value of 18 (Table I and Fig. 5). In addition, this study tested performance of the proposed PIRS using data for the 209 subjects in the PolyU dataset using the same P and R values used in the case of the 140 subjects (Table I). The results indicate that the best accuracy of recognition (94.14%) of the 140 subjects was associated with LBP (18, 5).

TABLE I. PERFORMANCE OF THE PROPOSED PRS BASED ON DIFFERENT VALUES OF THE LBP PARAMETERS (P AND R)

Neighborhood	Dimensional histogram	Recognition accuracy (%)	
		R=1	R=5
P=8	59	92.50	95.01
P=16	243	92.79	95.03
P=18	309	92.50	95.14
P=20	383	92.93	95.07
P=24	555	92.50	94.93
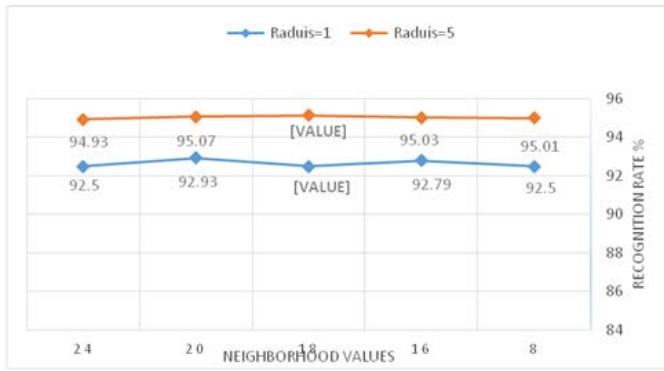
Fig. 5. Recognition Accuracy Values of the Proposed PRS based on Different Values of the LBP Parameters (P and R).

Table II displays the recognition accuracy values based on different bases, where the recognition accuracy was tested based on the number of images (209 and 140 subjects) used in matching by using the periocular images of right and left sides of the face, once separately, and once together. Lastly, levels of performance of the suggested PRS were examined for the two regions of the EMR spectrum separately, i.e., the VIS and NIR regions.

It is noted in Table II that the recognition rates are higher in all cases for the NIR region than for the VIS region since the former region relieves any physical discomfort resulting from the illumination, reduces the specular reflections, and increases the proportion of texture that is captured for some of the iris colors [14].

TABLE II.   RECOGNITION ACCURACY VALUES FOR THE NIR AND VIS SPECTRAL REGIONS

Side	Recognition accuracy (%)			
	VIS-VIS		NIR-NIR	
	140 Subjects	209 Subjects	140 Subjects	209 Subjects
Left	96.00	96.17	98.14	97.89
Right	95.00	94.07	98.43	97.13
Both	95.14	94.04	98.21	97.32

In the case of images taken in the same spectral regions, the levels of performance of the proposed PIRS were the best in the case of the NIR-NIR owing to that the NIR region captures the details of the periocular area, even in the case of a heavily-pigmented periocular image.

C. Results of Cross-Spectral Analysis

This sub-section highlights the results of the experiments performed on the cross-spectral periocular images. The PolyU dataset provides multispectral periocular images shot in the VIS and NIR regions. These images were processed in the training and testing steps much like the VIS-NIR images were.

TABLE III.   RECOGNITION ACCURACY VALUES FOR BOTH THE NIR AND VIS SPECTRAL REGIONS

Side	Recognition accuracy (%)	
	VIS-NIR	
	140 Subjects	209 Subjects
Left	100	100
Right	99.93	99.95
Both	97.07	96.41
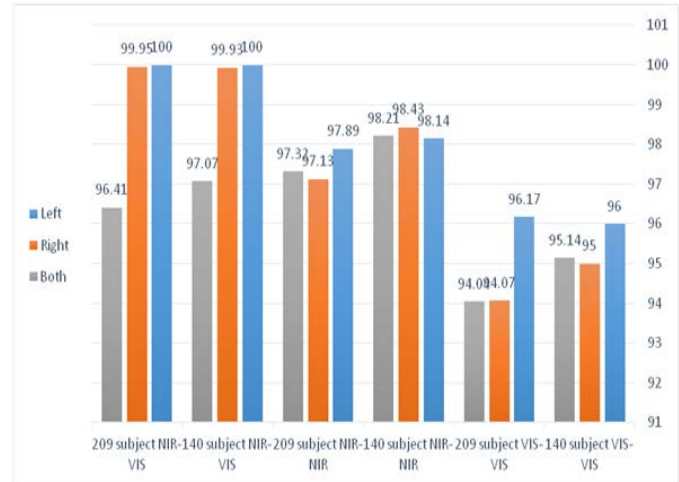
Fig. 6. Comparison of Recognition Percentages for both the VIS and NIR Spectral Regions and Cross-Spectral Analysis.

Table III shows the recognition percentages associated with cross-spectral matching for 140 and 209 subjects using single periocular images and pairs of periocular images. Additionally, Fig. 6 displays the results relating to all the cases of the EMR spectrum addressed by this study, including single periocular images and pairs of periocular images.

Fig. 6 unveils that the optimum recognition results were realized when matching was conducted for both the left and right periocular areas individually, then for their pairs. Moreover, the results disclose that there is remarkable difference in recognition between the left and right periocular regions of the same subject such that performance on the left side is better than that on the right side of the face.

D. Discussion with Comparison with Previous Approaches

Two metrics were used for performance evaluation in related previous works: the False Rejection Rate (FRR) and Guanine Accept Rate (GAR) [15]. The FRR is the proportion of genuine users who are rejected by the biometric system. In verification of a biometric system, the user will make claims of her/his identity and, therefore, the system must not reject enrolled user. That is, the number of false rejections (Equation Eq. (5)) must be kept at minimum.

$$FRR = \left(\frac{No.of\ rejected\ genuine}{Total\ no.genuine\ assessed} \right) * 100 \qquad (5)$$

TABLE IV.   PERFORMANCE COMPARISON

Approach	GAR (or Recognition accuracy)
Ramaiah et al. [9] (140 subjects)	73.20
Behera et al. [8] (209 subjects)	83.12
Our approach (140 / 209 subjects)	97.07 / 96.41

The GAR, on the other hand, is defined as the proportion of genuine users accepted by the system. It is given by Equation Eq. (6) [15]:

$$GAR = 100 - FRR \qquad (6)$$

With this simple definition, it can be said that GAR is equivalent to the recognition accuracy. On account of this, the performance evaluation measure is actually the recognition accuracy.

Table IV points out that the approach which Behera et al. [8] followed produced better results than the approach followed by Ramaiah et al. [9]. The results of the current study prove to be very good relative to the results of these two studies (Table IV), with noticeable differences. The proposed PRS succeeded in almost all cases.

V. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

In this study, a PRS was proposed, developed, and tested. This system progresses in three steps: pre-processing, feature selection, and classification. The major contribution of this study is optimizing the periocular identity recognition process.

The feature extraction step is based on a combined LBP method and PCA. In this step, the features are extracted from the entire periocular image. In the meantime, the LBP method, which was originally developed for texture analysis, was employed to extract the features of the iris and periocular area as vectors. The study then applied PCA. Lastly, the k-NN algorithm was applied in the Matlab environment in order to determine the optimum value of k for the best recognition outcomes.

The study results were categorized according to three foundations. The first foundation was the number of the periocular images (images of the right or left side of the periocular area, or of both sides). The second foundation was region of the EMR spectrum (the VIS and NIR regions). The third foundation was the number of subjects whose images were employed in the experiments (140 and 209 subjects). The study results bring to surface that the proposed PRS best performed with 140 subjects when using the periocular images shot in the NIR region for pairs of pictures. The concomitant recognition accuracy was 98.21%. Performance of the proposed PRS was compared with levels of performance of other systems. The comparison revealed that performance of the PRS proposed herein is better than levels of performance of other systems, with noticeable differences. Indeed, the proposed system gave the best recognition results in all the tested cases.

In view of the study results, the researchers suggest to testing the performance of the herein proposed PRS (i) on other dataset(s), (ii) in iris recognition, and (iii) using other classification methods like the ANN.

Lastly, periocular recognition can be further improved by considering clues such as eye shape and eye size.

REFERENCES

[1] Joke A. Badejo, Adekunle A. Akinrinmade, Emmanuel Adetiba, (2019)," Survey of Periocular Recognition Techniques", Journal of Engineering Science and Technology Review Vol.12 , No 5,pp 214 – 226.

[2] Kumari P., Seeja K.R. (2019), " Periocular biometrics: A survey", Elsevier B.V. on behalf of King Saud University.

[3] Fasca Gilgy Mary P., Sunitha Kency Paul P., Dheeba J., (2013)," Human Identification Using Periocular Biometrics", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 2, No. 5.

[4] Park U., Ross A., and Anil K. Jain, (2009) " Periocular Biometrics in the Visible Spectrum: A Feasibility Study", Theory, Applications and Systems (BTAS 09), Washington DC.

[5] Woodard D. L., Pundlik S., Miller P., Jillela R., and Ross A., (2010), "On the fusion of periocular and iris biometrics in non-ideal imagery", in Proceedings of International Conference on Pattern Recognition, pp. 201–204.

[6] Sharma A., Verma S., Vatsa M., Singh R., (2014), "On cross spectral periocular recognition", Image Processing (ICIP) 2014 IEEE International Conference on. IEEE, pp. 5007-5011.

[7] Nie L., Kumar A. and Zhan S., (2014), "Periocular Recognition Using Unsupervised Convolutional RBM Feature Learning", Proc. 22nd Intl. Conf. on Pattern Recognition, ICPR 2014, Stockholm, pp. 399-404.

[8] Behera S. S., Gour M., Kanhangad V., and Puhan N., (2017), "Periocular recognition in cross-spectral scenario", In IEEE IJCB, pp 681–687.

[9] Ramaiah N. P. and Kumar A., (2016), "On matching cross-spectral periocular images for accurate biometric identification," in 8th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS '16).

[10] Ramaiah N. P. and Kumar A., (2017), "Toward More Accurate Iris Recognition Using Cross-Spectral Matching", IEEE Transactions on Image Processing, Vol. 26, No. 1, pp. 208–221.

[11] Ojala T., Pietikainen M., and Maenpaa T., (2002), "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, pp. 971–987.

[12] Gautam D., (2014), "Facial expression detection using implemented PCA algorithm", International Journal of Computer Science and Technology Research, Vol. 2, No. 7, pp. 68-73.

[13] Chomboon K., Chujai P., Teerarassammee P., Kerdprasop K., Kerdprasop N., (2015). "An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm", The Proceedings of the 2nd Int. Conf. on Industrial Application Engineering 2015, 280–285.

[14] Burge MJ, Monaco MK (2009)," Multispectral iris fusion for enhancement, interoperability, and cross wavelength matching.", Proceeding of SPIE 7334, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV, Vol. 7334. SPIE. pp 73341D–1–73341D–8.

[15] Shodhganga,https://shodhganga.inflibnet.ac.in/bitstream/10603/55920/11/11_chapter6.pdf Webpage last accessed 18/7/2020.

Design and Performance Analysis of Different Dielectric Substrate based Microstrip Patch Antenna for 5G Applications

Nurulazlina Ramli[1], Shehab Khan Noor[2*], Taher Khalifa[3] and N. H. Abd Rahman[4]

Centre of Advanced Electrical and Electronic Systems (CAEES)[1, 2, 3]

The Faculty of Engineering, Built Environment and Information Technology, SEGi University, Kota Damansara.

47810 Petaling Jaya, Selangor, Malaysia

Antenna Research Centre (ARC), Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia[4]

Abstract—In this paper, a 3.5 GHz microstrip patch antenna using three different substrates materials with varying relative permittivity have been designed. However, the thickness of the substrates are slightly different from each other which is 1.6 mm for FR-4, 1.575 mm for RT-5880 and 1.58 mm for TLC-30 have been chosen to carry out this work. The three substrates materials are FR-4 (Design-1), RT-5880 (Design-2), and TLC-30 (Design-3) with the relative permittivity of 4.3, 2.2, and 3, respectively. The antennas' performances in terms of reflection coefficient, voltage standing wave ratio (VSWR), bandwidth, gain, and efficiency performance is simulated, analyzed and compared using CST Microwave studio (CST 2019). The findings reveal that there is a significant change in gain and bandwidth due to different relative permittivity and the thickness value of the substrate materials. The gains achieved were at 3.338 dB, 4.660 dB, and 5.083 dB for Design-1, Design-2 and Design-3 respectively. The efficiency of the antennas also showed that TLC-30 gave the best efficiency at 75.70% when compared to FR-4 which was at 60.13% and RT-5880 which was at 61.51% efficiency. All the proposed antennas have a bandwidth above 100 MHz where Design-1 had a bandwidth of 247.1 MHz whilst Design-2 had a bandwidth of 129.7 MHz and finally, Design-3 had a bandwidth of 177.2 MHz.

Keywords—Efficiency; gain; microstrip patch antenna; permittivity; substrates

I. INTRODUCTION

Since the development of First Generation technology: 1G, Second Generation: 2G, Third Generation: 3G, Fourth Generation: 4G and finally, Fifth Generation: 5G is being realized [1-2]. There have been significant and noteworthy improvements with every generation of communication systems. Fifth Generation (5G) technology is designed to offer alternatives to the limitations in the Fourth Generation (4G) technology. The limitations include limited bandwidth and speed. However, the benefits of 5G are shown in Fig. 1. In January 2020, the Malaysian Communications and Multimedia Commission (MCMC) established three bands, which are 700 MHz, 3.5 GHz and 26/28 GHz for the implementation of the 5G in Malaysia [3]. The 3.5 GHz frequency band is widely recognized and has been approved in most countries [4]. Accordingly, this research focuses on 3.5 GHz for the upcoming Fifth Generation (5G) applications.
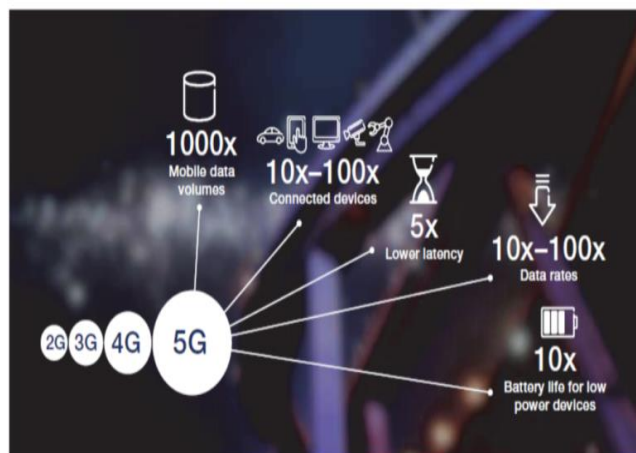
Fig 1. The Benefits of 5G in a Communication System [5].

Fifth Generation (5G) applications require a high-gain antenna to meet the demands of long-distance communications and this means higher data rates. This latest technology can be successfully implemented using a microstrip patch antenna. In general, an antenna is a radiating device developed for the transmission of electromagnetic waves [6]. The purpose of the microstrip patch antenna includes design convenience, low profile and is light weight. Microstrip patch antenna applications are used in a various industry such as medical, telecommunications and military systems. However, the narrow bandwidth and low gains are the greatest drawbacks of the microstrip patch antennas [7]. Nevertheless, the bandwidth can be increased using a thicker substrate which would increase the surface waves that move around the substrate and radiate the patch. As a result, the antenna gains reduce and this could affect the overall performance of the antenna. In addition, dimensions of the patch and feedline influences the performance of the antenna [8]. For microstrip patch antennas, the dielectric value ranges from 2 to 10 but this depends on the application for which the antenna has been designed for [9].

For this research, three microstrip patch antennas with three different dielectric substrates were designed and simulated to operate at 3.5 GHz as part of a 5G application. The antennas were simulated using Computer Simulation

---

*Corresponding Author

Technology (CST 2019) software which is widely recognized as user-friendly software [16] and the performance of the three antennas in terms of reflection coefficient, Voltage Standing Wave Ratio (VSWR), gains, bandwidth and efficiency was measured. The related works of FR-4, RT-5880 and TLC-30 substrate based antenna are discussed in Section II. The methodology and the design of the proposed antennas are discussed in Section III. The theoretical design of the antenna as well as the simulated design is discussed in Section IV. Section V focuses on the results obtained, the contrast of the proposed antennas and the discussion. The conclusion of this study is set out in Section VI.

II. RELATED WORKS

The two key factors that have to be considered when deciding on the antenna substrate are thickness and relative permittivity. These two factors play a vital role in influencing the performance of the antenna [10-15]. Several authors have previously used the FR-4 substrate to design the 3.5 GHz patch antennas. Authors in [10-11] designed a microstrip patch antenna with a permittivity value of 4.4 and a thickness of 1.6 mm. The antenna designed in [10] had achieved a gain of 3 dB and that the gain achieved in [11] was 2.24 dB. With regards to the antenna bandwidth, the antenna designed in [10] had a bandwidth of 300 MHz while the antenna in [11] had a bandwidth of 360 MHz. Based on [12-13], the authors have designed a rectangular patch antenna using RT-5880 as the substrate with a thickness 1.574 mm and 1.57 mm respectively. At 3.8 GHz, the bandwidth of the antenna was 72 MHz [12] and the bandwidth was 50 MHz [13] at 2.5 GHz. The gains achieved in [12] and [13] were 13.2 dB and 5.51 dB respectively. The authors in [14] and [15] have used TLC-30 as the dielectric substrate with a permittivity value of 3 to operate at 3.1 GHz and 2.4 GHz respectively. Nonetheless, the authors in [14] and [15] have used a substrate with a thickness of 1.575 mm and 1.5 mm respectively. In addition, the gains achieved were 2.9 dB and 2.7 dB respectively. However, the information related to bandwidth has not been mentioned in papers [14-15].

III. METHODOLOGY

This paper consists of three microstrip patch antennas using three different dielectric substrates. The specific substrate materials used are FR-4 for Design-1, RT-5880 for Design-2 and TLC-30 for Design-3. The overall objective of this project was achieved by following the steps shown in Fig. 2. The desired parameter is considered significant prior to pre-design, as it has a major impact on the overall performance of the antenna. The key concept was to design three antennas that would operate at 3.5 GHz. The antennas must have a reflection coefficient value (S11) of less than -10 dB, a VSWR of less than or equal to 2 dB, a line impedance match at 50 Ohms and a bandwidth greater than or equal to 100 MHz as shown in Table I below. In addition, series of optimization to achieve the desired results in terms of reflection coefficient and VSWR was conducted. Next, the gains and radiation patterns of the proposed antennas have been analyzed. Lastly, the simulated results of Design-1, Design-2 and Design-3 were compared.
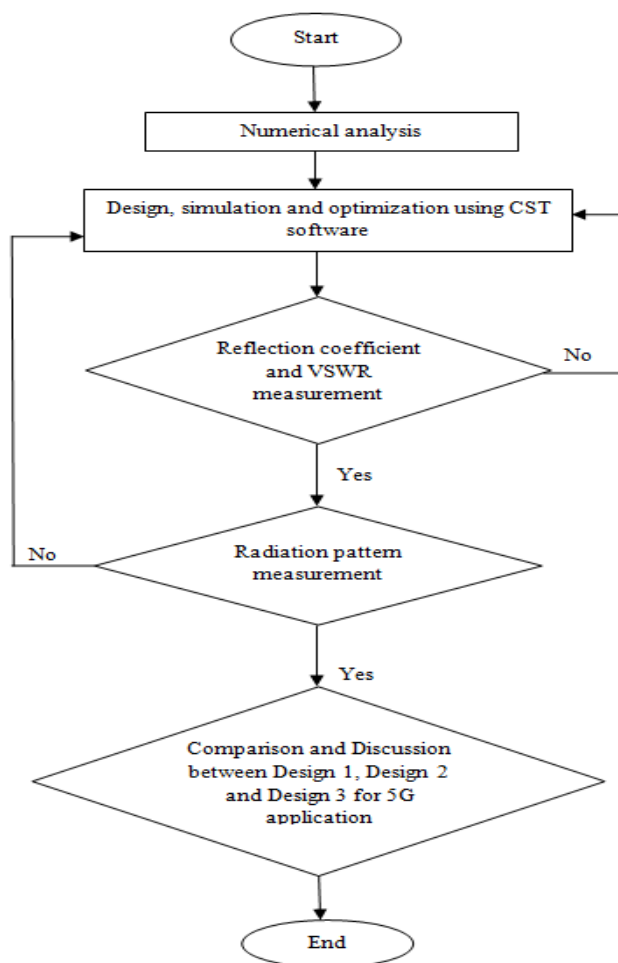
Fig 2. Flowchart for antenna Design-1, Design-2 and Design-3.

TABLE I. DESIGN SPECIFICATIONS OF THE ANTENNAS

Specification	Values
Frequency, f	3.5 GHz
Reflection coefficient, S11	Less than -10 dB
VSWR	1:2
Input Impedance, Z	50 Ohms
Copper thickness, mm	0.035 mm
Bandwidth	≥ 100 MHz

IV. ANTENNA DESIGN

A microstrip patch antenna consists of a radiating patch, a dielectric substrate and a ground plane. The patch is typically made of a conductive material such as copper or gold. Primarily, the substrate is required so as to give the antenna mechanical strength. The ground plane is a flat conductive material which acts as part of the antenna to reflect the radio waves emitted from the other components of the antenna. The basic structure of the microstrip patch antenna is illustrated in Fig. 3. The dimensions of the proposed antennas can be calculated using the formulas (1)-(9) shown below [17].
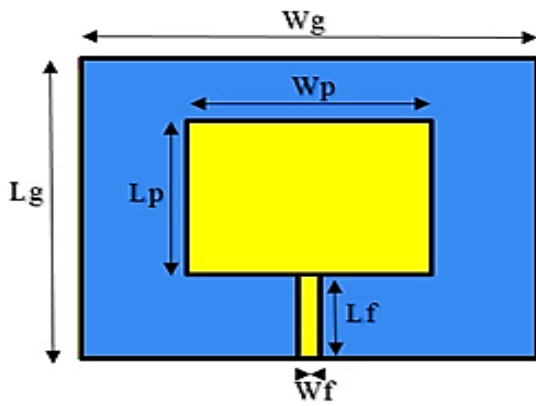
Fig 3.   Basic Dimensions of Microstrip Patch Antenna [6].

The width, Wp and the length,   Lp of the patch are calculated using the equations below:

$$Wp = \frac{c}{2fo\sqrt{\frac{(\varepsilon r+1)}{2}}} \quad (1)$$

Here C $=3\times10^8$ m/s (light speed), εr = permittivity of substrate and f_o = resonant frequency (GHz)

$$Lp = Lreff - 2\Delta L \quad (2)$$

Where L_{reff} can be found using:

$$L_{reff} = \frac{c}{2fo\sqrt{\varepsilon reff}} \quad (3)$$

$$\varepsilon reff = \frac{\varepsilon r+1}{2} + \frac{\varepsilon r-1}{2}\left[1 + 12\frac{h}{w}\right]^{\frac{-1}{2}} \quad (4)$$

Moreover, ΔL can be found using:

$$\Delta L = 0.412h\left(\frac{\varepsilon reff+3}{\varepsilon reff-0.258}\right)\left(\frac{\frac{w}{h+0.264}}{\frac{w}{h+0.813}}\right) \quad (5)$$

The width Wg and length Lg of the ground plane can be found using:

$$Wg = 6h + Wp \quad (6)$$

$$Lg = 6h + Lp \quad (7)$$

here h = height of the substrate.

The width of feed line, Wf:

$$Zo = [87/\sqrt{(\varepsilon r + 1.141)}]\ln(5.98h/0.8Wf) \quad (8)$$

The length of feedline, Lf:

$$Lf = \frac{Lg-Lp}{2} \quad (9)$$

Design-1 used a flame retardant (FR-4) substrate. This dielectric substrate has a relative permittivity of 4.3 and with a thickness of 1.6 mm. Design-2 used the Rogers (RT-5880) which offers a relative permittivity of 2.2 and a standard substrate thickness of 1.575 mm. Design-3 is completed used Taconic (TLC-30) and has a relative permittivity of 3 and a thickness range of 0.51 mm to 6.35 mm [18]. However, for this research work, 1.58 mm was selected to ensure that all proposed antennas were of similar thickness for performance comparison purposes. Using the formulas (1) - (9) above, the

dimensions of the three antennas were calculated. However, the numerical results were not achieved at 3.5 GHz. As a result, a series of optimizations were performed until the resonant frequency is 3.5 GHz was achieved.

The length and width of the ground plane are indicated by "Lg" and "Wg" while the length and width of the patch are represented by "Lp" and "Wp" respectively. The width and length of the feedline are represented by "Wf" and "Lf" respectively. It has been observed that by increasing the patch width of "Wp" and the feedline length of "Lf" led to the achievement of the resonant frequency. In addition, the reduction of patch length "Lp" and feedline width "Wf" enabled Design-1 and Design-2 antennas to achieve resonant frequency. To recapitulate, an increase in the patch area and feedline area has resulted in the desired output being achieved. The calculated and optimized dimensions of the antennas are shown in Table II to Table IV. In addition, the front and back views of Design-1 to Design-3 are shown in Fig. 4.

TABLE II.   CALCULATED AND OPTIMIZED DIMENSION OF DESIGN-1

Design-1 using FR-4 substrate		
Notation	Calculated Value (mm)	Optimized value (mm)
Wp-1	26.31	33.5
Lp-1	20.20	19.35
Wg-1	35.91	36
Lg-1	29.8	29.8
Wf-1	3.175	3
Lf-1	4.8	6.00

TABLE III.   CALCULATED AND OPTIMIZED DIMENSION OF DESIGN-2

Design-2 using RT-5880 substrate		
Notation	Calculated Value (mm)	Optimized value (mm)
Wp-2	35	46.45
Lp-2	26.4	27.6
Wg-2	36.908	47
Lg-2	31.698	31.698
Wf-2	4.8	4.46
Lf-2	1.4	1.9

TABLE IV.   CALCULATED AND OPTIMIZED DIMENSION OF DESIGN-3

Design-3 using TLC-30 substrate		
Notation	Calculated Value (mm)	Optimized value (mm)
Wp-3	30.28	40
Lp-3	24.12	23.46
Wg-3	39.79	42
Lg-3	33.6	33.71
Wf-3	2.75	3.85
Lf-3	4.74	5.45
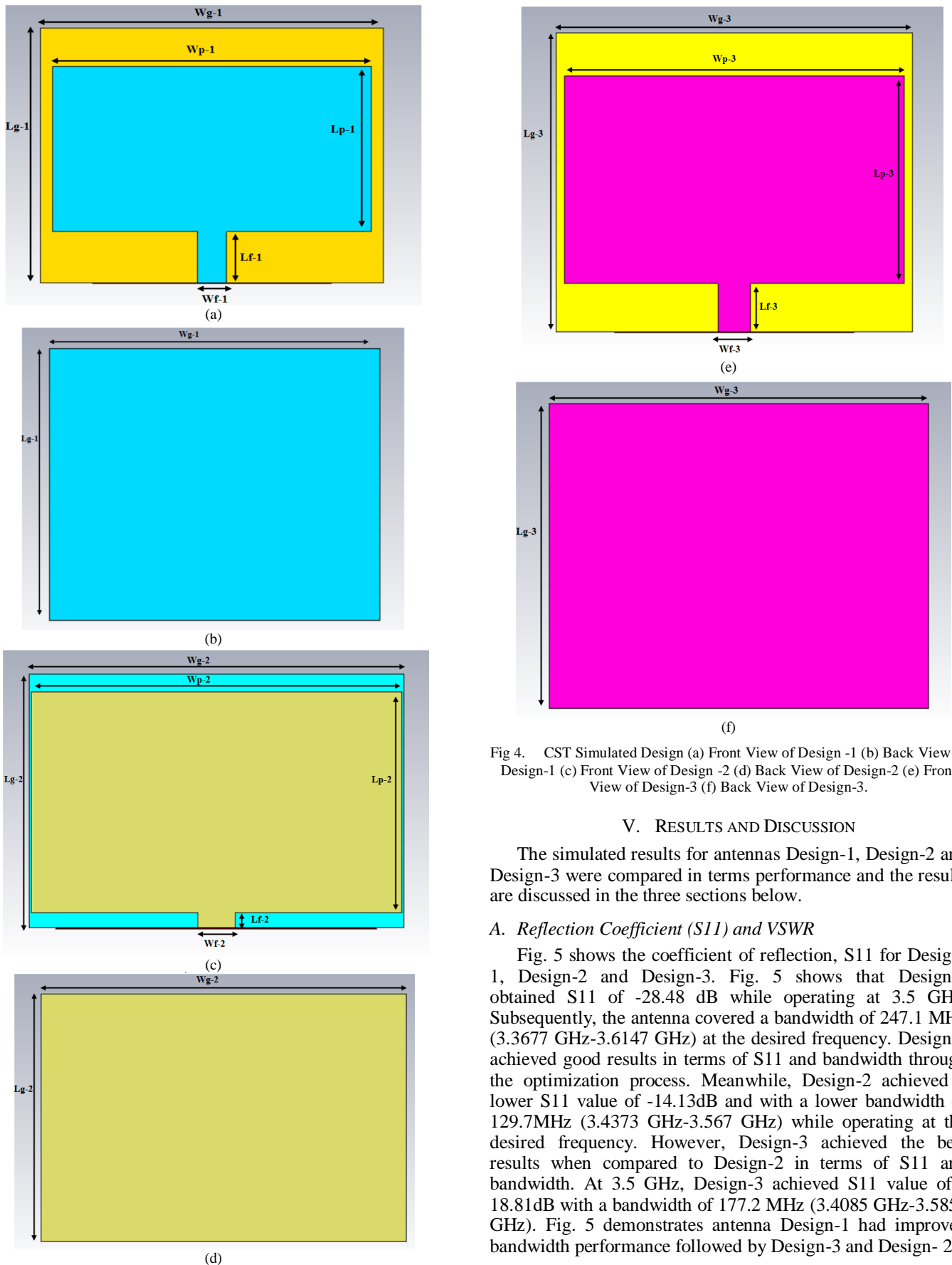
Fig 4. CST Simulated Design (a) Front View of Design -1 (b) Back View of Design-1 (c) Front View of Design -2 (d) Back View of Design-2 (e) Front View of Design-3 (f) Back View of Design-3.

V. RESULTS AND DISCUSSION

The simulated results for antennas Design-1, Design-2 and Design-3 were compared in terms performance and the results are discussed in the three sections below.

A. Reflection Coefficient (S11) and VSWR

Fig. 5 shows the coefficient of reflection, S11 for Design-1, Design-2 and Design-3. Fig. 5 shows that Design-1 obtained S11 of -28.48 dB while operating at 3.5 GHz. Subsequently, the antenna covered a bandwidth of 247.1 MHz (3.3677 GHz-3.6147 GHz) at the desired frequency. Design-1 achieved good results in terms of S11 and bandwidth through the optimization process. Meanwhile, Design-2 achieved a lower S11 value of -14.13dB and with a lower bandwidth of 129.7MHz (3.4373 GHz-3.567 GHz) while operating at the desired frequency. However, Design-3 achieved the best results when compared to Design-2 in terms of S11 and bandwidth. At 3.5 GHz, Design-3 achieved S11 value of -18.81dB with a bandwidth of 177.2 MHz (3.4085 GHz-3.5857 GHz). Fig. 5 demonstrates antenna Design-1 had improved bandwidth performance followed by Design-3 and Design- 2.
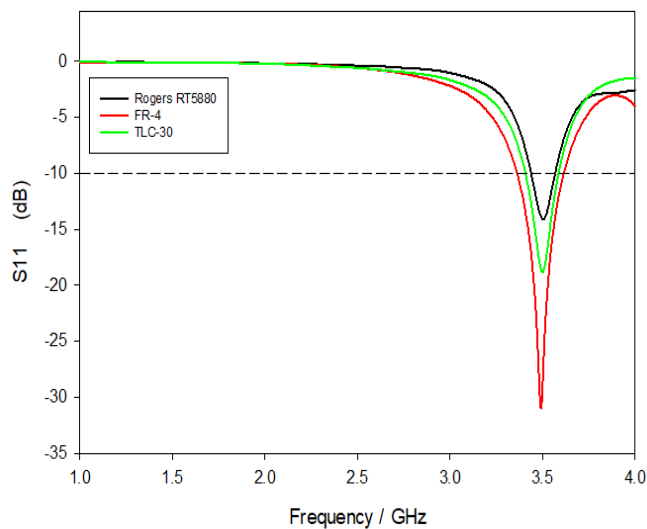
Fig 5.   Reflection Coefficient of Design-1, Design-2 and Design-3.
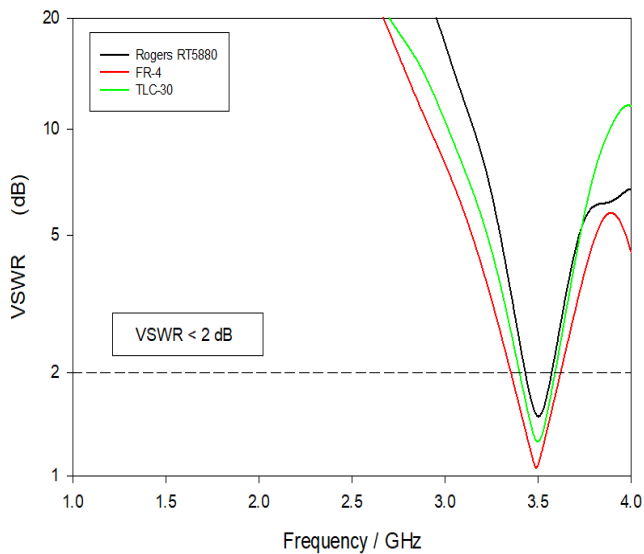
Fig 6.   VSWR of Design-1, Design-2 and Design-3.

Fig. 6 shows the VSWR for Design-1, Design-2 and Design-3. According to the ITU standards for 5G mobile communication system application [19], all three of the proposed antennas are eligible for 5G applications as the VSWR is less than 2. The VSWR for Design-1, Design-2 and Design-3 were 1.078, 1.48 and 1.259, respectively.

B. Radiation Pattern of the Proposed Antennas

Further investigation was conducted in relation to the radiation pattern of the proposed antennas using CST software. In the design and analysis of the antenna, the radiation pattern is considered to be one of the major aspects since it has a direct influence on the performance of the antenna. Fig. 7 depicts the three dimensional (3D) view, together with the gains achieved by the proposed Design-1, Design-2 and Design-3 respectively. The maximum gain of 5.083 dB was achieved by Design-3 while operating at the desired frequency of 3.5 GHz. Subsequently, Design-1 and Design-2 achieved gains of 3.338 dB and 4.660 dB at the

desired frequency respectively. In addition, the radiation pattern in the two dimensional (2D) view is shown in Fig. 8. All the three antennas have a reasonable linear directional behavior, which means that the antennas can cover long ranges in one particular direction.

(a)

(b)

(c)
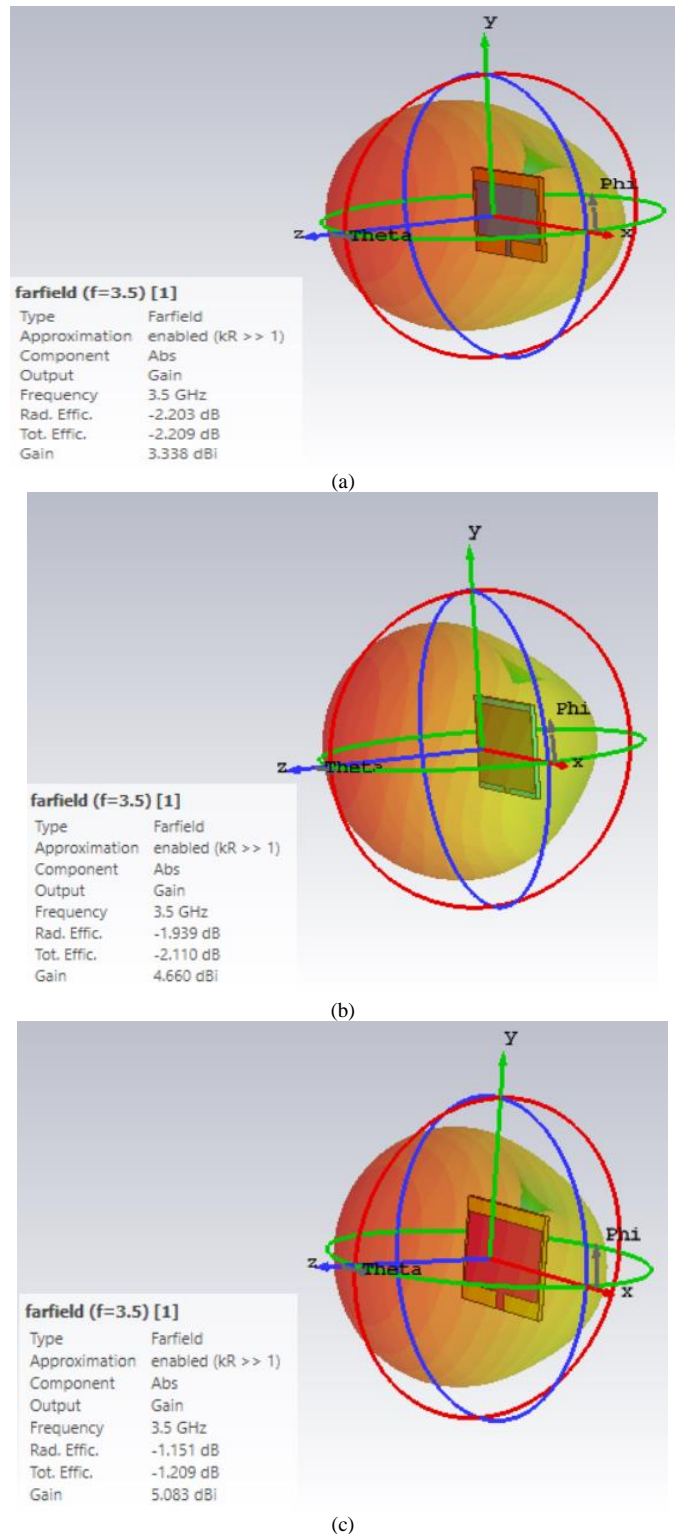
Fig 7.   Three Dimensional (3D) view of the Radiation Pattern (a) Design-1
(b) Design-2 (c) Design-3.

(a)

(b)

(c)
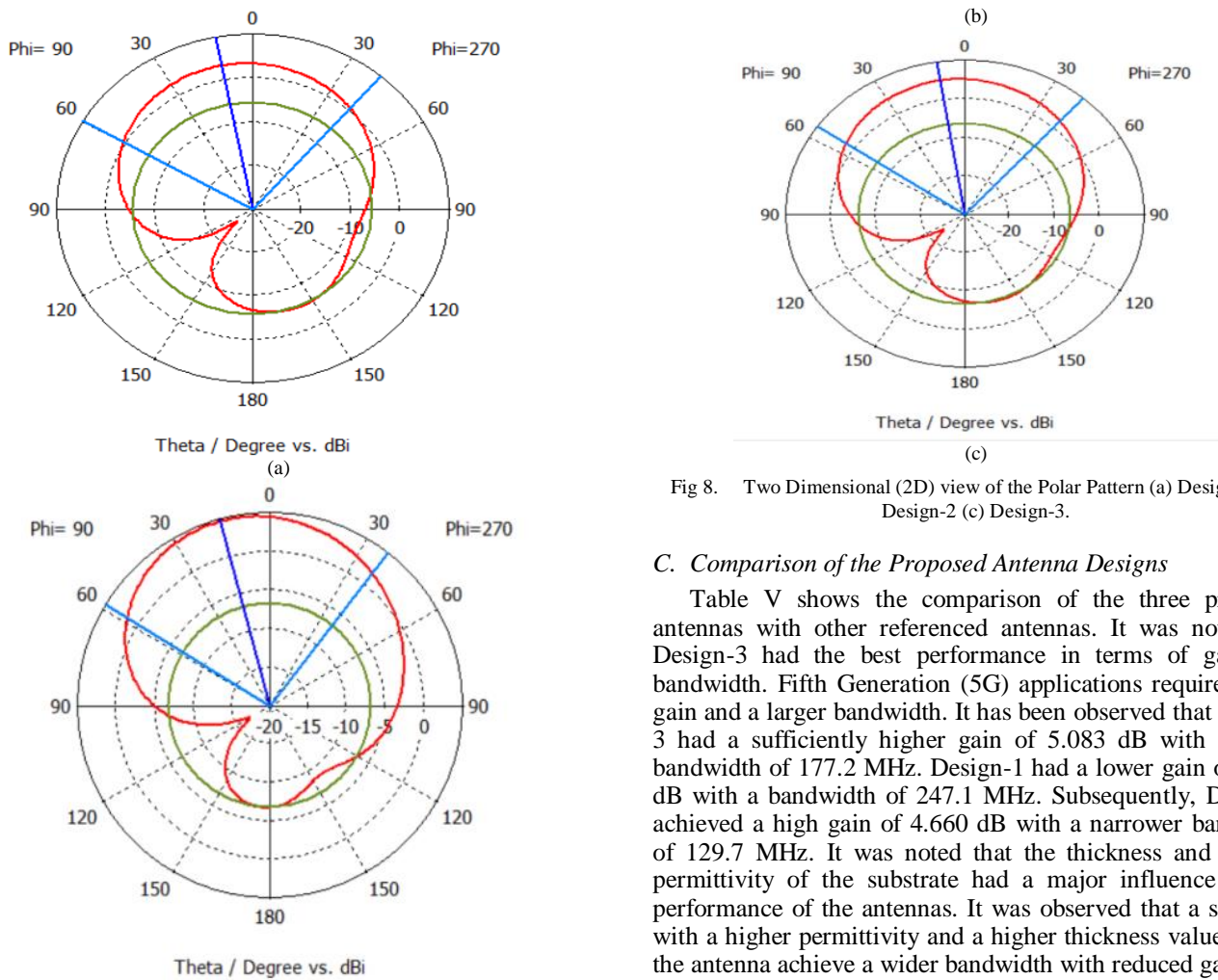
Fig 8. Two Dimensional (2D) view of the Polar Pattern (a) Design-1 (b) Design-2 (c) Design-3.

C. Comparison of the Proposed Antenna Designs

Table V shows the comparison of the three proposed antennas with other referenced antennas. It was noted that Design-3 had the best performance in terms of gain and bandwidth. Fifth Generation (5G) applications require higher gain and a larger bandwidth. It has been observed that Design-3 had a sufficiently higher gain of 5.083 dB with a wider bandwidth of 177.2 MHz. Design-1 had a lower gain of 3.338 dB with a bandwidth of 247.1 MHz. Subsequently, Design-2 achieved a high gain of 4.660 dB with a narrower bandwidth of 129.7 MHz. It was noted that the thickness and relative permittivity of the substrate had a major influence on the performance of the antennas. It was observed that a substrate with a higher permittivity and a higher thickness value helped the antenna achieve a wider bandwidth with reduced gains.

TABLE V. COMPARISON OF THE PROPOSED ANTENNAS WITH REFERENCED ANTENNAS

Reference	Dielectric substrate	Frequency, GHz	S11, dB	VSWR	Gain ,dB	Bandwidth, MHz	Patch Area (l × w), mm^2
[10]	FR-4	3.5	-29.5	1	3	300	2500
[11]	FR-4	3.5	< 10	≤ 2	2.24	360	1140
Design-1	FR-4	3.5	-28.48	1.078	3.338	247.1	648.225
[12]	RT-5880	3.8	-18.2	≤ 2	13.2	72	792.48
[13]	RT-5880	2.5	-12.105	1.4871	5.51	50	1571.92
Design-2	RT-5880	3.5	-14.13	1.48	4.660	129.7	1282.02
[14]	TLC-30	3.1	<10	≤ 2	2.9	-	182
[15]	TLC-30	2.4	<10	≤ 2	2.7	-	876 (rectangle + triangle)
Design-3	TLC-30	3.5	-18.81	1.259	5.083	177.2	938.4

Due to technological advancement, an antenna has to be small and be equipped with a wider bandwidth and higher gain. From the simulated results, it can be seen that Design-1 had a much smaller patch dimension of 648.225mm^2 compared to Design-2 with a dimension of 1282.02 mm^2 and finally, Design-3 with a dimension of 938.4 mm^2. This indicates that a smaller patch helped to achieve a wider bandwidth of 247.1 MHz. However, Design-3 achieved the highest gain of 5.08 dB compared to Design-1 gains 3.338 dB while Design-2 achieved gain of 4.660 dB. Design-2 had a narrow bandwidth which was advantageous for cancelling unwanted signals and was capable of transferring maximum energy. From the comparisons above, the proposed antennas can be deployed for use in Fifth Generation (5G) applications as they had performed better in terms of bandwidth and gain.

VI. Conclusion

The assessment of the results and the comparison between substrate materials FR-4, RT-5880 and TLC-30 has been extensively studied. The aim of the proposed antenna designs was to achieve good performance in terms of gain and bandwidth while maintaining a reflection coefficient below -10dB and VSWR ≤ 2. All the proposed antennas achieved good performance (higher gains and bandwidth ≥100 MHz) while meeting the reflection coefficient and VSWR requirements at 3.5 GHz. Moreover, Design-1, Design-2 and Design-3 have an efficiency of 60.13%, 61.51% and 75.70% at the desired frequency. This indicated that TLC-30 would be the best choice for 5G applications among the three proposed antennas. Further studies would need to be carried out to improve the efficiency of the proposed antennas, in particular to improve the gains and bandwidth while having a smaller antenna dimension.

Acknowledgment

References

[1] Zhang H, Dong Y, Cheng J, Hossain MJ, Leung VCM,"Fronthauling for 5G LTE-U ultra dense cloud smallcell networks," IEEE Wireless Commun. 2016. vol. 23(6), pp 48-53.

[2] S. Path, "A Straight Path Towards 5G," Straight Path Communications Inc, 2015, pp. 1–29.

[3] "Allocation of spectrum bands for mobilebroadband service in Malaysia," Malaysian Communications and Multimedia Commission, January 2020, pp 1-2.

[4] "Roadmap for C-band spectrum in ASEAN," GSMA, August 2019, pp 35-37.

[5] Olga Boric-Lubecke, Victor M. Lubecke, Branka Jokanovic, Aditya Singh, Ehsaneh Shahhaidar, and Bryson Padasdao, "Microwave and Wearable Technologies for 5G", 12th International Conference on Telecommunication in Modern Satellite, Cable and Broadcasting Services (TELSIKS), 14-17 October 2015.

[6] K. W. S. Al Kharusi, N. Ramli, S. Khan, M. T. Ali and M. H. Abdul Halim, "Gain Enhancement of Rectangular Microstrip Patch Antenna using Air Gap at 2.4 GHz," International Journal of Nanoelectronics and Materials Volume 13 (Special Issue), pp 211-224 (2020).

[7] Rajveer S Yaduvanshi, Harish Parthasarathy and Asok De, "Magneto-Hydrodynamic Antenna Design and Development Analysis with prototype," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No.2, February 2011.

[8] Nagapushpa K.P and Chitra Kiran N, "Studying Applicability Feasibility of OFDM in Upcoming 5G Network," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 8, No. 1, 2017.

[9] Zain Ul Abedin and Zahid Ullah, "Design of a Microstrip Patch Antenna with High Bandwidth and High Gain for UWB and Different Wireless Applications," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017.

[10] Archana R.Cheekatla and Pankaj S. Ashtankar, "Compact micro strip antenna for 5g mobile phone applications," International Journal of Applied Engineering Research. vol 14(2), pp.108-111, 2019.

[11] J. Ashish and A. P. Rao, "Design and Implementation of Compact Dual band U-slot Microstrip Antenna for 2.4GHz WLAN and 3.5GHz Wi-MAX Applications," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India. pp. 1084-1086.

[12] H. Sajjad, W. T. Sethi, K. Zeb and A. Mairaj, "Microstrip patch antenna array at 3.8 GHz for WiMax and UAV applications," 2014 International Workshop on Antenna Technology: Small Antennas, Novel EM Structures and Materials, and Applications (iWAT), Sydney, NSW. pp. 107-110.

[13] Jahariah Sampe , Noor Hidayah Mohd Yunus, Jumril Yunas, Alipah Pawi and Zeti Akma Rhazali, " Design and fabrication of a dual band 1.8/2.5 GHZ antenna for RF energy harvester" International Journal of Engineering & Technology. Vol. 7(4).

[14] K. Thana Pakkiam and J. S. Mandeep, "Design of multiband parallel strip patch antenna for DCS/WLAN/WIMAX and RFID applications," 2016 IEEE Industrial Electronics and Applications Conference (IEACon), Kota Kinabalu, 2016, pp. 348-353.

[15] K. T. Pakkiam, J. S. Mandeep and M. T. Islam, "Design of microstrip antenna for modern wireless communication," International Symposium on Telecommunication Technologies, Kuala Lumpur, 2012, pp. 42-46.

[16] H. Yon , N. H. Abd Rahman, M. A Aris and H. Jumaat, "Developed high gain microstrip antenna like microphone structure for 5G application," International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 3, June 2020, pp. 3086-3094.

[17] C. A. Balanis, "Antenna theory: a review," Proc. IEEE, vol. 80, no. 1,1992, pp. 7–23.

[18] "TLC" datasheet published by TACONIC, www.taconic-add.com (accessed on 25.05.2020)

[19] D. Andreev, "Overview of ITU-T activities on 5G/IMT-2020," Int. Telecommun. Union, 2017.

Monopole Antenna on Transparent Substrate and Rectifier for Energy Harvesting Applications in 5G

S. M. Kayser Azam[1], Md. Shazzadul Islam[2]
Department of Electrical and Computer Engineering
International Islamic University Malaysia
Kuala Lumpur, Malaysia

A. K. M. Zakir Hossain[3]
Fakulti Teknologi Kejuruteraan Elektrik & Elektronik
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

Mohamadariff Othman[4]
Department of Electrical Engineering
University of Malaya
Kuala Lumpur, Malaysia

Abstract—**In line with the harvested energy required for the emerging 5G technology, this article proposes a planar monopole antenna and a rectifier. The proposed Coplanar Waveguide (CPW)-fed antenna is printable on a transparent Poly-Ethylene Terephthalate (PET) substrate. The antenna has a center frequency at 3.51 GHz within a bandwidth of 307 MHz that covers the pioneer 5G band in Malaysia. The designed omnidirectional antenna exhibits the maximum gain of 1.51 dBi with a total efficiency of 95.17 percent. At the antenna frequency, a rectifier has been designed with the voltage doubler technique on a Rogers RO3003 substrate. Over an input RF power of 0 dBm, the rectifier has a power conversion efficiency around 42 percent. The proposed antenna rejects harmonics at least until 16 GHz frequency that makes it compatible with the rectifier to eliminate an additional bandpass filter or impedance matching network from the energy harvesting system.**

Keywords—*Monopole antenna; transparent substrate; 5G; energy harvesting; rectifier*

I. INTRODUCTION

The modern communication system is moving forward to employ different applications of antennas other than only for communication purposes. For example, antennas for wearable devices, industrial and biomedical applications [1], [2] are very popular nowadays. Meanwhile, the emerging 5G technology facilitates the Internet of Things (IoT) [3] based on numerous sensor devices that require a continuous intercommunication and remote power supply at the same time. Obviously, batteries are not the only choice for empowering numerous sensor devices; power can be supplied by other sources like solar cells. In terms of remote power supply, batteries are in fact not the best choice because they need regular maintenance and replacement. Whereas, solar cells do not benefit when the sunlight is absent, or devices remain remotely indoor. On the other hand, the ambient Radio Frequency (RF) sources are always available even in the remote corners. Therefore, antennas with rectifiers [4] have drawn a special attention as energy harvesters.

In the recent times, several research works have been introduced with antennas and rectifiers for RF energy harvesting applications. Mostly for Wireless Local Area Network (WLAN) and Industrial, Scientific, and Medical (ISM) bands, RF energy harvesters have been proposed as in [5]–[8]. The RF to DC Power Conversion Efficiency (PCE) is increased by following the hierarchically similar technique of tapered matching network. In [9], a voltage doubler circuit is adopted by using Schottky diodes that are often utilized in harvesting circuits. Nevertheless, improving the PCE is a major challenge while multiple diodes are used. In [10]–[14], multi-layered aperture-coupled antennas are designed and incorporated with interdigital capacitor-based voltage doubler circuits for energy harvesting applications. Some remarkable features of harmonic suppression, dual-band operation etc. are introduced through those techniques in [10]–[14]. However, the amount of input power needed to obtain a high PCE with sufficient DC output voltage in those works is not compatible for remote sensors and devices since the ambient RF power normally available in the environment is not large in amount

Antennas and rectifiers are incorporated together for energy harvesting applications [15] in frequencies other than ISM and WLAN bands. A circular loop antenna with linear polarization is utilized in [15] on an optically transparent poly-methyl methacrylate substrate at 940 MHz, 1.86 GHz, 2.14 GHz, and 2.49 GHz frequencies. However, designed antennas are large in dimension despite their simple structures and transparent substrates. Again, the single element shunt topology-based rectifier has limited PCEs even at the lower frequencies. Whereas, the RF power is generally captured more at low frequencies than that at high frequencies [15] because the effective area is larger at low frequencies. Also, at high frequencies, the harvested power is limited by the additional dielectric loss and the parasitic loss added by rectifying diodes. These issues can be addressed by the harmonic suppression ability [4] of antennas and rectifiers. Bandpass filters and further matching networks in typical energy harvesting systems can significantly decrease PCE and remote applicability.

To address these issues, in this article, a Planar Monopole Antenna (PMA) and a rectifier have been proposed for the pioneer 5G band in Malaysia. The antenna is designed on a transparent PET substrate that makes it suitable for 5G and IoT

applications by allowing the sunlight to the outdoor solar cells and the wearable feature to the indoor devices. The designed antenna suppresses harmonics at least up to 16 GHz frequency. The rectifier needs no additional impedance matching network. Therefore, the antenna with the rectifier can eliminate the need of an additional bandpass filter or impedance matching network from the energy harvester. Section II includes the antenna design, Section III demonstrates the rectifier design, Section IV discusses the results, and Section V ends the article.

II. ANTENNA DESIGN

In the PMA design, the CPW feeding technique has been adopted for a circular patch. In a conventional CPW-fed circular patch antenna, the radiating patch has the circular shape and the feedline is surrounded by the coplanar ground with a narrow spacing. The coplanar ground is typically large in dimension at both sides of the feedline. As a result, a high inductive path is formed on the coplanar ground. Also, the circular radiating patch is not surrounded by the ground. This results in wideband characteristics in the CPW-fed circular patch antenna. In this design, unlike the conventional circular patch CPW-fed monopole antenna, the coplanar ground is not large in size and the radiating patch is surrounded by a circular thin ground line. The design technique is illustrated in Fig. 1.

The coplanar ground has been designed in a way so that only a thin circular line is maintained as ground around the radiating patch. The antenna is symmetrically designed; thus, the circular ground line is joined to the truncated ground lines at the bottom through the thin strip-line along the feedline. The strip-line and the circular line have the same width of w_s, whereas, the strip-line has a length of S_L at both side of the feedline that has a width of W_F and a length of L_F. Whereas, each rectangular ground line has a width of W_G and a length of L_G. The circular radiating patch has a radius of r that is less than both the inner radius (R) and outer radius (R_{out}) of the circular ground line. The spacing between the radiating patch and the circular ground line is equally maintained by $R - r = \Delta r$ while $R_{out} - w_s = R$. The PMA has a length of L and a width of W while the gap between the strip-line and the feedline is g. Dimensional parameters of the proposed antenna are provided in Table I.
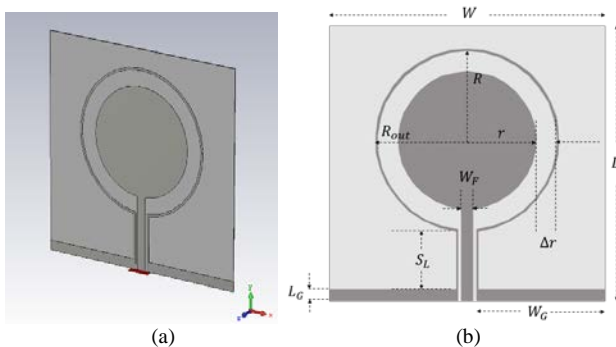
Fig 1.   Proposed CPW-fed PMA (a) 3D View in Design Environment (b) Dimensional Indications.

TABLE I.   PROPOSED ANTENNA DIMENSIONS

Name	W, L	w_s, g	r	L_F, R_{out}	Δr	W_F	S_L	L_G	W_G
Unit (mm)	24	0.25	6	8	1.75	1	4.9	1.1	11.25

TABLE II.   PET SUBSTRATE PROPERTIES

Optical Transparency (%)	Relative Permittivity, ε_r	Loss Tangent, $\tan \delta$	Substrate Height, h (mm)
90.4	3.2	0.002	0.1

The specialty of the proposed antenna is its truncated and thin coplanar ground lines surrounding the radiating patch and the feedline. It is known that conventional CPW-fed circular patch antennas have wideband characteristics. In case of the proposed antenna, the truncated and thin ground line significantly reduces the overall inductance. Such a reduced inductive characteristic results in narrowing the antenna bandwidth. At the same time, increasing the spacing between the radiating patch and the circular line of the ground decreases the overall mutual capacitance. As a result, chances of resonance at higher frequencies become less that can help the antenna to suppress harmonics. The center frequency is stably maintained by a tight coupling between the feedline and its adjacent coplanar ground lines. The antenna has been designed on an optically transparent Poly-Ethylene Terephthalate (PET) substrate by using CST Microwave Studio 2019. The designed antenna is printable and especially recommended for inkjet printing process for its fabrication. The importance of selecting the optically transparent substrate has already been discussed earlier. In addition, the low-cost PET substrates are extensively available in the market. Also, PET materials are flexible, water resistant and atmospheric-heat sustainable. Important properties of PET substrate [16] is given on following Table II.

III. RECTIFIER DESIGN

In conventional RF energy harvesting systems, as shown in Fig. 2(a), antennas are first connected to bandpass filters for harmonic suppression [4] so that the rectifying circuit constructed by non-linear diodes are protected from damage. Antenna with the bandpass filter is then connected to an impedance matching network for joining the rectifier. Lastly, a lowpass filter is followed by the rectifier to extract only the rectified DC power at the output terminal. However, there is a chance of power loss for the rectifier when the input RF signal pass through passive components like bandpass filters and impedance matching networks, especially at the high frequency. In fact, bandpass filters often suffer from high insertion loss [17]–[19] which is a major cause of low PCE for the rectifier. Thus, a rectifier has been proposed in this work which in and of itself can serve the purpose of bandpass filters and impedance matching networks. Hence, the proposed rectifier reduces components from the RF energy harvesting system as depicted in Fig. 2(b).

(a)

(b)
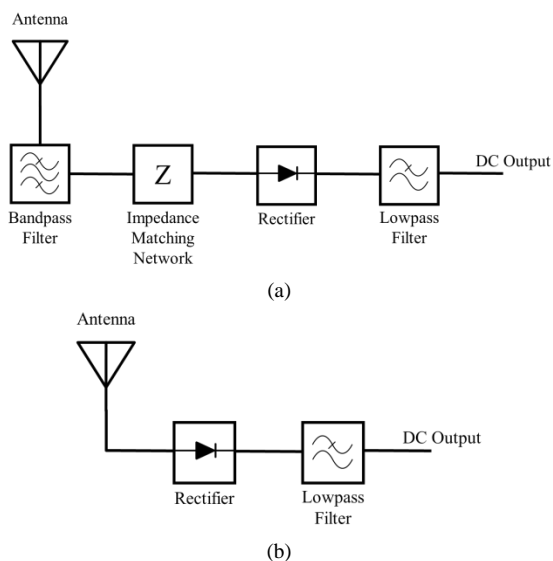
Fig 2.   RF Energy Harvesting System (a) Conventional (b) Proposed.

For the proposed rectifier, two Schottky diodes (SMS7630-079LF) and two capacitors (1000 pF) have been utilized on a Rogers RO3003 substrate ($\varepsilon_r = 3$, $\tan\delta = 0.001$, $h = 0.51$ mm). Voltage doubler circuit topology has been adopted to design the rectifier circuit since this topology offers a better tradeoff between the output voltage and overall PCE than that of other rectifier topologies. The rectifier has been designed by using Advanced Design System (ADS) 2017 for both schematic and layout design. Schematic circuit, electromagnetic (EM) co-simulation set up and final layout of the proposed rectifier are demonstrated in Fig. 3.

(a)

(b)

(c)
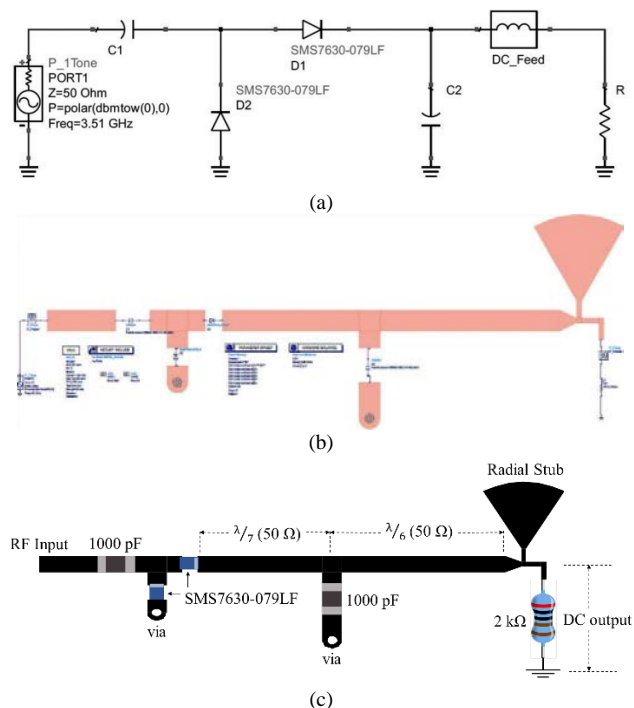
Fig 3.   Proposed Rectifier Design (a) Schematic Circuit (b) EM Co-Simulation Set Up (c) Final Layout.

Input terminal and core of the rectifier have been designed by employing the unequal distribution of microstrip 50 Ω transmission line. The smaller microstrip 50 Ω line is followed by the larger microstrip 50 Ω line connected to the output terminal. In between them, the grounding capacitor is placed. Both lines are shorter than the quarter-wave line; the smaller one is constructed by a $\lambda/7$ and the larger one is constructed by a $\lambda/6$ length of microstrip 50 Ω transmission line. This reduces the unwanted radiation loss of RF signal. At the output terminal, a tapered line has been utilized to facilitate a fine impedance matching between the 50 Ω lines and high impedance (100 Ω) output lines since tapered transmission lines are useful for preventing sudden impedance mismatch [20]. From output terminal, two high impedance lines have been designed – one for connecting the output resistance and another for employing a radial stub that serves the purpose of a lowpass filter. The rectifier layout is complete after connecting a 2 kΩ output resistance to extract the output DC power. The overall PCE of the designed rectifier can be obtained [5] from the following equation.

$$PCE = \frac{P_{DC}}{P_{RF}} \times 100 \ (\%) \tag{1}$$

Here, for the designed rectifier, the DC power collected from the output terminal is denoted as P_{DC} while the RF power received by the antenna is represented by P_{RF}.

IV. RESULTS AND DISCUSSION

The designed antenna has been simulated over a wide range of frequency from 2 GHz to 16 GHz by using CST Microwave Studio 2019. Impedance matching and radiation pattern of the antenna have been observed for the proposed antenna. On the other hand, ADS 2017 electromagnetic co-simulation has been performed for the final layout of the rectifier by employing the replicating models of Schottky diodes and capacitors as per the datasheet and toolkit provided by the corresponding manufacturers. Impedance matching of the rectifier has been observed to find out whether the rectifier bandwidth fits within the antenna bandwidth or not. Responses of the rectifier have been obtained analyzed in terms of its DC output voltage and PCE with respect to the change of RF input power and load resistance.

A. Antenna

One of the significant attributes of the proposed antenna is its ability to suppress the higher harmonics. S-parameter is observed to justify whether the antenna suppresses the harmonics or not. As presented in Fig. 4(a), the return loss of the antenna shows that no resonance is found below -10 dB until 16 GHz frequency. Hence, harmonic suppression has become an intrinsic characteristic of the proposed antenna which further helps the rectifier to eliminate the need of bandpass filters and impedance matching network from the RF energy harvesting system. This is one of the unique features of the proposed antenna. The minimum return loss is observed as -39.61 dB at the center frequency of 3.51 GHz with a -10 dB impedance bandwidth of 307 MHz. The stability of antenna impedance matching is further justified by observing its Voltage Standing Wave Ratio (VSWR) as shown in Fig. 4(b). The minimum VSWR is found as 1.02 at the center frequency.

(a)

(b)
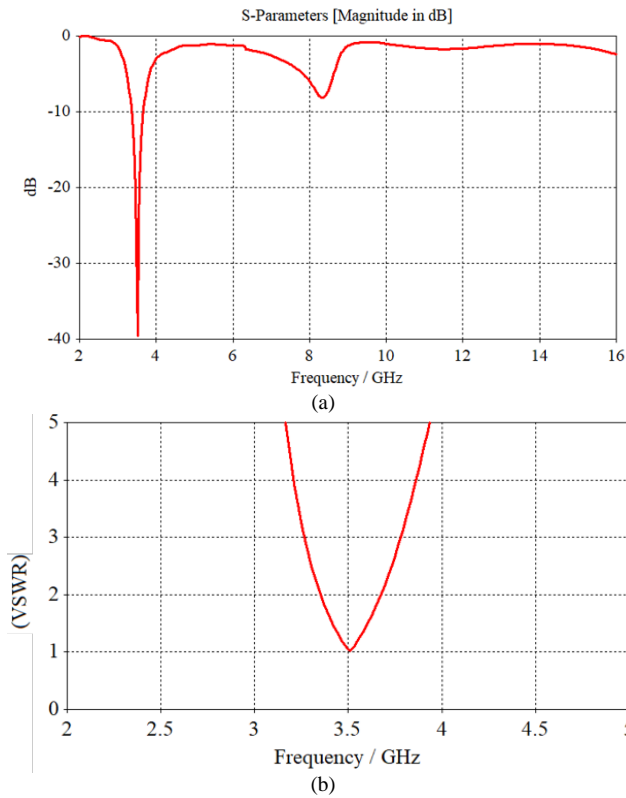
Fig 4.   Impedance Matching of the Antenna (a) Return Loss (b) VSWR.

According to the antenna frequency and bandwidth, it is evident that the pioneer 5G band in Malaysia suitably fits with the designed antenna. The proposed antenna offers a maximum gain of 1.51 dBi which is satisfactory on a transparent and flexible substrate in the 3.5 GHz frequency range for energy harvesting applications. Since the antenna is designed by truncating a significant amount of conductive area, the antenna can exhibit such a level of gain. The designed antenna has a typical monopole omnidirectional radiation pattern that is illustrated in Fig. 5 for its center frequency.

Such an omnidirectional radiation pattern is a useful feature of the proposed antenna to collect the ambient RF energy from any direction.

(a)   (b)
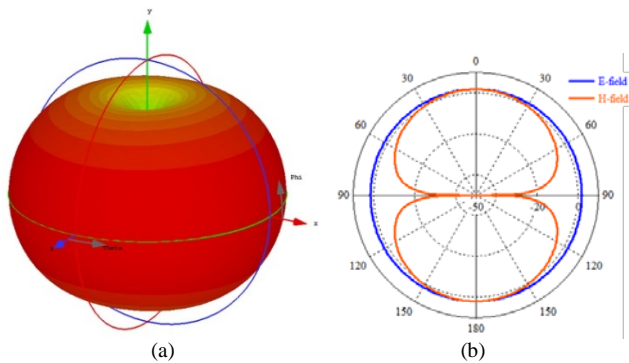
Fig 5.   Radiation Pattern of the Antenna (a) 3D View (b) Polar Plot.

B. Rectifier

Return loss (S_{11} parameter) of the rectifier is presented in Fig. 6. A minimum return loss of -23.35 dB is found at the rectifier center frequency of 3.51 GHz.

The -10 dB impedance bandwidth of the proposed rectifier is 59 MHz which lies within the 100 MHz bandwidth of 5G band in Malaysia. This is an indication that the proposed antenna with the rectifier can perform a simultaneous operation of RF energy harvesting in the 5G frequency band and data communication in other bands like WiMAX.

It is common in RF rectifiers that the output DC voltage varies with the different input RF power. In Fig. 7, output response of the rectifier is presented for different input power.
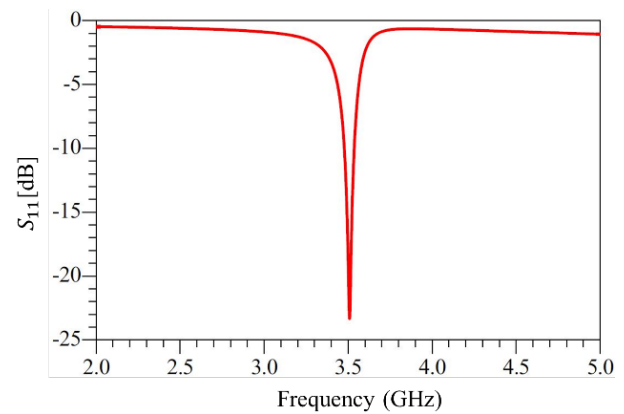
Fig 6.   Return Loss of the Proposed Rectifier.

(a)

(b)
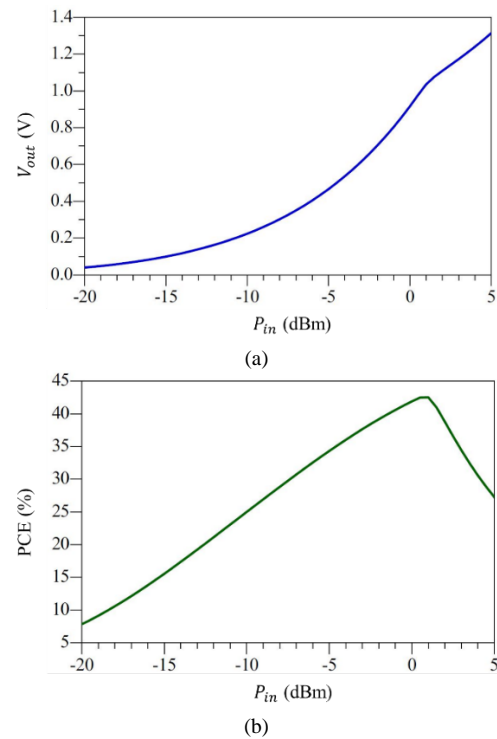
Fig 7.   Rectifier Output for Different Input Power (a) V_{out} (b) PCE.

(a)

(b)
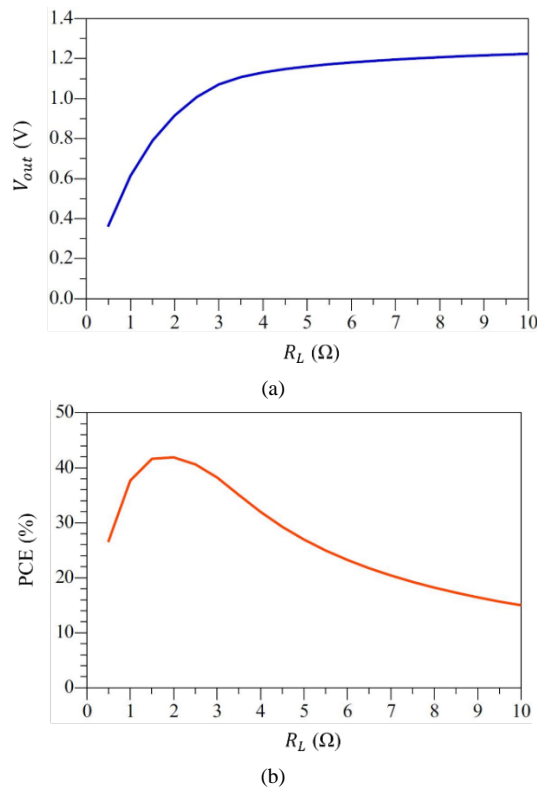
Fig 8.   Rectifier Output for Different Load Resistance (a) V_{out} (b) PCE.

As illustrated in Fig. 7(a), DC output voltage is 1.313 V when 5 dBm input power is applied. For 0 dBm input power, the output voltage becomes 0.915 V. For the usual range of input RF power from −10 dBm to −5 dBm, the output DC voltage varies from 224 mV to 466 mV that can drive low-power sensor devices for the 5G technology. In fact, for the variation of RF input power, DC output voltage changes approximately with a linear trend within the range. This can significantly save sensor devices from damages due to sudden fluctuations in the DC voltage. As observed from Fig. 7(b), the highest PCE is found around 42% for the optimum input RF power of 0 dBm. Whereas, in case of usual range of input RF power, PCE varies from 24% to 34% which is satisfactory enough for RF energy harvesting applications in 5G frequencies.

Stability of DC output voltage due to the change of load resistance is another important aspect of the designed rectifier. Fig. 8 depicts the rectifier response with different load resistances.

Unlike many conventional rectifiers, with different load resistances, DC output voltage of the proposed rectifier is quite stable as shown in Fig. 8(a). Over a large variation of the output resistance (from 1 kΩ to 10 kΩ), the DC output voltage changes from 0.614 V to 1.224 V. This stability of DC output voltage in contrast of output resistance variation is very useful for providing a consistent amount of voltage to different nodes of a sensor network simultaneously. In Fig 8(b), variation of PCE is shown for different load resistances. With resistances from 1 kΩ to 10 kΩ, it is evident that PCE does not change dramatically, rather, gradually decreases to not less than 15%.

PCE is found highest (around 42%) when 2 kΩ load resistance is connected at the output terminal. In terms of physical size, the rectifier is small with an area of 3.25×1.25 cm^2.

V.   CONCLUSION

A printable planar monopole antenna with harmonic suppressing feature has been introduced in this work. The omnidirectional antenna has been designed on a transparent PET substrate that allows the sunlight on outdoor and flexibility on indoor conditions for the IoT and 5G applications. A voltage doubler topology-based rectifier has been proposed with the antenna for eliminating the use of bandpass filters and separate impedance matching networks in RF energy harvesting systems. The proposed antenna and the rectifier are expected to be applied on a 5G-based practical energy harvesting system in future.

REFERENCES

[1]   M. S. Islam, M. I. Ibrahimy, S. M. A. Motakabber, A. K. M. Z. Hossain, and S. M. K. Azam, "Microstrip patch antenna with defected ground structure for biomedical application," Bull. Electr. Eng. Inform., vol. 8, no. 2, pp. 586–595, 2019.

[2]   M. S. Islam, M. I. Ibrahimy, S. M. A. Motakabber, and A. K. M. Z. Hossain, "A Rectangular Inset-Fed Patch Antenna with Defected Ground Structure for ISM Band," in 2018 7th International Conference on Computer and Communication Engineering (ICCCE), 2018, pp. 104–108.

[3]   M. S. Islam, M. I. Ibrahimy, S. M. A. Motakabber, A. K. M. Z. Hossain, and S. M. K. Azam, "A Wideband Millimeter-Wave Printable Antenna on Flexible Substrate for Breast Cancer Imaging," presented at the 2019 7th International Conference on Mechatronics Engineering (ICOM), Kuala Lumpur, Malaysia, Oct. 2019, [Online]. Available: https://www.researchgate.net/publication/337936173_A_Wideband_Mil limeter-Wave_Printable_Antenna_on_Flexible_Substrate_for_Breast_Cancer_I maging.

[4]   S. K. Divakaran and D. D. Krishna, "RF energy harvesting systems: An overview and design issues," Int. J. RF Microw. Comput.-Aided Eng., vol. 29, no. 1, p. e21633, 2019.

[5]   A. Eid, J. G. Hester, J. Costantine, Y. Tawk, A. H. Ramadan, and M. M. Tentzeris, "A Compact Source–Load Agnostic Flexible Rectenna Topology for IoT Devices," IEEE Trans. Antennas Propag., vol. 68, no. 4, pp. 2621–2629, 2019.

[6]   A. Eid et al., "A Flexible Compact Rectenna for 2.40Hz ISM Energy Harvesting Applications," in 2018 IEEE International Symposium on Antennas and Propagation USNC/URSI National Radio Science Meeting, Jul. 2018, pp. 1887–1888, doi: 10.1109/APUSNCURSINRSM.2018.8608525.

[7]   J. Costantine, A. Eid, M. Abdallah, Y. Tawk, and A. H. Ramadan, "A load independent tapered RF harvester," IEEE Microw. Wirel. Compon. Lett., vol. 27, no. 10, pp. 933–935, 2017.

[8]   A. Eid, J. Costantine, Y. Tawk, M. Abdallah, A. H. Ramadan, and C. G. Christodoulou, "Multi-port RF energy harvester with a tapered matching network," in 2017 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting, 2017, pp. 1611–1612.

[9]   M. M. Mansour and H. Kanaya, "Compact RF rectifier circuit for ambient energy harvesting," in 2017 IEEE International Symposium on Radio-Frequency Integration Technology (RFIT), 2017, pp. 220–222.

[10] S. Ahmed, Z. Zakaria, M. N. Husain, I. M. Ibrahim, and A. Alhegazi, "Efficient feeding geometries for rectenna design at 2.45 GHz," Electron. Lett., vol. 53, no. 24, pp. 1585–1587, Oct. 2017, doi: 10.1049/el.2017.2657.

[11] S. Ahmed, Z. Zakaria, M. N. Husain, and A. Alhegazi, "Integrated rectifying circuit and antenna design with harmonic rejection for RF energy harvesting," in 2017 11th European Conference on Antennas and

Propagation (EUCAP), Mar. 2017, pp. 1940–1944, doi: 10.23919/EuCAP.2017.7928106.

[12] S. Ahmed, Z. Zakaria, M. N. Husain, I. M. Ibrahim, N. A. Shairi, and A. Alhegazi, "High gain antenna design and doubler rectifier for microwave power transfer," in 2017 7th IEEE International Symposium on Microwave, Antenna, Propagation, and EMC Technologies (MAPE), Oct. 2017, pp. 296–299, doi: 10.1109/MAPE.2017.8250860.

[13] N. Hassan et al., "Design of dual-band microstrip patch antenna with right-angle triangular aperture slot for energy transfer application," Int. J. RF Microw. Comput.-Aided Eng., vol. 29, no. 1, p. e21666, 2019, doi: 10.1002/mmce.21666.

[14] F. S. M. Noor, Z. Zakaria, H. Lago, and M. A. M. Said, "Dual-band aperture-coupled rectenna for radio frequency energy harvesting," Int. J. RF Microw. Comput.-Aided Eng., vol. 29, no. 1, p. e21651, 2019, doi: 10.1002/mmce.21651.

[15] S. Bellal, H. Takhedmit, and L. Cirio, "Design and experiments of transparent rectennas for wireless power harvesting," in 2016 IEEE Wireless Power Transfer Conference (WPTC), 2016, pp. 1–4.

[16] H. Dong, X. Hou, Q. Zhang, and F. Wang, "Flexible slot-ring antenna for RF wireless energy harvesting," in 2018 International Workshop on Antenna Technology (iWAT), 2018, pp. 1–4.

[17] S. M. K. Azam, M. I. Ibrahimy, S. M. A. Motakabber, and A. K. M. Z. Hossain, "A compact bandpass filter using microstrip hairpin resonator for WLAN applications," in 2018 7th International Conference on Computer and Communication Engineering (ICCCE), 2018, pp. 313–316.

[18] S. M. K. Azam, M. I. Ibrahimy, S. M. A. Motakabber, A. K. M. Z. Hossain, and M. S. Islam, "A miniaturized hairpin resonator for the high selectivity of WLAN bandwidth," Bull. Electr. Eng. Inform., vol. 8, no. 3, pp. 916–922, 2019.

[19] S. M. K. Azam, M. I. Ibrahimy, S. M. A. Motakabber, and A. K. M. Z. Hossain, "Microstrip Coupled Line Bandpass Filter with Radial Stubs for Narrow-Band Applications," Intl J GEOMATE, vol. 13, no. 40, pp. 183–188, 2017.

[20] S. M. K. Azam, M. I. Ibrahimy, S. M. A. Motakabber, and A. K. M. Z. Hossain, "Plans for Planar: Phase-Noise Reduction Techniques in Voltage-Controlled Oscillators," IEEE Microw. Mag., vol. 20, no. 11, pp. 92–108, Nov. 2019, doi: 10.1109/MMM.2019.2935364.

Blockchain-based Global Travel Review Framework

Tanakorn Karode[1], Warodom Werapun[2]
College of Computing
Prince of Songkla University
Phuket, Thailand

Tanwa Arpornthip[3]
Faculty of Technology and Environment
Prince of Songkla University
Phuket, Thailand

Abstract—**An online review system is an important part of almost every e-commerce platform, especially a tourism e-commerce. However, various problems exist in the current online review systems. The review content is stored in a centralized database of each individual platform. Each platform differs in review management methods. In some cases, the review score of the same product disagrees across different platforms. Moreover, a centralized system has low transparency because it is difficult to trace individual actions within the system. As a result, some users are skeptical of the reliability of online reviews in centralized systems. This work proposes a global travel review framework based-on the blockchain technology. The incorporation of blockchain helps improve an online review system. The best practices for online review management from popular platforms, and the guidelines from trusted sources are used to develop the new system. The use of blockchain improves an online review system through its unique features of high transparency, security, and reliability. Additionally, the proposed framework relies on a community-driven environment. The accessibility level of users is controlled by using the smart contract. There is no single authoritative owner of the system. All participants in the system can exert controls on the system equally. This work illustrates the details of a blockchain-based global travel review framework. The advantages and disadvantages of such a system are discussed. The proposed framework can be easily integrated with any existing platforms since it can be accessed publicly.**

Keywords—Consumer online review; traveling; blockchain; smart contract

I. INTRODUCTION

From the Mintel report, 60% of travelers use online reviews to make a holiday trip plan [1]. People consider online reviews to reduce the risks of wasting money and time. It means that an online review much influences the travel-related products [2]. An online review impacts to the customer's trust and also the business's revenue. When people realized the effects of online review, they tried to manipulate it. As a result, fraudulent activities in online review systems are widespread over the internet [3]. It causes information that people perceive is misleading. The review readers need to take a longer time for online product assessment. And also, the business's credibility is undermined. Many platforms are struggling with those problems. Some methods they are using work very well. However, some strict procedures are not suited to the user's needs. For example, the automatic filtration scopes down the reviewer's opinion. Low-score reviews are frequently filtered out. The similar cases happen to the very high-score reviews. A blog author argues that his genuine review was filtered out from a platform, and the reason that he got was not making

sense [4]. Moreover, other customers do not know when a review is filtered out, updated, or removed since those actions are not exposed publicly. The online review Today's online review is controlled by the platform providers who can manage their platforms as they prefer. Each platform has its database and might have different practices in its system. According to the news, the score of a product is contrastive in two platforms [5]. This kind of circumstance critically affects the customer's trustworthiness toward the business. The score might partially reflect the quality of the product. However, the platform practice is one factor that impacts the overall score of a product. It occurs when a platform tries to maintain the attractive view of a partner business. The separated sources of reviews undermine user credibility. The businesses are also inconvenient to maintain the same score in every platform they use.

In this work, the blockchain-based global travel review framework is proposed. The main objective of this framework is to improve the credibility of the online review. The transparency of the system is much raised by utilizing the blockchain technology. Exposing actions in the system increases the user's credibility since it can be traced publicly. Blockchain helps in communication among the different systems that have its database much easier. All individuals have the same level of accessibility. The accessibility rule is constructed once by using a smart contract. The rule is displayed publicly and cannot be updated. Users can decide to participate in the system if they agree on the rules. Traveling reviews are stored in a single global database, which can be accessed publicly. There are many potential applications that can extend from this characteristic. For example, a business ranking application, a travel suggestion system, a fake review detection system, and many others.

The main contributions are not only the idea of the framework but also the implementation guidelines for the smart contract and decentralized applications (DAPPs). Additionally, the framework's performance in terms of cost and speed are analyzed to consider the potential of the real environment usage of the framework.

Although the proposed framework is implemented with blockchain technology, the centralized online review system's guidelines or practices are still needed. That information will be described in Section II. Then, we propose a framework in Section III. Section IV is the discussion about the advantages and some issues in the framework. Eventually, we conclude the work in Section V.

II. Literature Review

A. Online Review Components and its Impact

A consumer review contains the essential components that every platform presents, including review content, author's profile information, and review scores. That information describes the consumer's opinion of a product or service. There are the additional components that some platforms include them while some platforms do not use them, such as graphical information (i.e., pictures and videos), review timestamps, review helpfulness score, response to a review, and reviewer contribution history. The additional components provide further information, which makes the reader understand better and more accurate. Furthermore, the reader gets more information to decide whether to trust the review they are reading.

Prior researches investigated the effects of the review component as followings:

1) Review content: The descriptive text for the feedback explanation is the most important part of the review. There are many attributes inside the review content that the researchers use to analyze their characteristics. For example, readability represents how easily the reader can perceive the information from the review. The review content with low readability (i.e., difficult to read, using vague and improper words) can be treated as a low-quality review. Chen et al. proposed the manipulation review detection model by using a decision tree [6], where the critical attribute for review manipulation classifying was sentimental (both positive and negative sentimental). The informative direction is an important attribute for the quality review. As shown in [7], the review content that describes in one direction (either positive or negative) is likely to get more helpful score than the review that does not clearly present the aspect.

2) Rating: The numeric data that the reviewers provide to the products or services. This numeric data can be aggregated to create the summary data, which can minimize a huge load of information the readers comprehend. Also, the rating can be used to be the comparator among the businesses that participate in a platform. As reported in [8], a business with the higher review score and a greater number of reviews likely is to obtain more revenue.

3) Reviewer: The reviewer information is an essential source of data that the reader can use to consider the review credibility. The authors in [9] stated that the reviewers who disclose their identity more get more helpfulness scores. Trip Advisor is a platform that pays much attention to the reviewer's information by presenting the reviewer contribution system. Every user has the contribution score and history, which is very useful information for deciding the trustworthiness of a review posted by that user.

4) Graphical data: Photo or video is the attractive information that provides more details in the review content and raises the review credibility. The graphical data shows the real place, product, or situation is likely to get more trust. Especially, the one that appears the author identity. Thus, it can support the purchasing decision more than the only textual review [10].

5) Review helpfulness score: It is presented as the two-ways review. The scores that people give to a review reflect the level of trust among the readers. It supports reader consideration in trusting a review and its author. Many researchers use this component to analyze the perceived usefulness of the review readers [11]. Moreover, the review quality assessment is constructed by using helpfulness score. A high helpfulness score review is often used as a baseline for the quality review.

6) Review response: It expresses the opinion of other consumers or the business towards that review. It is a fair practice for the platform to provide a public replying channel. As a result, a reviewed business can dispute some misleading reviews.

7) Review timestamp: The time of writing review which can be used to identify some strange behaviors in the online review system. For example, the reviews posted to a business by the same user in a short period can be considered a spam.

B. The Guidelines for Online Consumer Review Platforms

Even though the consumer online review platforms contain similar components, the presenting information practice is quite different. Both consumers and businesses suffer from fraudulent activities in the online review. They need some guidelines for the best practice to address the credibility problem in the online review system. The Australian Competition and Consumer Commission (ACCC) published the online consumer review guidelines for consumers, businesses, and system providers [12]. We pick the guideline for the system provider to discuss the on-going platforms today. To create a consumer review system that plays the fairest to both consumer and business, ACCC provides the guideline for platform providers as followings:

1) Disclosure the commercial relationships: The platform and the business could have a commercial relationship that does not affect the review results, such as advertising or promoting searching results. The prominent disclosure about the commercial relationship should be presented to convey key information to any user. A guideline example is a prominent disclosure should be placed near the aggregated review results [13].

2) Do not publish misleading reviews: The misleading reviews cause perceived information about the reviewed business to mismatch the fact. The misleading reviews can be written by the reviewed business itself or the consumer who gets a benefit for posting an inflated review. On the other hand, the negative misleading reviews can also be generated by a competitor business or a third party that has a relationship with it.

3) The platform should not edit or delete the posted reviews: The information of review may be misleading if it does not actually come from the review authors. Removing or updating the review by the platform, particularly negative reviews cause decreasing the platform credibility.

4) Detecting and removing fake reviews: The platform should detect and manage the fake review properly to keep the platform's integrity and make it useful for the consumers. The ACCC recommends that online consumer review platforms remove the reviews that they know to be fake. There is no precise way to detect fake reviews. Here is the guideline from the ACCC. The fakes review may include those:

a) which are part of a significant frequent in reviews about a particular business in a period of time.

b) written from the same IP address or email of the reviewed business or the friendly business.

c) written about a particular business where each of those has abnormal similarities (e.g., username, IP address, email).

d) which use overly positive marketing-style language.

e) which do not make sense.

f) which use the same language style as the other reviews of the same business or product.

A practical way to detect fake reviews is to include a flag button, which provides users with an opportunity to point out fake reviews to the attention of platforms.

5) Incentivized user review: The ACCC recognizes that the potential of incentivized reviews may lead to biased reviews, inflating review results, and misleading consumers in some circumstances. When an online review platform offers an incentive, it should do following these recommendations:

a) The platform should disclose any incentive prominently on the review page of the businesses whose reviews are affected by that incentive.

b) The platform should be alert to the customer incentivizes provided by the businesses in order to manage it properly. If there is some review clearly and dramatically inflated by an incentive, the platform should remove that review to prevent misleading.

c) The additional best practice is to ask the reviewer whether they have received any incentives from the business when writing the review.

The incentive offers should not affect the review result. Namely, the incentives are offered regardless of the review score given by the reviewer.

6) The overall score of review is important: The overall review score (e.g. star ratings) of businesses should be presented in the platform. In many cases, users do not read many reviews before relying on the overall review score. It is the responsibility of the review platform to keep the review results be genuine. Deleting or hiding reviews suspected of being fake is not misleading as the consumers prefer improving the quality of reviews. When the platform provides

the aggregated rating. It is recommended to show the number of reviews next to the average scores (e.g., 5 stars, 20 reviews). Content moderation policy ensures users and businesses have a clear understanding of the reason for removing reviews. It is recommended that the providers inform the policy to the businesses and consumers.

7) Dealing with the businesses who have got unfavorable reviews: Businesses may receive negative reviews, which is understandable to be concerned. It is recommended that the platform should provide the public reply channel for the business to substantiate publicly that the review is fake. The review platform should follow the investigation. When there is enough evidence that suggests that the review does not come from the truth, the platform should remove that review as soon as possible.

8) A review should be written by a person who has already consumed a product: Businesses should not post a review to competitor businesses and their own business too. If there is some friendly business review, the platform should disclose the relationship between them. Importantly, businesses must avoid paying for posting reviews. This practice could induce misleading reviews.

C. The Current Online Review Systems

Nowadays, there are many tourism related platforms that provide the channel for the consumers to give some feedback and suggestion to the businesses publicly. In this work, we investigate the top six tourism related platforms on their consumer review systems by comparing with the ACCC guidelines. The studied platforms are Trip Advisor, Yelp, Booking.com, Expedia, AirBnB, and Agoda. The comparisons are presented in Table I.

When we compare the service practice in the popular platforms, we can clearly notice that the third-party platforms like Trip Advisor and Yelp (i.e., the platforms that do not provide the booking or payment process) are likely to contain more information than the booking or purchasing platforms. Also, the conquering fake reviews procedures are stricter in the third-party platforms. For example, Trip Advisor and Yelp contain the automatic filtration system to remove the reviews pointed out to be fake reviews. Additionally, the flag system that allows users to point out fake reviews is available on those platforms. More importantly, readers can consider trusting a review by considering its author's contributing history. In contrast, the booking platforms do not make much earnest about the online review system. While posting images and videos is the common practice for the user-generated content application, these platforms do not provide this feature. However, a huge advantage of the booking platforms is the review posting privilege. Only users who have consumed a service can post a review. This rule causes constructing fake reviews more difficult and costly.

TABLE I.   ONLINE REVIEW PLATFORM COMPARISON

Guidelines / Components	Platforms					
	Trip Advisor	*Yelp*	*Booking.com*	*Expedia*	*AirBnB*	*Agoda*
Review verification	The posted review is verified by automatic and manual system	The reviews are posted immediately	The reviews are verified before listed in the platform	The reviews are posted immediately	The reviews are posted immediately	The reviews are posted immediately
Removing or updating reviews by plaform	In many cases, the user reviews are filtered out	Many reviews are filtered out by the automatic system	Reviews could be removed if they do not follow the rules	The reviews are removed by platform if they contain the violent content	Reviews could be removed if they do not follow the rules	Reviews could be removed if they do not follow the rules
Fake or violent reviews report channel	Flag system	Flag system	None	None	None	None
Review response	None	The business can reply a review, but it does not show publicly.	None	None	None	None
Review posting allowance	Any user can post review	Any user can post review	Only user who has already purchased	Only user who has already purchased	Only user who has already purchased	Only user who has already purchased
Images or videos	Available	Available	Available	None	None	None
Reviewer contribution history	Available	Available	None	None	None	None

In summary, today's service practice in popular platforms is acceptable in some aspect, whereas some practices are questionable. If an online review platform is created with the best practice, it can eliminate a large amount of fake and violent reviews. However, detecting an actual fake review is not a simple task. In the current platforms, many genuine reviews are filtered out by automatic systems. Such a situation hurts the reviewers who spend some time to provide true information, but their reviews still get filtered out. As a result, the reviews are scoped in the narrow space. In addition, some platforms do not provide acceptable rules for review filtration. Thus, the very strict automatic filtration might not be suited for online reviews in some situations. A key for the quality review platform obtained from a research published by Wu, Ruhai, et al. is transparency [14]. It can be achieved by exposing the actions taking place in the platform as much as possible. The rest processes are just to let users find the evidence and consider trusting the information by themselves.

D. Online Review with Blockchain

As mentioned in the previous section, transparency is key to achieve a high credibility online review system. Sometimes, the reviewers might prefer to change or delete their reviews after posted. Allowing users to manage their reviews is a good practice because there might be mistakes in writing a review. However, the readers should be able to see all previous versions of the manipulated reviews. This practice offers more information to consider reviews. Some current platforms allow users to manipulate their reviews, but they do not provide all previous versions of reviews. Even though some platforms reveal all versions of the reviews, the platform owner can still manipulate them without the awareness of users since the data is stored in the centralized database. It is a trust issue that users need to rely on the terms and conditions provided by the platform owner. A centralized database cannot serve the full

transparency, a feature that is supported natively in a blockchain-based database.

The first emergence of blockchain technology was the time where Satoshi Nakamoto proposed an electronic cash system that can be done without intermediary [15]. Instead of storing all ledger data in the central authority, the blockchain ecosystem distributes all transaction history to every participated node. By doing so, individuals can validate their data with each other. Thus, tempering data is more difficult for more nodes in the system and cannot be done eventually. In the technical term, there are several steps to make it work. When a transaction is sent, it needs to be verified by the special nodes called "miner" before getting recorded. The miners check whether a transaction is valid by considering its correctness, such as whether the sender balance enough for the transferred amount plus the transaction fee, the receiver address, and the other aspects. Then miners gather the validated transactions into a set of data called "block". While miners simultaneously verify the transactions and put them to block, only a block from a miner gets recorded for each round. With that procedure, the blockchain data is sorted in the same way among all participated nodes. The blockchain network uses a consensus algorithm to select a miner. Once the consensus is succeeded, a miner who completes the consensus broadcasts its block to all other nodes and gets a reward from the blockchain network (coin base) plus a transaction fee. The name "blockchain" comes from the data structure using in its ecosystem. Every block of data is chained with its previous block using the cryptographic hash. With mathematical power, the hash function gives the same value for the same input every time. If a small piece of data in blockchain is changed, an individual can simply check by comparing its hash value.

The first version of blockchain uses Proof-of-work (POW) as a consensus algorithm to select a miner who has the right to use his/her block append to the latest block. Miners need to

find the hash value that has a lower value than the target value. The problem can be illustrated as the following equation:

$$SHA256(v, h_{prev}, h_{merkle}, t, b, n) = hash_{cur} \tag{1}$$

POW uses the SHA256 algorithm to calculate the block hash ($hash_{cur}$), as shown in (1). It gets the blockchain version (v), the hash value of the previous block (h_{prev}), the hash value of the Merkle root tree (h_{merkle}), timestamp (t), the target bytes (b), and a nonce (n) as the function parameters. The only way to solve the problem is to attempt changing n value until the $hash_{cur}$ value is lower than b value. A miner who seizes more computational power has faster attempting and more chance to win. Once a miner solves the problem, it broadcasts the n value throughout the network to proof. Other miners can simply use that value to check the answer and record the block data. This procedure provides very strong security when there are a large number of miners in the network. This is how Bitcoin was presented its security.

However, POW consumes much energy and spend a long time to complete a transaction. Moreover, most miners gather and create mining pools to get a higher possibility of winning. Thus, the computational power is not distributed and can be attacked with the 51% attack much easier. As a result, other consensus algorithms are proposed by many groups of developers. An interesting one is Proof-of-Stake (POS), that does not require a lot of computation power but using the amount of token miners staking instead. At the time of writing, the Ethereum blockchain uses POW as a consensus algorithm, and will be upgraded to POS in Ethereum 2.0 soon. Thus, the cost and transaction duration illustrated in this paper might be much cheaper and faster in Ethereum 2.0.

A smart contract is a small computer program that is running on the blockchain environment. It can be treated as the agreement that is created once and used in the long term without updating. The only way to update the agreement is to create a new one and use it instead of the deprecated ones. It plays fair to every participant since the agreement is exposed publicly. Anyone who wants to join the contract can estimate risk before participating in the contract. The Ethereum team firstly proposed the utilizing of the smart contract with blockchain. It improves the capability of blockchain to be able to work in general applications. According to its capability to connect the separated parties, it is utilized in the provenance application for product tracking [16]. It is also used in reward and loyalty programs since it can securely protect the data and cannot easily be hacked [17]. Moreover, the applications that need high transparency and the integrity of data history, such as voting, and ballot are the first usage of the smart contract [18].

Ethereum virtual machine (EVM) is a place that miners execute the transactions. A transaction sender needs to pay a cost if it contains state updating operations. The transaction cost is needed to keep the miners in the network and for the terminating mechanism. Since the smart contract can be developed similar to the general program, it might contain any mistakes such as infinity loop. According to Ethereum Yellow Paper [19], the transaction cost is positively correlated to the amount of data and the operation in that transaction. Such that, reducing those things in a transaction is a practical way to reduce the cost. In our previous work [20], we proposed ways to reduce the amount of data and the operations by using the Interplanetary File System (IPFS) and the smart contract event. IPFS is the distributed file system that can keep the permanently non-tempering data. It can store the huge amount of data and return the IPFS hash, which is the 48-character string refers to the actual data in the IPFS network. The proposed approach keeps IPFS hash in the blockchain history by emitting the event instead of using the smart contract state. The result showed that it could reduce approximately 20 times of cost compared to the smart contract that does not use IPFS and keeps the review content inside the smart contract state. That is a guideline for storing content data in the blockchain system. This approach will be used in this work for the smart contract implementation.

Currently, there are some groups of developers who are working on the blockchain-based online review system. We investigate some of the top of them listed below:

1) Revain [21]: The crypto community online review platform utilizes blockchain technology. This project is currently on the production state. It contains many businesses and users who generate a lot of reviews everyday. This project introduces the advantages of using blockchain with the online review system such as transparency, reward token, and the system cooperation. The businesses can easily integrate this review system in their platform by using the completed user interface provided by Revain or directly accessing review data in the blockchain storage. Moreover, this platform provides the user contribution history feature in the name of "karma". Every action of the reviewers reflects the karma point, which can increase their credibility. This platform is a pioneer project for the blockchain-based online review system that introduces various practical ways to improve online review credibility. There is something different from our work. Revain platform targets the crypto businesses review, but this work aims to the travel businesses. There is no purchasing verification before posting a review in the Revain platform, where anyone can still generate review without the real consuming experience. Even though the platform contains automatic filtration, it mainly focuses on the sentimental analysis where the extremely negative or positive reviews are filtered out. As a result, the overall reviews in the platform are scoped to be moderate tone.

2) Lina.Review [22]: Lina is a platform that utilizes blockchain technology in various fields, including review, supply chain, individual identity, healthcare, and education. Lina.Review is a module in the Lina platform. It stores the submitted reviews from users in a secure manner and rewards quality reviews. The individuals or businesses can build their systems on Lina.review platform. A special feature of Lina.Review is reviews from experts. A reviewer who proves the domain knowledge by providing CV to the platform becomes an expert. An alternative is getting promoted from other users by posting many accepted reviews. The experts and helpers write reviews to the businesses regarding their knowledge domain. They are entitled to receive the reward

from advertisement revenue and registration fee. Lina.review stores the review content in their private blockchain to reduce the transaction cost. As a result, the platform is under controlled by the Lina team. This point is the main difference between Lina and our work.

3) Dentacoin [23]: Dentacoin (DCN) claimed to be the first blockchain-based online review system for dental services. The dentists can register their dental offices to get patient feedbacks and display in public. Additionally, the patient can also register the dental clinics they have serviced. It uses the concept of "trusted reviews", where only the patients who have received service from the corresponding dentists can post the review. These patients are verified by the dentists sent a link for posting reviews via email. The reviews posted via those links are marked as trusted reviews. The reviewer gets the reward in DCN token after posting a review. DCN token can be used to get discounts or promotions for further treatment. Moreover, DCN token will be distributed through the industries which means that the users who are holding DCN token can use it as a cryptocurrency. The trusted reviews procedure presented in the DCN platform is an interesting way to reduce fake reviews. However, this approach still lacks of transparency. The dentist or even platform can send email to any reviewers, who might not have taken a service. While those actions cannot be traced publicly. Unlike in our work, the validating step relies on the transactions that take place in the blockchain. Thus, every action can be perceived in public.

III. PROPOSED FRAMEWORK

In this work, we create a global travel review system by using the Ethereum smart contract. The main goal of the system is to improve the review credibility. The travel review contents are stored in the distributed database among the participants. Every user sees the same version of reviews regardless of the platform or business they are using. The system does not need the central authority. It manages the accessibility by using the smart contract. The review authors have the most privilege in their reviews. Updating and deleting the posted reviews are possible but with a traceable history. This system is a community-driven environment. The incentive and fake review omission features are done by the user community. A review author is incentivized by the helpful score they get. This approach, customers are encouraged to write a quality review to get much helpful score from other users. Fake reviews can be pointed out by a user who has some clues. A user can get help from the community to judge a fake review by opening an issue. Other users can vote to agree or disagree with the argument provided by the issuer. Once the voting time is up, the winners get the reward, and the review is stamped to be fake if the "Yes" score is higher than the "No" score. All of the actions are recorded in blockchain. Thus, people can know all movements in the system and can consider trusting information presented in the platform. The framework details are described in the following sections:

A. System Architecture

The core component of the system is the smart contract. It enforces a rule that every participant needs to follow up. All participants can see the rules in the smart contract code and decide whether to join the system. The deployed smart contract cannot be updated. As a result, the rules do not change after the system starts operation. Every user in the Ethereum blockchain is a customer at the beginning state. When a user records his/her product in the smart contract, that user becomes a seller. While both customer and seller can retrieve data in the same way, there are the differences in actions they can take to the smart contract as followings:

1) Seller: Sellers can record their travel-related products (e.g., hotel room, food, tour guide, accommodation, and other travel products.) in the smart contract. Once a product is recorded in the smart contract, customers can use Ether to purchase it through the smart contract. The sellers are not allowed to purchase a product and do other actions requiring purchasing status, including posting a review, giving a helpful score, opening issue, and voting to remove a fake review. However, sellers can reply to a review that might be misleading review to clarify some misleading statements. The replying message is exposed publicly.

2) Customer: The customer can purchase products that are listed in the smart contract. After that, some actions are unlocked by using the purchasing state of the particular product. For example, a customer needs to refer to a purchased order to post a review. Every action the customer interacts with the system is recorded permanently in blockchain. All actions can be treated as user contribution history. It improves user's credibility and provides many opportunities for being an influencer in the traveling community.

The possible structure of the system can be presented as Fig. 1. Every participant can directly interact with the smart contract. A technical user who has some experience about the smart contract interaction might implement a channel to retrieve data or take some actions directly. However, creating a product order needs to be done by the product seller. Thus, businesses need to provide a channel for the user to request a product order for purchasing. It could be a business own website or a third-party platform. Presenting reviews with the correct content is the responsibility of the platform provider. If a platform does not play fair by changing some content in the presentation layer, the customers can be aware of that action by comparing it with other platforms or the smart contract. With the proposed structure, there can be many possibilities to support the travel industry. For example, a third-party can directly retrieve all customer feedbacks from the smart contract to create a travel suggestion website. The mentioned application is very useful if the review content is factual information. Thus, the community-driven approach presented in this work is a way to make it possible.
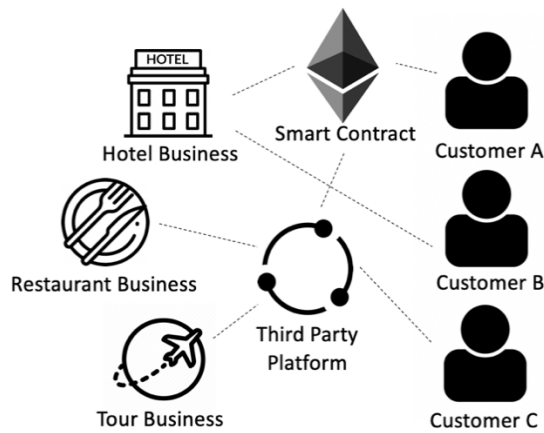
Fig. 1. System Architecture.

B. System Functions

The smart contract contains several functions that let users interact with the system. The overall system workflow is described below:

1) Recording a product: Sellers need to record product information and its review value in the smart contract. The product information needs to be stored in IPFS and get its hash to stamp in the smart contract. The schema of product information might be varied, but every user needs to follow up on the same format. The simplest schema contains the product name and a URL link for more information. Since updating data in the smart contract requires cost, the frequently changed data should be avoided storing in the smart contract. The review value is the amount of Ether that the business willing to pay for the user community to incentivize quality reviews and handle fake reviews. The review value can be changed afterward, but not too frequently, because it might not motivate users to post reviews.

2) Recording user profile: Like the product information, the user profile needs to be stored in IPFS, and its schema should be the same for all users. Otherwise, it might cause problems in the presentation layer. The users have alternatives to put their information into the system or not. According to prior research, the helpful score of a review is directly affected by the user identity disclosure. As a result, users are allowed to record their information to improve their credibility. The recorded information is permanently bound with the user address. Users can update their profile data, but all of its versions are available in blockchain.

3) Purchasing a product: After a product is recorded, it appears to every user. The product purchasing steps are illustrated in Fig. 2. A customer can buy a product by requesting the seller to create a product order. A seller creates an order by determining product price in Ether and the buyer's address. Since the product price can be varied by factors such as the number of products, promotion, etc., the system is designed to handle the dynamic pricing. When an order is already created, it got an identification number (order id). The order id is returned to the customer, and he/she can use it to

purchase a product with the defined price in the order. The net amount of Ether that the business gets is the total amount deducted by the review value (i.e., net amount = price – review value). The review value is kept in the smart contract. A customer can refer to a purchased order to get the right to do these actions including posting a review, give a helpful score to another's review, open issues, and vote for issues.
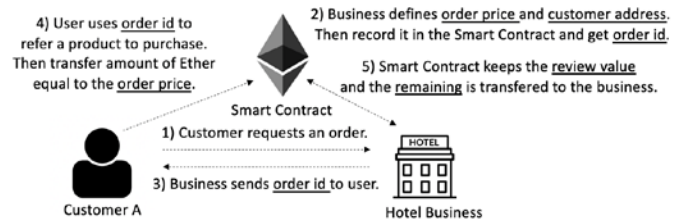
Fig. 2. Purchasing Procedure.

4) Managing a review: A review posted to the system needs to be stored in the IPFS to reduce the transaction cost. When a review is recorded in the blockchain, it can be updated or deleted by its author. In practice, the smart contract events for deleting and updating a review are emitted to update the review version, while the old versions of the review are still accessible. In the presentation layer, a platform that utilizes this system should only present the latest version of the reviews. And also, other versions should be available for users because they can be proofs of fake reviews.

5) Review Replying: The system allows businesses or other consumers to reply to a review if there is some misleading information. The replying content needs to be stored in IPFS too. The smart contract emits an event to record the replying message. It is the responsibility of a platform to display all the replied messages appending to the particular reviews.

6) Helpful score: After finish purchasing, a customer has two choices about the review value. Firstly, giving a helpful score to another's review within a time. This way, the review value is transferred from the smart contract to the author of a review that gets a helpful score. Lastly, omitting a helpful score giving. If a customer does not take any action to the helpful score until the end of the time, the customer cannot use that order to give a helpful score anymore. The review value of that order can be picked by anyone who opens an issue. The review value is used to be the additional reward for the issue winners.

7) Openning an issue: Once a customer notices that a review in the product he/she has ever purchased might be misleading, the customer can open an issue for that review. For opening an issue, the customer needs to define voting timeout, the maximum value to vote, and the argument text to point out the mistakes of that review. If there is an order that is not spent on giving a helpful score, it can be picked up as the additional reward for the issue winners. The customer can define the amount of Ether in voting that review at the time of opening an issue.

8) Voting an issue: Customers who purchased a product that contains an issue can participate in the issue by voting. A voter can decide to vote "yes" to agree or "no" to disagree with the issuer's argument. The level of confidence depends on the amount of Ether the voter spends. While everyone can see the realtime voting result, there might occur the last-minute voting problem. In order to prevent it, the voting timeout is extended in a period of time if there still exists voting near the ending time. When the extended time is up without additional votes, the voting result is summarized. If "yes" score wins, the review is marked as a fake review. Otherwise, it is still in the same place. The reward calculation can be illustrated as (2), (3), and (4). Firstly, the reward portion (P) is calculated by dividing the user's voting amount (V_{own}) with the total voting amount in the same side ($V_{total\ same}$). If there exist the additional reward ($R_{addition\ total}$) from an unused order, it is shared among the winners by ($R_{addition}$) value. Finally, the voting reward (R) is the partition of total votes in the opposite side ($V_{total\ opposite}$) plus the voter's vote amount (V_{own}) and the portion of additional reward ($R_{addition}$). The reward is distributed amog the winners. Investing more Ether increases the portion of the earned reward. However, losing an issue means wasting all invested Ethers too. With this condition, voters need to be confident in their votes.

$$P = \frac{Vown}{Vtotal\ same} \qquad (2)$$

$$R_{addition} = P \times R_{addition\ total} \qquad (3)$$

$$R = (P \times V_{total\ opposite}) + V_{own} + R_{addition} \qquad (4)$$

C. Performance

We tested the system on the Ethereum Ropsten test network because it uses POW as a consensus algorithm. Thus, the result is nearest the main network. Objectives of the experiment is to evaluate cost and response time for a transaction. The Gas price used in the experiment is 3 GWEI since it is the average speed of the transaction suggested by the Metamask wallet. We tried submitting 20 transactions for each method. The experiment results are presented as followings:

1) Response time: We focus on two types of response time; hash time and block time. Hash time is the duration starting from a transaction is submitted until it gets the transaction hash. Block time is when a transaction is submitted until the time that it is recorded in the blockchain. From the experiment, the average hash time ranges from 1 to 2 seconds. At the same time, the block time is in the range of 20 to 30 seconds.

2) Transaction cost: Each method consumes a different amount of cost as shown in Table II.

The result presented in Table II is the number of gas used for each transaction. The actual cost can be calculated by defining gas price value and multiply with the gas used value. For example, the gas price is 3 GWEI. The transaction cost of the "Add product" method is 78,379 x 3 GWEI, which is 235,137 GWEI (1 GWEI = 10^{-9} Ether). The transaction cost in fiat can be calculated by considering the Ether value comparing

to fiat. For instance, the Ether value at the time of writing is 244.55 USD. As a result, the transaction cost for the "Add product" method in fiat is 235,137 x 10^{-9} x 244.55 USD, which is 0.0575 USD.

TABLE II.   GAS USED FOR EACH METHOD

Method Name	Gas Used
Add product	78,379 (± 7,246)
Update product	26,024 (± 0)
Update profile	25,442 (± 0)
Create order	117,374 (± 16,807)
Purchase	81,233 (± 2,589)
Post review	50,216 (± 4,085)
Delete review	27,148 (± 0)
Update review	30,080 (± 0)
Reply review	27,725 (± 0)
Give helpful	50,203 (± 3,338)
Open issue	133,041 (± 16,432)
Vote	75,277 (± 7,131)
Get reward	71,743 (± 11,475)

IV. DISCUSSION

As mentioned in Section II, today's popular online review platforms have their strengths, but they lack some important components. The proposed work tries to gather all advantages to raise the highest credibility to the world of online review as much as possible. With the blockchain, there are many advantages compared to the existing solutions. However, some weaknesses still exist. Both advantages and disadvantages are described below:

A. Advantages

1) Global database: Today online reviews are separated in each platform. A business that participates in more than one platforms might have many versions of its reviews. It much improves their credibility, if all of those are the same trend. In contrast, if those versions have big differences, the business loses its credibility from the consumers. The platform might be one main factor of the difference score because of the platform providers' different practices. Keeping the review in one place handles this problem and helps consumers to focus on one point of trust. The business gets the benefit from this practice too. By using the blockchain as a global database, every platform can use the same version of the reviews. The businesses do not need to begin with zero reviews when they move to a new platform. The global database is an ability that cannot easily be achieved with the centralized system since it needs trust among the database providers.

2) The value of review: Before posting a review, the customer needs to purchase a product. Moreover, posting a review to the smart contract needs transaction costs. Thus, a review presented in the system is more valuable because of the effort needed to generate it. A review can also be treated as a

property of the review author and the business that gets reviewed. While the review is recorded permanently in the blockchain, it is bound to the author and the business addresses. This framework does not rely on any authority. As a result, the reviews are still valuable if at least one platform uses the system. In contrast, the reviews are unworthy if a platform that is maintained by an authority stops the service.

3) Communinty-driven ecosystem: As described in the previous sections, the community promotes quality reviews and tackles fake reviews. In reality, defining good or fake reviews is not a simple task. It cannot be handled by just using text analysis or some automatic filtration because it depends on the real experience that a customer met. The people who can give the best answer are those who get the same experience, not the automatic filtration system or any specialist. The extremely positive reviews can be fake or honest, and in the same case for the extremely negative reviews. The majority controls the direction of the system. Thus, the proposed work lets the full control of the ecosystem to the users. This feature cannot be accomplished without trust in the centralized system since a system needs the maintainer who has the highest privilege to control everything in the platform.

4) Real purchasing validation: Some current platforms do not allow customers to post a review without purchasing a product. The validation procedure uses the data in the centralized database to check the user's purchasing status. Obviously, the status can be changed by the platform owner. There might exist the case that a business pays to the platform to boost its review score. Additionally, the purchasing status might get fake by the platform provider itself. These practices cannot be done in the proposed framework. All purchasing states are the real transactions generated by the actual purchasing between the customer and the seller. Thus, generating fake negative reviews needs much effort to do. On the other hand, fake positive reviews might be generated by friendly businesses or third customers. However, these actions can still be detected by users. Since the transaction history in the blockchain is exposed publicly, anyone can find the relationship between those accounts. Consequently, it provides useful clues to judge the fraudulent activities in the system.

5) Two levels promotion: In the popular Trip Advisor platform where the incentive does not exist, plenty of reviews are still generated every day. One of the factors is that Trip Advisor is a popular platform with a large traveling community. The user's contribution history impacts the review credibility, and it is an opportunity for the reviewer to make their profile. In the proposed framework, the user contribution creates more impact since every action needs cost to do. Moreover, the incentive is available in the proposed system. This incentive directly promotes the reviewer to create a review that gives benefit to the readers. Unlike the general platforms that give reward to every reviewer who follows up some conditions such as the minimum word count, the graphical information. The result from that practice is low-

quality reviews from the users who only need a reward from posting a review. In the proposed framework, the customers are promoted in two levels to keep the best practice in their reviews and avoid the punishment. The first is the reward from the helpful score and the last is the contribution history which directly impacts the reviewer's credibility.

6) On-demand review: The businesses can express their requirement for reviews by adjusting the review value. Increasing the review value motivates the reviewer to post review more. The business can also minimize the review value if it does not need more reviews in a period of time.

7) High-level data protection: As mentioned in the basic knowledge about the blockchain, the possible attack that can manipulate the newly recorded data in the blockchain is 51% attack. It needs extremely high effort and cost. Hackers receive little gain compared to the effort they need use to attack the system. In addition, hacking the blockchain network is almost impossible because the difficulty increase along with the number of transactions in the network.

B. Other Issues

1) Changing of Ether value: In the proposed framework, the main currency that is used to drive every action in the system is Ether. Since the Ether is a volatile currency, its value change over time. As a result, the transaction cost can be very cheap and expensive sometimes. To deal with the high transaction cost, the sender might reduce the gas price value when the Ether value is very high. Another case is in the purchasing procedure, where a business defines the product price for each order. The volatility of Ether might affect the business that mainly uses fiat to define the product price. The created order needs to be purchased immediately before the Ether value gets much different. Otherwise, the business needs to create a new order where the additional transaction cost is needed. This issue can be solved by using a stable coin for purchasing procedures. However, the transaction cost might be higher than those which appear in this work. The best solution is to use the cryptocurrency as the main currency in the system. It might be possible for a blockchain network with enough speed and cheap transaction cost, where people prefer to use it as the main currency.

2) Uncommon user interface: Like every new technology in the early phase, general users are not familiar with its processes. In the proposed framework that utilizes the blockchain as the core system, users must keep their private key secretly. Besides, they might also need to know about the gas price and gas limit they define every transaction. Moreover, the submitted transaction is not complete until a bunch of time for Block confirmation. This experience needs some time for people to get familiar. Fortunately, many DAPPs are being developed and coming closer to general users every day. The day that people use DAPPs as the primary applications is not so far from now.

V. Conclusion and Future Work

The main contribution of this work is to discover the potential features of a blockchain-based online review system. We proposed a global-scale online review for tourism system. The same version of consumer reviews can be displayed seamlessly in any platform. The proposed framework supports consumers and businesses better than the centralized approach. Consumers can be confident that the review score is not affected by the platform providers. Beside the businesses can maintain the same rating score regardless the platform they take part. Low-quality review handling is a challenge for the global-scale review system. The automatic filtration is a popular solution that the existing platforms use to address the problem. However, such complex solution cannot easily operate on-chain. Thus, we enforce a set of rules for maintaining the review trustworthiness. A customer needs to purchase a product before posting a review to that product. The posted reviews can be marked as "fake" or "low-quality" when the majority users agree. Every action in the system is recorded permanently and exposed publicly. Violent actions undermine the reviewer credibility. In contrast, well actions increase the author faithfulness. The review authors get incentivized from those actions. As a result, the system environment motivates users to take actions properly. Eventually, this framework can be easily integrated into the existing platforms and acts as an alternative way for the users who need more credible information. This way, people can get familiar with the uncommon interface of the framework.

The proposed work still has some features that can be improved more, such as the ability to check purchasing status for fiat, consuming validation without payment. These abilities provide the wider potential of usage since people still use fiat as the main currency today. Moreover, the reviews that does not need to be purchased are the general use cases in the online review system, such as attraction reviews. It needs off-chain verification to validate those events. A possible way to do is to use an Oracle. These issues will be investigated in future work to improve the capability of this work.

Acknowledgment

References

[1] J. Worthington, "Executive Summary: Holiday Planning and Booking Process- UK-May 2018," 2018.

[2] U. Gretzel and K. H. Yoo, "Use and Impact of Online Travel Reviews," Inf. Commun. Technol. Tour. 2008, pp. 35–46, 2008, doi: 10.1007/978-3-211-77280-5_4.

[3] S. Gossling, C. M. Hall, and A.-C. Andersson, "The Manager's Dilemma: A conceptualization of online review manipulation strategies," Curr. Issues Tour., vol. 21, no. 5, pp. 484–503, 2018.

[4] CliveNicol, "Review censorship!!!??," 2015. [Online]. Available: https://www.tripadvisor.com/ShowTopic-g1-i12105-k8465871-o60-Review_censorship-Tripadvisor_Support.html. [Accessed: 13-Jul-2020].

[5] M. McGrath, "Why Consumer Reports Says You Can't Trust Angie's List," 2013. [Online]. Available: https://www.forbes.com/sites/maggiemcgrath/2013/09/18/why-consumer-reports-says-you-cant-trust-angies-list/#647bc5531bfa. [Accessed: 10-Jan-2020].

[6] L. S. Chen and J. Y. Lin, "A study on review manipulation classification using decision tree," in 2013 10th International Conference on Service Systems and Service Management - Proceedings of ICSSSM 2013, 2013, pp. 680–685, doi: 10.1109/ICSSSM.2013.6602538.

[7] I. E. Vermeulen and D. Seegers, "Tried and tested: The impact of online hotel reviews on consumer consideration," Tour. Manag., vol. 30, no. 1, pp. 123–127, 2009, doi: 10.1016/j.tourman.2008.04.008.

[8] S. Park and J. L. Nicolau, "Asymmetric effects of online consumer reviews," Ann. Tour. Res., vol. 50, pp. 67–83, 2015, doi: 10.1016/j.annals.2014.10.007.

[9] Z. Liu and S. Park, "What makes a useful online review? Implication for travel product websites," Tour. Manag., 2015, doi: 10.1016/j.tourman.2014.09.020.

[10] E. J. Lee and S. Y. Shin, "When do consumers buy online product reviews? Effects of review quality, product type, and reviewer's photo," Comput. Human Behav., vol. 31, pp. 356–366, 2014, doi: 10.1016/j.chb.2013.10.050.

[11] B. Fang, Q. Ye, D. Kucukusta, and R. Law, "Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics," Tour. Manag., vol. 52, pp. 498–506, 2016, doi: 10.1016/j.tourman.2015.07.018.

[12] C. Commision, "WHAT YOU NEED TO KNOW ABOUT: Online reviews—a guide for business and review platforms," Canberra, Australien Captial Territory, 2013.

[13] Y. Shen, W. Shan, and J. Luan, "Influence of aggregated ratings on purchase decisions: an event-related potential study," Eur. J. Mark., 2018, doi: 10.1108/EJM-12-2016-0871.

[14] R. Wu and C. Qiu, "Seller manipulation of consumer reviews under competition," Proc. Annu. Hawaii Int. Conf. Syst. Sci., vol. 2016-March, pp. 3575–3583, 2016, doi: 10.1109/HICSS.2016.447.

[15] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," Consulted. 2008, doi: 10.1007/s10838-008-9062-0.

[16] Q. Lu and X. Xu, "Adaptable Blockchain-Based Systems: A Case Study for Product Traceability," IEEE Softw., vol. 34, no. 6, pp. 21–27, 2017, doi: 10.1109/MS.2017.4121227.

[17] L. Wang, X. Luo, Y. Hua, and J. Wang, "Exploring How Blockchain Impacts Loyalty Program Participation Behaviors: An Exploratory Case Study," in Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019, doi: 10.24251/hicss.2019.553.

[18] F. P. Hjalmarsson, G. K. Hreioarsson, M. Hamdaqa, and G. Hjalmtysson, "Blockchain-Based E-Voting System," in IEEE International Conference on Cloud Computing, CLOUD, 2018, pp. 983–986, doi: 10.1109/CLOUD.2018.00151.

[19] G. Wood, "Ethereum: a secure decentralised generalised transaction ledger," Ethereum Proj. Yellow Pap., vol. 151, pp. 1–32, 2014, doi: 10.1017/CBO9781107415324.004.

[20] T. Karode and W. Werapun, "Performance Analysis of Trustworthy Online Review System using Blockchain," in The 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2020), 2020, pp. 510–513.

[21] Revain LLP., "Revain Full Whitepaper," 2018.

[22] L. Network, "(LINA) Blockchain Based Application for Innovation LINA Network," pp. 1–34, 2019.

[23] T. M. Switch, "Dentacoin: The Blockchain Solution for the Global Dental Industry," p. 37, 2017, doi: 10.1108/09604521111113465.

Assessing Vietnamese Text Readability using Multi-Level Linguistic Features

An-Vinh Luong[1], Dien Dinh[3]
Computational Linguistics Center
University of Science
Ho Chi Minh City, Vietnam

Diep Nguyen[2]
Department of Linguistics
University of Social Sciences &
Humanities
Ho Chi Minh City, Vietnam

Thuy Bui[4]
Faculty of Foreign Languages
Hoa Sen University
Ho Chi Minh City, Vietnam

Abstract—**Text readability is the problem of determining whether a text is suitable for a certain group of readers, and thus building a model to assess the readability of text yields great significance across the disciplines of science, publishing, and education. While text readability has attracted attention since the late nineteenth century for English and other popular languages, it remains relatively underexplored in Vietnamese. Previous studies on this topic in Vietnamese have only focused on the examination of shallow word-level features using surface statistics such as frequency and ratio. Hence, features at higher levels like sentence structure and meaning are still untapped. In this study, we propose the most comprehensive analysis of Vietnamese text readability to date, targeting features at all linguistic levels, ranging from the lexical and phrasal elements to syntactic and semantic factors. This work pioneers the investigation on the effects of multi-level linguistic features on text readability in the Vietnamese language.**

Keywords—*Text readability; text difficulty; readability formula; linguistics features; Vietnamese*

I. INTRODUCTION

Text readability is a measure of how easy or difficult a text is to be read [1], effectively guiding the process of comprehending that text. The readability of a document heavily depends on its linguistic features such as word usage, phrasal structures, and sentence meaning. Not only does text readability help readers determine whether a document is suitable to read, but it also assists authors in adjusting their writing for the target audience.

Building a model to assess the readability of texts yields great significance across various disciplines. In academia, researchers can rely on text readability to improve their scientific communications, while curriculum designers can be assured in developing appropriate course outlines for each age group of students, and language teachers can effectively create or select relevant second language learning materials for foreigners. Moreover, text readability plays a key role in aiding publishers in establishing varied audiences, supporting policy makers in drafting legal documents that accommodates all citizens with different literacy levels, and supporting manufacturers in preparing product manuals.

Research on the readability of text has been conducted since the late nineteenth century, with a special focus on English and other resource-rich languages. These studies are generally divided into two main approaches: the statistical approach and the machine learning approach. The statistics-oriented works mainly examine how the features of a text affect that text's readability using correlation and regression analyses. These analyses determine features that are highly correlated with readability and calculate the weight of those features, respectively, to develop formulas that predict the readability of that text. Representative works of this approach include the Dale-Chall formula [2], the SMOG formula [3], among others. Meanwhile, studies that follow the machine learning approach seek to exploit neural network algorithms with great computational power that enable the manipulation of a broader range of features and at a deeper level to create text classifiers based on the readability level. Works that demonstrate this approach are Si and Callan [4], Collins-Thompson and Callan [5], Pitler and Nenkova [6], Vajjala and Meurers [7], Sinha and Basu [8], Vajjala and Lučić [9], and Al Khalil, et al. [10], among others.

In Vietnamese, research on text readability remains relatively limited. First, Nguyen and Henkin [11] pioneered this vein of research for overseas Vietnamese people. Then, in 2017, when examining the features of text in linguistic textbooks, Luong, et al. [12] showed that the text length significantly influences the classification of these grammatical texts by readability level. In another study in 2018, Luong, et al. [13] further argued that Sino-Vietnamese elements and dialect features also plays a critical role in evaluating the readability of texts in Vietnamese textbooks.

Besides the relatively small number of studies on this topic in Vietnamese, the features examined are only at shallow levels, with surface statistics such as word frequency and type-token ratio. Features at higher levels like syntax and semantics remain still untapped, mainly due to the lack of survey resources and the low accuracy rates yielded by in-depth word processing tools. Recently, more extensive studies on Vietnamese texts have gained increasing attention and promising results, leading to their application to the problems of natural language processing in general and the question of text readability in particular. Therefore, in this study, we investigate the effects of linguistic features on the readability of text in Vietnamese. These linguistic features range from word-level (word frequency, language, sentence length, etc.) and Language model features (bi-gram, tri-gram, etc.) to syntactic (parsing tree height/width, number of clauses, etc.)

and fundamental semantic features (average of semantic numbers of words/sentences). Not only is this work the most comprehensive study on this topic in Vietnamese as of the time of publication, it is also the first to exploit the deepest linguistic level of Vietnamese texts for the readability question.

The rest of the paper will be structured as follows: Section 2 presents relevant previous works addressing the text readability problem. Section 3 introduces the features examined, the dataset used for the examinations, the methods, and the results of our study. Finally, Section 4 contains the bulk of discussions and conclusions drawn from the experimental process.

II. RELATED WORKS

In this section, we will introduce previous studies on the text readability problem in the world as well as in Vietnamese. As introduced in Section 1, the study of text readability has begun since the end of the nineteenth century. While a great deal of works has been published since then, the research focus has been on English and other resource-rich languages.

There are two main approaches in the study of text readability: (1) statistical approach and (2) machine learning approach. In the statistical approach, researchers focus mainly on identifying features closely related to the difficulty of a text through correlation analysis. Then, the selected features are used to construct the readability measurement formulas of the text. This approach has been implemented in a broad range of studies, including but not limited to Chall and Dale [2], Kincaid, et al. [14], Zeno, et al. [15], as well as Lee and Hasebe [16]. In Vietnamese, there have been four studies based on this approach: three of which are by Nguyen and Henkin [17], Nguyen and Henkin [11], Luong, et al. [18], and one of which is by Nguyễn, et al. [19].

TABLE I. SOME NOTABLE STUDIES IN RECENT YEARS ON TEXT READABILITY FOR RESOURCE-RICH LANGUAGES AND FOR VIETNAMESE

Work	Dataset	Features
Statistical approach		
Kincaid, et al. [14]	531 subjects from four schools at two Navy bases	Average length of sentences and average number of syllables per word
Chall and Dale [2]		Percentage of difficult words and average length of sentences
Lee and Hasebe [16]	A combination of texts from 83 introductory to advanced Japanese textbooks and texts from National Diet meeting transcripts, categorized into 6 scale levels	Average length of sentences, proportion of kango, proportion of wago, proportion of verbs, and proportion of auxiliary verbs
Machine learning approach		
Sun, et al. [23]	637 documents extracted from textbooks for grades one to six in mainland China	76 text features from surface features, Part-of-Speech features, parse tree features, and Entropy features
De Clercq, et al. [21]	105 paragraphs from the Dutch LassyKlein corpus	Fundamental level, language model features, and deeper level features
Chen and Daowadung [24]	720 texts from six subjects of elementary school textbooks in Thailand	Term frequency features, shallow features, and language model features
Berendes, et al. [22]	2,928 readings in the geographic textbooks of four publishers in Germany from grades 5 to 10	Vocabulary, syntax, morphology, and cohesion-related features
Tseng, et al. [25]	1,441 social science articles and 772 natural science articles	LSA features
Vietnamese		
Statistical approach		
Nguyen and Henkin [17]	20 text paragraphs with about 300 words each from Vietnamese novels and magazines, as well as textbooks of Vietnamese students in the United States from grade 4 to college	Average length of sentences and average length of words
Nguyen and Henkin [11]	24 text paragraphs with about 300 words each from Vietnamese novels and magazines, as well as textbooks of Vietnamese students in the United States from grade 4 to college	Word difficulty and average length of sentences
Luong, et al. [18]	996 texts collected from stories for children, sample essays, fairytales, textbooks, newspapers, political theory articles, language and literary articles, law, and legal documents,…	Average length of sentences, average length of words, and percentage of difficult words
Nguyễn, et al. [19]	209 prose texts in Vietnamese textbooks for elementary school children from grades 2 to 5	25 Part-of-Speech elements
Machine learning approach		
Luong, et al. [12]	288 texts from Vietnamese textbooks for elementary students and Literature textbooks for junior high school students in Vietnam	Average length of sentences, average length of words, and percentage of difficult words, and the length of text
Luong, et al. [13]	372 texts from Vietnamese textbooks for general students in Vietnam	Percentage of Sino-Vietnamese words, percentage of dialect words, and percentage of proper nouns

Meanwhile, in the machine learning approach, features are included in machine learning classifiers to evaluate which features help increase the accuracy of the classification process. Some pre-graded reference texts are utilized to train the model and evaluate the classification accuracy. Some of the notable studies on this approach are Dell'Orletta, et al. [20], De Clercq, et al. [21], and Berendes, et al. [22], etc. In Vietnamese, studies based on this approach have only been carried out in recent years like those of Luong, et al. [12], Luong, et al. [13].

Table I presents a summary of some influential studies on text readability from both approaches in recent years along with information about the dataset and features examined for a range of languages, including Vietnamese.

III. RESEARCH DESIGN AND METHODOLOGY

In this section, we will present our examinations on linguistic features of documents that can be extracted automatically by word processing tools for Vietnamese (up to the present time) to address the question of assessing the readability of Vietnamese writings.

A. Features

In this study, we examined 271 linguistic features listed in Table II. These features range from superficial features such as the average sentence length, the ratio of Sino-Vietnamese words, and the local word ratio, etc. (21 features in total) and word-type (Part-of-speech – POS) level features, such as the ratio of proper nouns, the average number of word-types, etc. (150 features in total) to syntax-level features such as the depth of syntactic trees, the numbers of clauses and of connected words per sentence, etc. (31 features in total) and basic semantic features such as the ratios of monosemous words and of polysyllabic words, the average number of meaningful units per sentence, etc. (10 features in total). Regarding features at the shallow level, we examined 30 language model features such as the average rank, the average frequency, and the average perplexity value of n-grams. These n-grams include character n-gram, syllable n-gram, word n-gram at bi- and tri-grams levels. Meanwhile, for features at the word-type level, we focus on the language model features at word bi-grams and word tri-grams (12 features in total). At the semantic level, given that research on automatic semantic labeling in Vietnamese text is still limited, we only extracted 17 basic statistical features such as the ratios of monosemous words and of polysemous words, the average meaningful units per word in the text, as well as the geometric mean of meaning of sentences in text, etc.

TABLE II. LIST OF FEATURES EXAMINED

RAW FEATURES		
distinct easy syllables/distinct syllables	ratio of monosyllabic words	
distinct easy word/distinct words	ratio of polyphonic words	average word length in character
ratio of 2-syllable words	average sentence length in character	average word length in syllable
ratio of 3-syllable words	average sentence length in syllable	ratio of long sentence (in syllable)
ratio of distinct easy syllables	average sentence length in word	ratio of long sentence (in word)
ratio of distinct easy words	average sentence lengths in syllable (remove duplicate)	ratio of short sentence (in syllable)
ratio of easy syllables		ratio of short sentence (in word)
ratio of easy words	average sentence lengths in word (remove duplicate)	
PART-OF-SPEECH FEATURES		
POS tags/sentences	distinct directional verbs/distinct words	emotion words/distinct words
POS tags/words	distinct directional verbs/sentences	emotion words/sentences
ratio of 2-POS tag words	distinct directional verbs/words	emotion words/words
ratio of 3-POS tag words	distinct emotion words/distinct words	foreign words/distinct words
ratio of multi POS tag words	distinct emotion words/sentences	foreign words/sentences
ratio of single POS tag words	distinct emotion words/words	foreign words/words
adverbs/distinct words	distinct foreign words/distinct words	idioms/distinct words
adverbs/sentences	distinct foreign words/sentences	idioms/sentences
adverbs/words	distinct foreign words/words	idioms/words
common nouns/distinct words	distinct idioms/distinct words	modifiers/distinct words
common nouns/sentences	distinct idioms/sentences	modifiers/sentences
common nouns/words	distinct idioms/words	modifiers/words
comparative verbs/distinct words	distinct modifiers/distinct words	numerals/distinct words
comparative verbs/sentences	distinct modifiers/sentences	numerals/sentences
comparative verbs/words	distinct modifiers/words	numerals/words
concrete nouns/distinct words	distinct numerals/distinct words	onomatopoeia/distinct words
concrete nouns/sentences	distinct numerals/sentences	onomatopoeia/sentences
concrete nouns/words	distinct numerals/words	onomatopoeia/words
countable nouns/distinct words	distinct onomatopoeia/distinct words	parallel conjunctions/distinct words
countable nouns/sentences	distinct onomatopoeia/sentences	parallel conjunctions/sentences

countable nouns/words	distinct onomatopoeia/words	parallel conjunctions/words
demonstrative pronouns/distinct words	distinct parallel conjunctions/distinct words	personal pronouns/distinct words
demonstrative pronouns/sentences	distinct parallel conjunctions/sentences	personal pronouns/sentences
demonstrative pronouns/words	distinct parallel conjunctions/words	personal pronouns/words
directional co-verb/distinct words	distinct personal pronouns/distinct words	prepositions/distinct words
directional co-verb/sentences	distinct personal pronouns/sentences	prepositions/sentences
directional co-verb/words	distinct personal pronouns/words	prepositions/words
directional verbs/distinct words	distinct prepositions/distinct words	proper nouns/distinct words
directional verbs/sentences	distinct prepositions/sentences	proper nouns/sentences
directional verbs/words	distinct prepositions/words	proper nouns/words
distinct adverbs/distinct words	distinct proper nouns/distinct words	quality adjectives/distinct words
distinct adverbs/sentences	distinct proper nouns/sentences	quality adjectives/sentences
distinct adverbs/words	distinct proper nouns/words	quality adjectives/words
distinct common nouns/distinct words	distinct quality adjectives/distinct words	quantity adjectives/distinct words
distinct common nouns/sentences	distinct quality adjectives/sentences	quantity adjectives/sentences
distinct common nouns/words	distinct quality adjectives/words	quantity adjectives/words
distinct comparative verbs/distinct words	distinct quantity adjectives/distinct words	state verbs/distinct words
distinct comparative verbs/sentences	distinct quantity adjectives/sentences	state verbs/sentences
distinct comparative verbs/words	distinct quantity adjectives/words	state verbs/words
distinct concrete nouns/distinct words	distinct state verbs/distinct words	subordinating conjunctions/distinct words
distinct concrete nouns/sentences	distinct state verbs/sentences	subordinating conjunctions/sentences
distinct concrete nouns/words	distinct state verbs/words	subordinating conjunctions/words
distinct countable nouns/distinct words	distinct subordinating conjunctions/distinct-words	temporal nouns/distinct words
distinct countable nouns/sentences	distinct subordinating conjunctions/sentences	temporal nouns/sentences
distinct countable nouns/words	distinct subordinating conjunctions/words	temporal nouns/words
distinct demonstrative pronouns/distinct words	distinct temporal nouns/distinct words	volatile verbs/distinct words
distinct demonstrative pronouns/sentences	distinct temporal nouns/sentences	volatile verbs/sentences
distinct demonstrative pronouns/words	distinct temporal nouns/words	volatile verbs/words
distinct directional co-verb/distinct words	distinct volatile verbs/distinct words	
distinct directional co-verb/sentences	distinct volatile verbs/sentences	
distinct directional co-verb/words	distinct volatile verbs/words	

SYNTAX-LEVEL FEATURES

average height of clauses (parse tree)	average no. clauses	average number of nonterminal nodes (parse tree)
average height of level 1 branches (parse tree)	average no. clauses (remove duplicate)	average number of noun phrases
average highest clauses (parse tree)	average no. conjunction word	average number of prepositional phrases
average length of clauses	average no. content words	average number of terminal nodes (parse tree)
average longest clauses	average no. distinct conjunction word	average number of verb phrase
average longest noun phrases	average no. function words	average tree breadths (parse tree - remove duplicate)
average longest preposition phrases	average no. level 1 branches (parse tree)	average tree breadths (parse tree)
average longest verb phrases	average no. level 1 nonterminal nodes (parse tree)	average tree depths (parse tree - remove duplicate)
average no. brackets (parse tree)	average no. nodes (parse tree - remove duplicate)	average tree depths (parse tree)
average no. branches (parse tree - remove duplicate)	average no. nodes (parse tree)	ratio of simple sentences
average no. branches (parse tree)		

BASIC SEMANTIC FEATURES

ratio of 2-semantic words	average of word semantic/sentences	product of word semantics/words
ratio of 3-semantic words	geometric mean of word semantic/sentences	semantics/sentences
ratio of monosemous words	product of word semantics/sentences	semantics/words
ratio of polysemous words		

RAW-LEVEL LANGUAGE MODEL FEATURES

average character bigram frequencies	average syllable bigram frequencies	average word bigram frequencies
average character bigram perplexity	average syllable bigram perplexity	average word bigram perplexity
average character bigram rankings	average syllable bigram rankings	average word bigram rankings
average character trigram frequencies	average syllable list frequencies	average word list frequencies
average character trigram perplexity	average syllable rankings	average word rankings
average character trigram rankings	average syllable set frequencies	average word set frequencies

average distinct syllable frequency	average syllable set rankings	average word set rankings
average distinct word frequency	average syllable trigram frequencies	average word trigram frequencies
average frequency of sentence length in syllable (remove duplicate)	average syllable trigram perplexity	average word trigram perplexity
average frequency of sentence length in word (remove duplicate)	average syllable trigram rankings	average word trigram rankings
POS-LEVEL LANGUAGE MODEL FEATURES		
average POS bigram frequencies	average POS trigram perplexity	average word with POS bigram rankings
average POS bigram perplexity	average POS trigram rankings	average word with POS trigram frequencies
average POS bigram rankings	average word with POS bigram frequencies	average word with POS trigram perplexity
average POS trigram frequencies	average word with POS bigram perplexity	average word with POS trigram rankings
VIETNAMESE-SPECIFIC FEATURES		
distinct borrowed words/distinct words	ratio of Sino-Vietnamese words	3-syllable Sino-Vietnamese words/words
distinct local words/distinct words	ratio of distinct Sino-Vietnamese words	monosyllabic Sino-Vietnamese-words/Sino-Vietnamese words
distinct Sino-Vietnamese words/distinct words	2-syllable Sino-Vietnamese words/Sino-Vietnamese words	monosyllabic Sino-Vietnamese-words/words
ratio of borrowed words	2-syllable Sino-Vietnamese words/words	polyphonic Sino-Vietnamese words/Sino-Vietnamese words
ratio of distinct borrowed words	3-syllable Sino-Vietnamese words/Sino-Vietnamese words	polyphonic Sino-Vietnamese-words/words
ratio of local words		
ratio of distinct local words		

B. Corpus

Following most of the previous studies on text readability in Vietnamese, this study also used the corpus of 371 literature texts by Luong, et al. [13]. Moreover, the collection and construction of a new dataset for the survey are extremely costly in terms of time and labor, and thus utilizing this existing corpus the optimal option. The research on texts of other domains will be carried out in future studies.

These documents were collected from Vietnamese and Literature textbooks for students in Vietnam. All of these textbooks are written in Vietnamese and published by Vietnam Education Publishing House under the resolution to renovating the program for general education of the National Assembly, The Socialist Republic of Vietnam, in 2000 [26].

In Vietnam, primary education is divided into five years – from grade 1 to grade 5. However, the Vietnamese textbooks for first grade students only include reading and writing exercises for simple characters and words, and thus they were not included in the surveys. The textbook for junior high school students is categorized into four levels, corresponding to four school years – from grade 6 to grade 9. For high school students, the Literature textbooks are partitioned into three levels corresponding to three school years – from grade 10 to grade 12. The Literature textbooks for high school students are also classified into two different sets: (i) a general set for most students and (ii) an advanced set, with more reading, for students specialized in Literature. Table III presents the statistics of the corpus.

To extract the features that we mentioned in Section 3.1 for each document, we took steps to process and label the text. This process consists of the following steps:

Encoding standardization: We standardized the data because the texts were collected from various sources with different encoding methods. For instance, the Vietnamese word "học" (study) consists of three characters – "h", "o", "c" – when this word is encoded in the pre-built Unicode. However, if it is done in the composite Unicode, this word includes 4 characters: "h", "o", "c", and "." (drop-tone). In this article, we converted all the documents into the pre-built Unicode.

TABLE III.   CORPUS STATISTICS

	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12
Number of documents	67	62	40	40	28	13	17	21	15	19	49
Average number of sentences	18.3	19.6	21.5	21.4	54.8	46.4	65.8	107.3	60.7	105.2	111.7
Average number of words	158	192	231	244	680	677	969	1447	862	1360	1710
Average number of syllables	178	222	276	288	784	821	1131	1710	1006	1579	2179
Average number of characters	827	1065	1335	1396	3709	3942	5402	8160	4860	7535	10761
Average number of distinct words	100.6	125.6	144.3	152.8	304.9	329.7	394.3	526.3	368.4	510	576
Average number of distinct syllables	111.4	141.5	164.8	173.4	327.5	372.5	428.4	555.5	390.1	534.9	594.2

Punctuation standardization: Punctuation like the dot (.), comma (,), semi-colon (;), colon (:), exclamation (!), question (?), single quotation ('), double quotation ("), brackets ([ ], (), {}), hyphen (-), slash (/), etc. were separated from their previous words by a space (" "). This enable the texts to appear clearer and the statistical operations in these texts to be more exact.

1) Tone standardization: Similar to encoding, in Vietnamese, there are two ways to place the tone mark. First, the "old style" emphasizes aesthetics by placing the tone mark as close as possible to the center of the word, by placing the tone mark on the last vowel if an ending consonant part exists, and on the next-to-last vowel if the ending consonant does not exist, as in "hóa", "hủy"). Meanwhile, the "new style" emphasizes linguistic principles and applies the tone mark on the main vowel (as in "hoá", "huỷ"). In this work, we converted all texts to the "old style".

2) Sentence segmentation and word segmentation: Sentences and words are two common features of readability research, often being examined in most readability studies – especially in readability formulas. They are also the basic features for other elements, such as part-of-speech (POS), named-entity (NE), dependency tree, or lexical chain, etc. Consequently, the texts were segmented into sentences, which, in turn, were segmented into words.

3) POS tagging: POS features are commonly used in text readability studies, such as Vogel and Washburne [27], Bormuth [28], Al Khalil, et al. [10], among others. Therefore, in this study, we conducted the POS tagging for documents in preparation for extracting features in Section 3.1.

4) Constituency parsing: Syntactic features have been widely exploited in the literature on text readability in the world. However, for Vietnamese, due to limitations on syntax labeling tools and methods, the syntax features remain relatively unexplored, not only with the readability of the text, but also with various other problems in the field of the Vietnamese language. However, recently, the accuracy of studies on automatic constituency parsing in Vietnamese has been significantly improved. In particular, Uyen, et al. [29] has achieved an accuracy rate of 79%. In this study, to effectively examine the syntactic features that affect the level of text readability, we used the results of Phan et al.'s research to parse documents in the corpus.

In this study, we used the CLC_VN_TOOLKIT of the Computational Linguistics Center (CLC)[1] to preprocess, split sentences, separate words, and tag POS. The tool's accuracy data was not disclosed, but our experiments indicates that the accuracy achieved was over 99% for the sentence and word tokenization tasks and over 97% for the POS tagging task.

After all the documents were processed and the necessary labels were assigned, we proceeded to extract the features for the examinations. The extraction of most of the features mentioned in Section 3.1 could be achieved straightforwardly

---

[1] http://www.clc.hcmus.edu.vn

from the processing and labeling steps. However, there were some features require additional support of external corpora, as follows:

1) Easy words and syllables features: In various studies, the ratio of easy words in a text remains a crucially dominant feature in the evaluation of the readability of that text. However, constructing a list of easy words is remarkably costly, as it requires a large number of readers to examine a large number of words. Hence, most studies commonly utilize frequency word lists instead. That is, if a word has a high frequency of use, it is likely that native speakers perceive that word as easy to understand, and vice versa. Likewise, easy syllable features were also implemented in this study. Our target is the readings in Vietnamese and Literature textbooks for students in Vietnam, and thus we used the list of the 3,000 most common words and the 3,000 most common syllables in Vietnamese of Dinh, et al. [30]. If a word appeared in this list of 3,000 common words, it would be treated as an easy word. Other words (including out-of-vocabulary (OOV) words) were treated as not-easy words. Similarly, a syllable was considered an easy syllable if it appeared in this list of 3,000 common syllables. It is possible for a word or syllable to appear more often only in a specific domain of text, and hence, are easier to comprehend to only a particular group of readers, but not to other text domains or other reader groups. In those cases, the list of frequent/easy words/syllables should be different.

2) Sino-Vietnamese features: The Vietnamese culture is strongly influenced by the Chinese culture. The Vietnamese language is also affected, as more than 60% of Vietnamese vocabulary is derived from Chinese, known as Chinese-Vietnamese words. Sino-Vietnamese words are frequently used in scientific texts, technical texts, and formal texts, and they are often considered more difficult than other pure Vietnamese words. Therefore, the ratio of Sino-Vietnamese words was additionally used in this study. We extracted features of Sino-Vietnamese words in the documents using the list of Sino-Vietnamese words from the Vietnamese Dictionary by Phe [31]. Words (including OOV words) that did not appear in this list were not treated as Sino-Vietnamese words.

3) Dialect features: The country of Vietnam stretches over 3,000 km with various diverse regions, each of which has its own culture and language usage. Many regions retain private words habitually used in that region, but not in other places. Therefore, with general texts, especially textbooks, the appearance of the dialect words might affect the readability of the text. Similar to Sino-Vietnamese words, in this study, we also extracted dialect words from the Vietnamese Dictionary by Phe [31] for statistics. Words (including OOV words) that did not appear in this list were not treated as dialect words.

4) Language model features: Language models are often implemented in a broad range of studies on NLP in general and on text readability in particular. Simply stated, a language model is a probability distribution over text sets, indicating how likely a sentence or phrase occurs in a language. The

higher the probability of a sentence or phrase is, the more familiar that sentence or phrase is to the readers. Consequently, that sentence or phrase may be easier to read than the low probability sentence or phrase. In this study, to extract features for the text difficulty problem, we built several language models, which include characters, syllables, words, words with POS, POS-only bi-grams, and tri-grams. The corpus that we utilized to construct the language model is VCor (Vietnamese Corpus) [30]. This corpus consists of 805,000 documents, extracted from a broad range of sources such as news sites, books, and Vietnamese newspapers, etc.

5) Semantic features: Since there is no semantic corpus with sufficiently large quantity to conduct the examination and experiment, no previous studies have focused on the processing or automatic semantic labeling of sentences in Vietnamese. In this study, we extracted basic statistical semantic features, such as the average number of meanings of words in a sentence and Geometric Mean of meanings of sentences in a text, among others. We also used the Vietnamese Dictionary by Phe [31] to conduct statistics on the meaning of words and extract the features that we mentioned in Section 3.1.

6) Text grouping: In this study, we grouped documents in two ways to fit each approach of the text readability assessment problem and match our examination method:

a) By school track: Texts were grouped into three school tracks, which were elementary, middle, and high schools. We grouped documents in this way to conduct features examinations according to the feature evaluation method of the text classification problem.

b) By grade level: Texts were grouped into 11 grade levels according to the curriculum of the general textbook in Vietnam. With this grouping, we investigated the role of the features using correlation and regression analyses.

C. Features Examination

In this study, we conducted surveys that evaluate the impact of NLP features introduced in Section 3.1 on text readability. These evaluations were based on the examinations on the textbook materials for Vietnamese students mentioned in Section 3.2.

We implemented two examination methods corresponding to two approaches of the text readability assessment problem:

1) Statistical approach

This approach mainly implements correlation analysis to identify the features highly correlated with the readability level, thereby extracting the weight of these features through regression analysis method to build formula(s) to predict the difficulty of the texts. This was also the approach used for developing famous text readability formulas such as Dale-Chall [2], Flesch Reading Ease [14], SMOG [3], as well as the first and second formulas for Vietnamese text in Nguyen, et al. [11, 17].

Correlation analysis determines the linear relationship between the quantitative variables in this study, which are the features of the text and the readability level of that text. The higher the correlation coefficient between the two variables is, the higher the degree of their correlation is. The correlation coefficient ranges from -1 to 1. A correlation coefficient of 0 (or nearly 0) indicates that the two variables almost have no contact with each other. Conversely, a coefficient of -1 or 1 signals that the two variables have an absolute relationship. If the value of the correlation coefficient is negative ($r < 0$), it suggests that when the value of one variable increases, the value of the other decreases (and vice versa, when one variable decreases, the other increases). Meanwhile, if the correlation coefficient value is positive ($r > 0$), it means that when one variable increases, the other increases, and vice versa. In this study, we use the Pearson correlation coefficient. Table IV presents a list of features that are highly correlated with the readability level of the text (with a correlation coefficient greater than or equal to 0.8 or less than or equal to -0.8). These features consisted of 13 raw text features, 7 POS features, 2 syntax-level features, 3 basic semantic features, 21 raw-level language model features, 6 POS-level language model features, and 4 Vietnamese specific features. The raw-level language model features and 15 raw text features were most strongly correlated with the readability level of Vietnamese texts with the highest correlation coefficients being 0.91 and 0.85, respectively. Other features like POS, syntax, basic semantic, or Vietnamese specific features were not as strongly correlated as raw-level language model and raw features, but also had high correlation coefficients, from 0.80 to 0.84.

After correlation analysis, we selected features closely related to the difficulty of the text to perform regression analysis. Regression analysis is a statistical technique used to estimate the equation that best fits the set of observations of the dependent variable, which is the text readability level in this study, and the independent variables, which are the features used. Regression analysis allows the best estimation of the true relationship between variables. From this estimating equation, we can predict the dependent variable (the readability level of the text – unknown) based on the given value of the independent variable (the features – known). In regression analysis, if independent variables strongly correlated with each other (high correlation coefficient), multi-collinearity phenomenon will occur. Therefore, independent variables that are strongly correlated with each other are typically removed before the regression analysis. However, during the process of correlation analysis, we found that all the features in Table IV were strongly correlated with each other (the correlation coefficients were $\geq$ 0.7), and thus we conducted two experiments: (1) regression analysis with features in Table IV, with no exclusion of any strongly correlated features, and (2) regression analysis with features that correlate with the readability of text greater than or equal to 0.7, eliminating features that were strongly correlated with each other. We did not remove the strongly correlated features in the first experiment because the feature that had the highest correlation with the text readability level – average word set rankings – was also strongly correlated with the remaining features, with correlation coefficient values $\geq$ 0.8. For the second experiment, we selected the features with the correlation coefficient with text difficulty $\geq$ 0.7 and removed the features

that correlated with the selected features ≥ 0.8. As a result, the remaining number of features is only three. If we were to lower the elimination threshold to 0.7, only one feature with the highest correlation coefficient would have been chosen. Table V and Table VI present the intercept scores the coefficients of the features in the estimation equation after regression analysis of both experiments. Table VII shows the correlation of the two estimation equations in our experiments with the text difficulty along with (i) the most correlated feature in our experiment (average word set rankings), (ii) the two text readability measurement formulas of Nguyen and Henkin [11, 17] and their revised version on our experiment corpus, and (iii) the revised version of the formula of Luong, et al. [18]. The correlations of the estimation equations with the text difficulty of the first experiment, the second experiment,

and the highest feature (average word set rankings) were 0.95, 0.92 and 0.91, respectively. Hence, while the elimination of strongly correlated features reduced the number of features to be analyzed and minimized processing costs in the text evaluation process, it also lowered the correlation between the estimated equation and the readability of the text. Meanwhile, the experimentation using the two formulas of Nguyen and Henkin [11, 17] on the set of readings in Vietnamese textbooks and Literature in Vietnam at the present yielded the correlation results of only about 0.51 and 0.58, respectively. When we updated the weights of Nguyen and Henkin's formulas [11, 17] and Luong, et al. [18] using our corpus, the correlation with the text readability increased, but it was not as high as the result in our first experiment.

TABLE IV.   LIST OF FEATURES HIGHLY CORRELATED WITH THE TEXT READABILITY LEVE

RAW FEATURES			
average word length in syllable	**0.853269**	distinct easy word/distinct words	-0.84908
average word length in character	0.844346	ratio of easy syllables	-0.85065
distinct easy syllables/distinct syllables	0.835926	ratio of easy words	-0.86098
ratio of long sentence (in syllable)	0.818193	ratio of monosyllabic words	-0.86667
ratio of long sentence (in word)	0.809846	ratio of distinct easy syllables	-0.86977
ratio of short sentence (in word)	-0.80448	ratio of distinct easy words	-0.8816
ratio of short sentence (in syllable)	-0.81497		
PART-OF-SPEECH FEATURES			
POS tags/words	-0.8304	adverbs/words	**-0.80988**
ratio of 2-POS tag words	-0.81505	distinct volatile verbs/words	-0.817
ratio of 3-POS tag words	-0.81994	distinct adverbs/words	-0.81554
ratio of multi POS tag words	-0.84525		
SYNTAX-LEVEL FEATURES			
average tree depths (parse tree)	**0.822985**	ratio of simple sentences	-0.81698
BASIC SEMANTIC FEATURES			
semantics/words	-0.82351	ratio of polysemous words	**-0.83606**
ratio of 3-semantic words	-0.82913		
RAW-LEVEL LANGUAGE MODEL FEATURES			
average word set rankings	**0.911331**	average distinct word frequency	-0.83279
average word set frequencies	0.895034	average syllable bigram frequencies	-0.8403
average word list frequencies	0.885074	average frequency of sentence length in word (remove duplicate)	-0.84562
average word rankings	0.863239	average syllable set frequencies	-0.84672
average word trigram frequencies	0.843268	average frequency of sentence length in syllable (remove duplicate)	-0.8502
average syllable trigram frequencies	0.842053	average syllable list frequencies	-0.8535
average syllable set rankings	-0.81599	average character bigram frequencies	-0.86795
average word bigram frequencies	-0.81744	average character trigram frequencies	-0.86852
average syllable bigram rankings	-0.82157	average character bigram rankings	-0.86854
average syllable rankings	-0.82241	average character trigram rankings	-0.86937
average distinct syllable frequency	-0.82974		
POS-LEVEL LANGUAGE MODEL FEATURES			
average word with POS trigram frequencies	**0.846458**	average POS trigram perplexity	-0.82213
average POS bigram perplexity	-0.81658	average POS trigram frequencies	-0.82706
average word with POS bigram frequencies	-0.8171	average POS bigram frequencies	-0.83434
VIETNAMESE-SPECIFIC FEATURES			
distinct borrowed words/distinct words	0.824652	ratio of borrowed words	0.814249
distinct Sino-Vietnamese words/distinct words	0.819849	monosyllabic Sino-Vietnamese words/Sino-Vietnamese words	**-0.83381**

TABLE V.   EXPERIMENTAL RESULTS OF THE FIRST REGRESSION ANALYSIS

Intercept	76.76817		
RAW FEATURES			
average word length in character	0.062244	ratio of easy words	-25.0171
average word length in syllable	-4.98138	ratio of long sentence (in syllable)	0.805849
distinct easy syllables/distinct syllables	0.382055	ratio of long sentence (in word)	0.186385
distinct easy word/distinct words	2.023699	ratio of monosyllabic words	-40.62
ratio of distinct easy syllables	-20.0669	ratio of short sentence (in syllable)	8.068247
ratio of distinct easy words	7.342935	ratio of short sentence (in word)	5.744779
ratio of easy syllables	43.08403		
PART-OF-SPEECH FEATURES			
POS tags/words	9.515384	ratio of 2-POS tag words	-4.60083
adverbs/words	-59.9567	ratio of 3-POS tag words	-0.61682
distinct adverbs/words	-25.5114	ratio of multi POS tag words	-16.419
distinct volatile verbs/words	-1.47296		
SYNTAX-LEVEL FEATURES			
average tree depths (parse tree)	0.034768	ratio of simple sentences	-5.72399
BASIC SEMANTIC FEATURES			
ratio of 3-semantic words	19.51943	semantics/words	0.923996
ratio of polysemous words	-21.6204		
RAW-LEVEL LANGUAGE MODEL FEATURES			
average character bigram frequencies	10.45311	average syllable rankings	3.716677
average character bigram rankings	6.02605	average syllable set frequencies	-44.0778
average character trigram frequencies	-3.62059	average syllable set rankings	8.422091
average character trigram rankings	-4.58448	average syllable trigram frequencies	-0.10974
average distinct syllable frequency	0.009618	average word bigram frequencies	38.36505
average distinct word frequency	-0.06197	average word list frequencies	0.070119
average frequency of sentence length in syllable (remove duplicate)	0.005918	average word rankings	4.84E-05
average frequency of sentence length in word (remove duplicate)	-0.00603	average word set frequencies	-3.646
average syllable bigram frequencies	0.208769	average word set rankings	0.002689
average syllable bigram rankings	1.633263	average word trigram frequencies	0.342254
average syllable list frequencies	-22.6624		
POS-LEVEL LANGUAGE MODEL FEATURES			
average POS bigram frequencies	-3.89573	average POS trigram perplexity	-15.3692
average POS bigram perplexity	7.087669	average word with POS bigram frequencies	-14.0184
average POS trigram frequencies	30.56821	average word with POS trigram frequencies	-0.21931
VIETNAMESE-SPECIFIC FEATURES			
distinct borrowed words/distinct words	-0.6103	ratio of borrowed words	0.269794
distinct Sino-Vietnamese words/distinct words	0.232131	monosyllabic Sino-Vietnamese words/Sino-Vietnamese words	-10.2129

TABLE VI.   EXPERIMENTAL RESULTS OF THE SECOND REGRESSION ANALYSIS

Intercept	2.808379
volatile verbs/sentences	0.003871
common nouns/words	-73.0814
average word set rankings	0.001179

TABLE VII.   CORRELATION COEFFICIENTS OF TWO EXPERIMENTS AND TWO READABILITY FORMULAS OF NGUYEN AND HENKIN [11, 17]

Nguyen and Henkin (1982)	0.51
Nguyen and Henkin (1985)	0.58
Nguyen and Henkin (1982) (revised)	0.85
Nguyen and Henkin (1985) (revised)	0.82
Luong et al. 2018 (revised)	0.87
Only use "average word set rankings"	0.91
Experiment 1	**0.95**
Experiment 2	0.92

2) Machine learning approach

This approach evaluates the role of features in the text classification problem according to the difficulty level. In this study, we used an algorithm called Feature ranking with recursive feature elimination and cross-validated selection of the best number of features (RFECV). Initially, all the features that are examined will be used to classify texts by readability level. The documents will be classified and evaluated by an SVM classification algorithm, using k-fold cross-validation, which splits the corpus into k parts, and then takes k - 1 part for training and the rest part for testing. The features are then removed gradually to test the accuracy of the combination of each feature. Finally, the algorithm evaluates the best combination of documents to classify documents according to their difficulty level. This algorithm has been implemented in the sklean library [32] in Python.

In this experiment, we eliminated from 1 to n-1 number of features, with n being the number of examined features, k = 5, and the evaluation criterion was the classification accuracy. Fig. 1 presents the results of the examination on the number of features and the accuracy achieved through the RFECV algorithm. With about 7 features, the accuracy of the classification process was the highest (85.7%). Table VIII presents the most highly ranked features surveyed by the RFECV algorithm. Out of these 7 features, 6 were raw-level (including 4 language model features), and 1 was Vietnamese-specific feature, with no semantic level features. When compared with the results of 85.17% in the work of Luong, et al. [13] for Vietnamese text, this combination of the seven features achieved slightly higher results with the rate of 85.7%. However, Luong, et al. [13] used some non-standardized text length features, such as numbers of sentences, words, syllables, characters, distinct words, and distinct syllables. These characteristics have proven to be valuable in assessing the difficulty of text in textbooks, when reading time is limited

within the framework of a lesson [12]. Therefore, we also conducted an empirical evaluation of the features mentioned in 3.1 together with non-standard text length features. Fig. 2 presents the ranking result and Table IX lists most highly ranked features in this experiment, including a non-standardized feature (number of words), 16 raw-level features, 5 POS-level features, 2 syntax-level features, 9 language model features, 4 Vietnamese-specific features, and no semantic level characteristics. It was possible that the semantic-level features were highly correlated with the readability level but were not suitable for the construction of a readability evaluation model. Another possibility would be that the features examined were too simple or inappropriate with the corpus in question. Other in-depth studies on these characteristics are needed to evaluate these possibilities. Table X presents the accuracy rates of the recent publications of Luong, et al. [12, 13] and of our two experiments on text readability classification on the corpus of Vietnamese and Literature textbooks. With 24 features (including non-standardized length features), the accuracy rate of the classification process was 88.14%, which was higher than those of Luong, et al. [12] and Luong, et al. [13] by 3% to 4%.
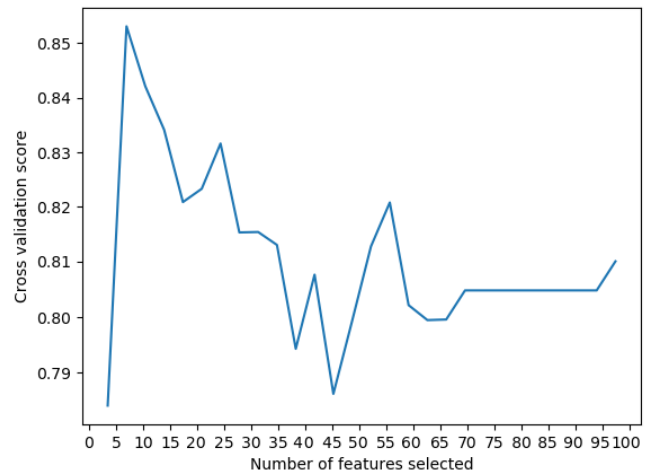
Fig. 1.   Experiment Result on the Numbers of Features (without Non-Standardized Length Features).
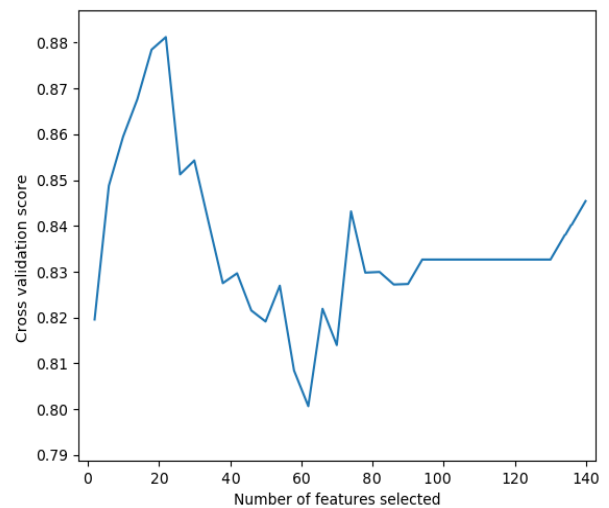
Fig. 2.   Experiment Result on the Numbers Of Features (with Non-Standardized Length Features).

TABLE VIII. MOST HIGHLY RANKED FEATURES (WITHOUT NON-STANDARDIZED LENGTH FEATURES)

average word length in syllable
distinct easy syllables/distinct syllables
average word set frequencies
average word list frequencies
average syllable trigram frequencies
average syllable bigram rankings
distinct Sino-Vietnamese words/distinct words

TABLE IX. MOST HIGHLY RANKED FEATURES (WITH NON-STANDARDIZED LENGTH FEATURES)

number of words
average word length in character
ratio of long sentence (in syllable)
ratio of long sentence (in word)
distinct common nouns/distinct words
distinct parallel conjunctions/distinct words
ratio of single POS tag words
adverbs/sentences
average no. distinct conjunction word
average no. conjunction word
average word set frequencies
average word list frequencies
average word trigram frequencies
average syllable trigram frequencies
average word bigram frequencies
average syllable rankings
average syllable set rankings
average syllable bigram rankings
average word with POS trigram frequencies
ratio of borrowed words
ratio of Sino-Vietnamese words
ratio of distinct borrowed words
ratio of distinct Sino-Vietnamese words
polyphonic Sino-Vietnamese words/Sino-Vietnamese words

TABLE X. ACCURACY RATES OF THE TEXT CLASSIFICATION MODELS BY READABILITY, USING 69 SELECTED FEATURES, COMPARED WITH PREVIOUS WORKS

Luong et al. (2017)	84.34
Luong et al. (2018)	85.17
Our experiment (without non-standardized length features)	85.70
Our experiment (with non-standardized length features)	88.14

IV. CONCLUSIONS

In this study, we examined the effects of linguistic features at all levels on the readability assessment of Vietnamese texts. We extracted a total of 271 features from Vietnamese textbooks for primary school students and Literature for middle and high school students in Vietnam to explore. These features range from superficial and word-level features to grammatical and fundamental semantic features. We also surveyed the n-gram features to evaluate the role that the language model plays in determining the difficulty of Vietnamese text.

We conducted the examinations in two main approaches to the readability problem: the statistical approach and the machine learning approach. For the statistical approach, we performed a correlation analysis of 271 features with the difficulty of the surveyed documents and selected 56 highly correlated features, with the correlation coefficient values $\geq$ 0.8. Next, we used these 56 features to perform a regression analysis to find the coefficients of the features in the formula to predict the readability of the text. Empirical results indicated that the estimation equation built from these 56 features was highly correlated with the difficulty of the text, with the correlation coefficient of 0.95, significantly higher than previous studies of Nguyen and Henkin [11, 17]. Regarding the machine learning approach, we evaluated the role of features in text classification according to the readability level. The evaluating algorithm used was feature ranking with recursive feature elimination and cross-validated selection of the best number of features (RFECV). This algorithm examined specific combinations in the text classification problem to ranked features, utilizing SVM to model classification and K-fold cross-validation to avoid over-fitting. Experimental results show that, with seven features, most of which were shallow features and language model features, the accuracy of the classification model obtained the highest accuracy (~85.7%). When experimenting with additional non-standardized text length features, the classification results showed a significant improvement over the existing features of Luong et al. [12, 13].

For future works, we will collect additional corpora on different domains to explore the features that would be useful in evaluating the readability of documents in those domains. Deeper features at the semantic level such as coherence and cohesion will also be investigated to detect better combinations for assessing the readability of Vietnamese text.

ACKNOWLEDGMENT

CONFLICT OF INTEREST

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

REFERENCES

[1] A. Bailin and A. Grafstein, *Readability: Text and Context*. Palgrave Macmillan UK, 2016.

[2] J. S. Chall and E. Dale, *Readability Revisited: The New Dale-Chall Readability Formula*. Northampton, Massachusetts: Brookline Books, 1995.

[3] G. H. Mc Laughlin, "SMOG grading-a new readability formula," *Journal of reading,* vol. 12, no. 8, pp. 639-646, 1969.

[4] L. Si and J. Callan, "A Statistical Model for Scientific Readability," in *Proceedings of the Tenth International Conference on Information and Knowledge Management*, New York, NY, USA, 2001, pp. 574-576: ACM.

[5] K. Collins-Thompson and J. Callan, "Predicting Reading Difficulty with Statistical Language Models," *J. Am. Soc. Inf. Sci. Technol.,* vol. 56, no. 13, pp. 1448-1462, November 2005.

[6] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," in *Proceedings of the conference on empirical methods in natural language processing*, 2008, pp. 186-195.

[7] S. Vajjala and D. Meurers, "On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, Montréal, Canada, 2012, pp. 163-173: Association for Computational Linguistics.

[8] M. Sinha and A. Basu, "A study of readability of texts in Bangla through machine learning approaches," *Education and Information Technologies,* vol. 21, no. 5, pp. 1071-1094, 2016.

[9] S. Vajjala and I. Lučić, "OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, 2018, pp. 297-304: Association for Computational Linguistics.

[10] M. Al Khalil, H. Saddiki, N. Habash, and L. Alfalasi, "A Leveled Reading Corpus of Modern Standard Arabic," in *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan, 2018: European Language Resource Association.

[11] L. T. Nguyen and A. B. Henkin, "A Second Generation Readability Formula for Vietnamese," *Journal of Reading,* vol. 29, no. 3, pp. 219-225, 1985.

[12] A.-V. Luong, D. Nguyen, and D. Dinh, "Examining the text-length factor in evaluating the readability of literary texts in Vietnamese textbooks," in *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, 2017, pp. 36-41.

[13] A.-V. Luong, D. Nguyen, and D. Dinh, "Assessing the Readability of Literary Texts in Vietnamese Textbooks," in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, 2018, pp. 231-236.

[14] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Technical Training,* vol. Research B, no. February, p. 49, 1975.

[15] S. Zeno, T. A. S. Associates, R. T. Millard, and R. Duvvuri, *The Educator's Word Frequency Guide*. Touchstone Applied Science Associates, 1995.

[16] J. H. Lee and Y. Hasebe, "Readability measurement for Japanese text based on leveled corpora," *Papers on Japanese Language from an Empirical Perspective, Ljubljana: Academic Publishing Division of the Faculty of Arts, Univ. of Ljubljana,* 2016.

[17] L. T. Nguyen and A. B. Henkin, "A Readability Formula for Vietnamese," *Journal of Reading,* vol. 26, no. 3, pp. 243-251, 1982.

[18] A.-V. Luong, D. Nguyen, and D. Dinh, "A New Formula for Vietnamese Text Readability Assessment," in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 2018, pp. 198-202.

[19] Đ. T. N. Nguyễn, A.-V. Lương, and Đ. Đinh, "Affection of the part of speech elements in Vietnamese text readability," *Acta Linguistica Asiatica,* vol. 9, no. 1, 01/30 2019.

[20] F. Dell'Orletta, M. Wieling, G. Venturi, A. Cimino, and S. Montemagni, "Assessing the Readability of Sentences: Which Corpora and Features?," in *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, Baltimore, Maryland, 2014, pp. 163-173, references_files/DellOrletta2014.pdf: Association for Computational Linguistics.

[21] O. De Clercq, V. Hoste, B. Desmet, P. Van Oosten, M. De Cock, and L. Macken, "Using the crowd for readability prediction," *Natural Language Engineering,* vol. 20, no. 3, pp. 293-325, 2014.

[22] K. Berendes *et al.*, "Reading Demands in Secondary School: Does the Linguistic Complexity of Textbooks Increase With Grade Level and the Academic Orientation of the School Track?," *Journal of Educational Psychology,* vol. 110(4), pp. 518–543, November 2017.

[23] G. Sun, Z. Jiang, Q. Gu, and D. Chen, "Linear model incorporating feature ranking for Chinese documents readability," in *The 9th International Symposium on Chinese Spoken Language Processing*, Singapore, 2014, pp. 29-33, references_files/Sun2014.pdf: IEEE.

[24] Y.-H. Chen and P. Daowadung, "Assessing readability of Thai text using support vector machines," *Maejo International Journal of Science and Technology,* vol. 09, no. 3, pp. 355-369, 2015.

[25] H.-C. Tseng, B. Chen, T.-H. Chang, and Y.-T. Sung, "Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts," *Natural Language Engineering,* vol. 25, no. 3, pp. 331-361, 2019.

[26] *Nghị quyết số 40/2000/NQ-QH10 của Quốc hội Khóa 10 Nước Cộng hòa Xã hội Chủ nghĩa Việt Nam về đổi mới chương trình giáo dục phổ thông (Resolution No. 40/2000/NQ-QH10 of the 10th National Assembly of the Socialist Republic of Vietnam on Renovating the Program for General Education)*, 2000.

[27] M. Vogel and C. Washburne, "An Objective Method of Determining Grade Placement of Children's Reading Material," *The Elementary School Journal,* vol. 28, no. 5, pp. 373-381, 1928.

[28] J. R. Bormuth, "Readability: A New Approach," *Reading Research Quarterly,* vol. 1, no. 3, pp. 79-132, 1966.

[29] P. T. P. Uyen, N. T. H. Tung, T. Thinh, and D. An, "Vietnamese Span-based Constituency Parsing with BERT Embedding," in *The 11th International Conference on Knowledge and Systems Engineering (KSE 2019)*, Da Nang, Vietnam.

[30] D. Dinh, T. N. Nguyen, and H. T. Ho, "Building a corpus-based frequency dictionary of Vietnamese," ed, 2018, pp. 72-98.

[31] H. Phe, *Từ điển tiếng Việt (Vietnamese dictionary)*, 8th ed. Da Nang Publishing House, 2017.

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

An ACM\IEEE and ABET Compliant Curriculum and Accreditation Management Framework

Manar Salamah Ali

Computer Science Department
King Abdulaziz University
Jeddah, Saudi Arabia

Abstract—Following methodological and systemized approaches in creating course syllabi and program curriculums are very crucial for assuring the coherence (correctness, completeness, consistency, and validity) of curriculums. Furthermore, designing coherent curriculums have a direct impact on achieving curriculum outcomes. For institutions seeking accreditation, presenting evidence of curriculum coherence is mandatory. In this paper, a general framework architecture for curriculum and accreditation management is proposed. Furthermore, we propose a detailed design for a knowledge base that comprises of: a) the ACM\IEEE body of knowledge for the Computer Science Department, b) course syllabi, and c) course articulation matrices. We show how to utilize the proposed knowledge base in the quality improvement life cycle, in ABET accreditation, and as a significant step towards curriculum coherence.

Keywords—Curriculum coherence; body of knowledge; accreditation; knowledge base design; ABET

I. INTRODUCTION

In recent years, educational institutions are much concerned with following world standards for creating their educational program curriculums. Also, most institutions are seeking world accreditations from world accrediting institutions such as the Accreditation Council for Business Schools and Programs (ACBSP) [1] or the Accreditation Board for Engineering and Technology (ABET) [2].

The Body of Knowledge (BoK) in undergraduate programs provides core knowledge areas in the field and guidelines for creating curriculums. It also provides a detailed specification of what content should be included in an undergraduate program [3].

Periodical assessment of curriculums based on learning outcomes, accreditation standards, technological changes, and professional requirements benefits significantly from the bodies of knowledge [22, 23]. Furthermore, coherent alignment of curriculum components such as learning outcomes, assessment methods, teaching methods, and program outcomes increases student success as much as two standard deviations [13].

For these reasons, it is necessary to follow methodological approaches for the course and curriculum designs. These approaches must be compliant with the standards of the bodies of knowledge and accrediting institutions. Furthermore, automating the necessary processes for creating curriculums

and evaluating its outcomes is significant for accurate and consistent assessments and evaluations.

A plethora of approaches for assessing and evaluating curriculum content and outcomes can be found in the literature. Nevertheless, most of these approaches assume manual-conducted processes when it comes to building curriculum content, which can result in the poor design of curriculums and jeopardize its coherence. There appears to be a lack of focus on proposing methods that measure the extent of compatibility and alignment of the curriculum content with the BoK. The BoK provides a consistent platform for curriculum and student outcome evaluation. Hence, there is a need for management frameworks that provide holistic and systemized solutions and utilize the BoK as a source for curriculum design, assessment, and accreditation processes.

The purpose of this work is three-fold. The first objective is to introduce a BoK-based curriculum management framework purposely designed to fine-tune the coherence of program curricula. The second objective is to provide the design of a knowledge-based system for supporting curriculum and accreditation management. The third objective is to demonstrate the use of the proposed management framework for accreditation and continuous improvement of the curriculum.

The first computer science BoK, as a standard, was delivered as a result of a joint task between ACM and IEEE organizations back in June of 1882 as an attempt to set standards and facilitate the assessment of educational programs. Since then, the ACM\IEEE produced three consecutive standards for the CS program: 2001, 2008, and 2013, respectively.

In this paper, we provide general framework architecture for managing the CS2013-BoK, course\curriculum design, and course articulation matrices. An articulation matrix design is proposed as the first step towards fully automated accreditation management. In this work, we propose a general knowledge base that is necessary for the continuous quality improvement of curriculums and for accreditation tasks. The work in this paper is compliant with both ACM|IEEE and ABET standards. The proposed management framework and the knowledge base will have a direct impact on curriculum coherence and a direct role in auto-generating reports for ABET: criteria 3 = "Student Outcomes", 4 = "Continuous Improvement", and 5 = "Curriculum" of the Self Study Report (SSR) [4].

The paper is organized as follows: Literature review is presented in Section II. The primary motivation of this research is discussed in Section III. An enhanced continuous improvement cycle model is introduced in Section IV. The proposed framework architecture is discussed in Section V. In Sections VI and VII, the requirements and design of the BoK and the course design modules are discussed. The articulation matrix as a tool for assessment is discussed in Section VIII. Related systems and conclusions are discussed in Sections IX and X, respectively.

II. LITERATURE REVIEW

The quality of academic programs is highly related to efficient and coherent curriculums where optimal content delivery is achieved[12]. Also, empirical results show that there is a direct relationship between curriculum coherence and student achievements of outcomes [13, 16, 24]. Many factors directly affect curriculum coherence, such as: 1) misalignment of program outcomes and accumulative course learning outcomes, 2) content redundancy, 3) content inconsistency, and 4) incomplete content with regard to core knowledge in the underlying area and required competencies. Achieving curriculum coherence requires different types of analysis applied to its content to ensure that content gaps and overlaps are addressed and remedied.

Applying improvements to the content of curriculums is driven by two interrelated processes: continuous improvement process and accreditation. The continuous improvement process is an ongoing sustainable method to improve the overall quality of an academic program in terms of the performance of individual courses, student learning, and program outcomes. The improvement process is incremental and iterative. According to ABET, the cycle encompasses iterative enactments of three main tasks: assessment, evaluation, and improvement. In assessment, data that is necessary for evaluation is collected from different resources such as direct and indirect assessment methods. In evaluation, the collected data is interpreted and analyzed to measure the level of student attainment with regard to outcomes. Consequently, improvements are applied at the course and program level according to the evaluation results [14]. At the course level or curriculum level, improvements may necessitate revisiting course topics and learning outcomes when a lack of curriculum coherence is evident.

Accreditation is a review process where institutions are required to provide evidence that their educational programs meet defined standards of quality. The continuous improvement of curriculums is an integral part of accreditation. For example, institutions must provide proof of the different analysis methods they are using to improve the quality of their program curriculums.

One of the main approaches for increasing the quality of education is the standardization of bodies of knowledge from which program curriculums are constructed. Some examples include Common BoK (CBK) for security management[5], the Business Analysis BoK (BABOK)[6], Landscape Architecture BoK (LABOK)[7], and ACM\IEEE (CS2013) for computer science[3].

Different methods have been proposed to analyze and assess curriculum content. For example, in [21], the authors used a network model to represent curriculum mapping. Curriculum mapping is a process used to identify gaps, overlaps, and misalignments by indexing the curriculum's constituent entities and applying appropriate methods to analyze its content. In comparison, the work in [19] follows a mixed method for evaluating program performance using tree representations for different curriculum and program associated entities. The authors in [25] introduced a risk management framework for providing data-driven decision-making to curriculum and program quality improvement. Curriculum coherence is evaluated in [12] using ontologies and natural language processing techniques.

Current approaches and analysis techniques focus on the existing content of curriculums without measuring the compatibility of the content with the standard requirements of the scientific knowledge in the underlying domain. The alignment of the curriculum with the BoK is significant to the development and improvement processes [18]. As a result, there is an evident gap in the literature. There is a need for proposing general and flexible frameworks, which uses the BoK as the source for curriculum contents and a benchmark for any further improvements and assessments.

III. MOTIVATION

The primary motivation of this work is to encourage curriculum and accreditation committees within academic programs to: 1) integrate the BoK in their processes of course and curriculum design, 2) create a knowledge base for the BoK, and 3) automate the process of course and curriculum design, 4) automate the process of assessment, evaluation, and improvement, and 5) enrich the knowledge base with periodical assessment and evaluation data.

The manual construction or update of the course\curriculum design might suffer from the following serious problems:

1) Curriculum incoherence: the curriculum can suffer from inconsistencies, redundancies, or incompleteness in terms of topics and learning outcomes. For example, the curriculum might suffer from overlapped topics, outdated topics, or missing core topics.

2) Difficulty or possible inaccuracy in retrieving statistics about courses, topics, or outcomes.

3) Difficulty or possible inaccuracy in writing ABET Criterion 3, 4, and 5 in the SSR report.

4) Potential difficulty\inaccuracy in complying with a BoK standard such as ACM\IEEE.

5) Difficulty in regular revisions of the program curricula

The first problem could be caused by course designers during the process of course design, while problems 2-5 are challenging issues that face program accreditation committees and quality assurance personnel.

To prevent such potential problems and inaccuracies, it is necessary to automate the process of creating the course

design. In addition, automation is an essential step towards the full automation of the assessment and evaluation processes.

In the following sections, the proposed framework architecture and design are presented.

IV. CONTINUOUS IMPROVEMENT CYCLE

In this work, we suggest an enhanced model of the continuous improvement cycle. As seen in Fig. 1, the model includes the BoK as a compulsory source for the course and curriculum improvements. The BoK acts as a source for core knowledge areas, topics, and learning outcomes as well as a guideline for coherent and correct curricula.
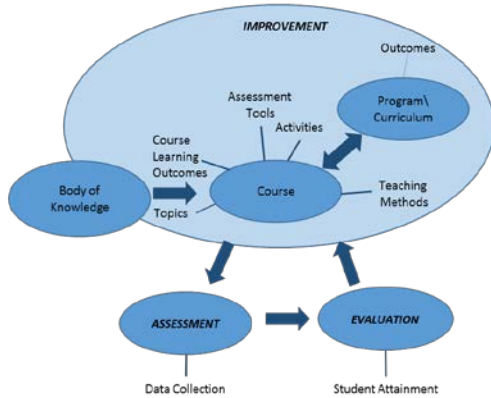
Fig. 1.   An Enhanced Continuous Improvement Cycle.

V. FRAMEWORK ARCHITECTURE

The necessary modules of the proposed framework are illustrated in Fig. 2. In the BoK module, the content of the BoK is stored in a database. The curriculum builder module automates the process of designing courses and stores individual course designs in a database. These modules must be supported with flexible and user-friendly interfaces for entering and editing data related to the BoK and course design. The stored data will hold as a knowledge base for assessment and evaluation-related tasks. The accreditation management module may include different tools that support the accreditation process according to institutions' needs.
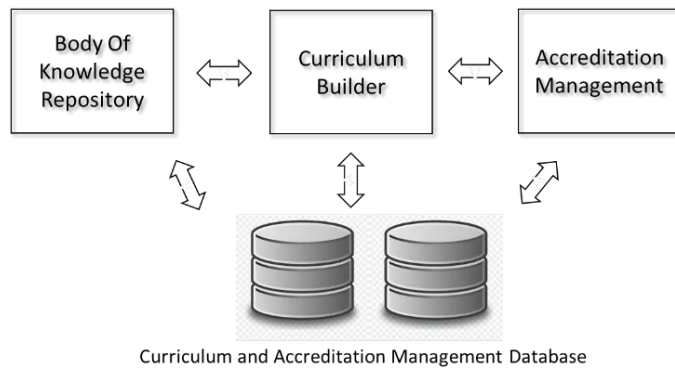
Fig. 2.   Framework Architecture.

A group of senior students has developed a prototype system for storing the BoK and for building course syllabi (see Acknowledgement). The system was developed on Microsoft SQL Server 2014 Management Studio, user interfaces designed with Visual Studio 2015, and the back-code programming in ASP.Net. Snapshots of the system are listed in Appendix A (Fig. 9 to 11).

For increased generalization, in this paper, we provide a general Database (DB) design for the BoK and curriculum builder modules to help interested developers implement them in their chosen platform environment. The proposed design is applicable to CS programs. However, it can be easily modified to serve other programs such as Information Technology, Information systems, Software Engineering, or Computer Engineering.

In the following sections, the design of the BoK and the curriculum builder are discussed in detail. For each module, the data requirements, the DB design in ER diagrams and relational notation, and implementation tips are provided.

VI. BODY OF KNOWLEDGE REPOSITORY MODULE

A. Data Requirements

The data requirements for the BoK has been elicited from the computer science CS2013 report [3]. The detailed discussion of the requirements is listed in the following subsections.

1) Body of knowledge structure

In the latest CS2013 report, the CS BoK is organized into a hierarchical structure. The BoK has 18 Knowledge Areas (KA), and each KA is decomposed into Knowledge Units (KU). Within each KU a set of topics and their expected Course Learning Outcomes (CLO) are provided. The topics and their CLOs are categorized into core-tier1, core-tier2, and electives. The CS2013 guidelines offer great flexibility in consolidating topics into courses and curricula, and institutions must develop their methods of combining topics from the BoK into courses.

The curriculum-topic recommended coverage is as follows: 100% coverage of topics in core-tier-1, 90-100% coverage of topics in core-tier-2, and significant coverage of elective topics.

2) Topic coverage

The basic unit for topic coverage is an "hour", which represents the time required to present the topic material in a traditional lecture. For each KA and KU, the minimum required coverage hours for tier1 and tier2 are provided. In addition, it is stated whether the KU has elective topics or not. The topic coverage scheme is illustrated in Fig. 3.

3) Mastery of CLOs

The CLOs in the CS2013 are classified according to three mastery levels: familiarity, Usage, or Assessment. However, each institution can apply different classification for the CLO mastery levels, such as blooms taxonomy. In Fig. 4, the topics and CLOs of the "Networked Applications" Knowledge Unit is demonstrated.

NC. Networking and Communication
(3 Core-Tier1 hours, 7 Core-Tier2 hours)

KU	Core-Tier1 hours	Core-Tier2 hours	Includes Electives
NC/Introduction	1.5		N
NC/Networked Applications	1.5		N
NC/Reliable Data Delivery		2	N
NC/Routing And Forwarding		1.5	N
NC/Local Area Networks		1.5	N
NC/Resource Allocation		1	N
NC/Mobility		1	N
NC/Social Networking			Y

Fig. 3. Coverage hours for KA="NC" and its Constituent KUs. (CS2013).

NC/Networked Applications
[1.5 Core-Tier1 hours]
Topics:
• Naming and address schemes (DNS, IP addresses, Uniform Resource Identifiers, etc.)
• Distributed applications (client/server, peer-to-peer, cloud, etc.)
• HTTP as an application layer protocol
• Multiplexing with TCP and UDP
• Socket APIs
Learning Outcomes:
1. List the differences and the relations between names and addresses in a network. **[Familiarity]**
2. Define the principles behind naming schemes and resource location. **[Familiarity]**
3. Implement a simple client-server socket-based application. **[Usage]**

Fig. 4. Topics and CLOs in KA="NC and KU="Networked Applications". (CS2013).

4) Cross referencing

In the BoK, topics can be shared between more than one KU. Hence, some of the KUs\topics are cross-referenced with other KUs in the BoK. For example, the Information Assurance and Security (IAS) Knowledge Unit has 9 combined core hours. However, there are 63.5 more hours distributed over other KUs such as OS\security or SDF\Development Methods. Hence, related and relevant topics may exist in different KUs. Thus, this an important issue that must be put into consideration when designing the DB.

B. Database Design

The ER diagram for the BoK is provided in Fig. 5, and its corresponding mapping in relational notation is represented in Table I. All the relations are normalized to third normal form.
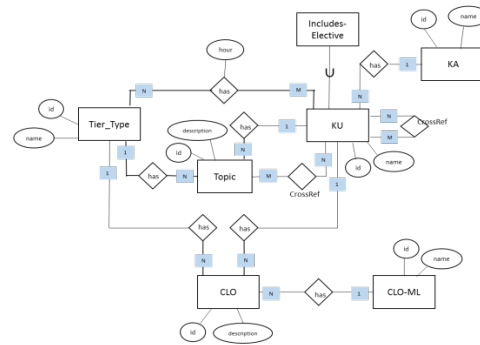
Fig. 5. The ER Diagram for Storing the BoK.

TABLE I. DESIGN OF BODY OF KNOWLEDGE DB IN RELATIONAL NOTATION

Relation	description	Foreign Key	Referenced relation
TIER-TYPE(id,name)	An enumeration relation for storing the types of tiers: core-tier1, core-tier2, and elective.		
CLO-MasteryLevel(id,name)	An enumeration relation for storing the different CLO mastery levels like Familiarity, Usage, and Assessment.		
KA(id, abbreviation, name)	Master Relation for storing the KAs, their abbreviations, and complete name.		
KU(id, name , ka-id)	Master Relation for storing the knowledge units: universal id in the system, name of unit, and the KA it belongs to.	ka-id	KA.id
TOPIC(id, description, tt-id, ku-id)	Master relation for storing the topics, with their universal system id, description, the tier type of the topic, and the KU it belongs to.	ku-id	KU.id
		tt-id	TIER-TYPE.id
CLO(id, description,tt-id,ku-id, cloMl-id)	Master relation for storing course learning outcomes, its universal id in the system, its tier type, the KU it belongs to, and the course mastery level.	ku-id	KU.id
		tt-id	TIER-TYPE.id
		cloMl-id	CLO-MasteryLevel.id
KU-TIER-HOUR (ku-id,tt-id,hour)	Cross Reference relation between KU and TIER-TYPE to store the number of hour coverage for each tier within a specific KU.	ku-id	KU.id
		tt-id	TIER-TYPE.id
Include-Elective(ku-id)	Specialization of KU, stores the id of KUs that have elective topics	ku-id	KU.id
KU-CrossRef(ku-id, ku-cr-id)	Cross Reference relation-Recursive M:N relationship of KU	ku-id, ku-cr-id	KU.id
Topic-Crossref(t-id, Ku-id)	Cross Reference relation between TOPIC and KU	t-id	TOPIC.id
		ku-id	KU.id

C. Implementation Tips

The following are some implementation tips for the development of the BoK repository:

- The interfaces for storing the BoK shall allow inserting, editing, deleting KAs, KUs, topics.

- For cross-referenced KUs and topics, users must follow the guidelines provided by the CS2013 report.

It is important to note that topics and CLOs outside the CS2013 report can be added to the BoK. Also, CLO mastery levels can be different than the ones provided in the report.

VII. COURSE\CURRICULUM DESIGN

Each institution retains its own policies for course\curriculum design and auditing. For example, how the courses are formed, who is responsible for creating and revising the syllabi, what are the auditing guidelines, and how often auditing is conducted. In Fig. 6, we present a workflow for course design. We also suggest the knowledge base necessary for automating the process. The design is represented in two parts: course syllabus design and articulation matrix.

In the proposed design, the main users are:

1) The curriculum committee is the group of people allocated by the college or department for designing and overseeing curriculum-related tasks.

2) Course coordinators\course designers are instructors allocated by the college or department for designing individual courses.

3) An accreditation committee is a group of people allocated by the college or department for conducting accreditation related tasks.
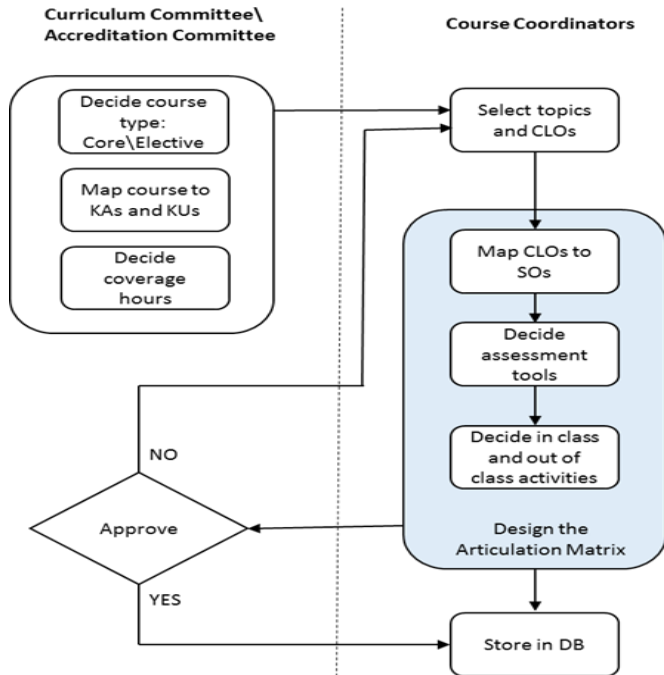
Fig. 6.   Course Design Workflow.

D. Course Syllabus Design

In the following subsections, the design of the course syllabus is discussed.

1) Data requirements

For each course, the system shall keep a record of its id, name, number, type (core\elective), credit hours, and the prerequisite course(s). For each course, the course is mapped to specific KAs, KUs, topics, and CLOs from the BoK.

2) DB design

The ER diagram for the course syllabus is provided in Fig. 7, and its corresponding mapping in relational notation is represented in Table II. All relations are normalized to third normal form.
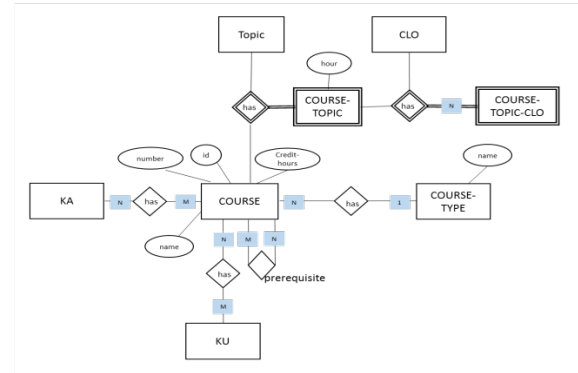
Fig. 7.   The ER Diagram for Course Syllabus.

TABLE II.   DESIGN OF THE COURSE SYLLABUS DB IN RELATIONAL NOTATION

Relation	description	Foreign Key	Referenced Relation
COURSE-TYPE(id, name)	An enumeration relation for course types "Core" or "Elective"		
COURSE(id, code, number, name, ct-id)	Master relation for storing course information, a universal system id, code, number, name, credit hours, and course type.	ct-id	COURSE-TYPE.id
COURSE-KA(c-id,ka-id)	Cross reference relation for storing the mapping of course to KAs	c-id	COURSE.id
		ka-id	KA.id
COURSE-KU(c-id,ku-id)	cross reference relation for storing the mapping of course to KUs.	c-id	COURSE.id
		ku-id	KU.id
COURSE-TOPIC(c-id, t-id,hour)	Weak relation for storing the mapping of course to topics and coverage hours for each topic	c-id	COURSE,id
		t-id	TOPIC.id
COURSE-TOPIC-CLO(c-id,t-id,clo-id)	Weak relation for storing the mapping of topics to CLOs which are specific for the designated course	(c-id,t-id)	COURSE-TOPIC(c-id,t-id)
		clo-id	CLO.id
COURSE-PREREQUISITE(c-id, c-pre-id)	cross reference relation for storing the prerequisite of courses	c-id, c-pre-id	COURSE.id

3) Design tips

Depending on the specific institutional and accreditation requirements, modifications can be applied to the design in Fig. 7. For example:

1) If the institution has a different way of calculating taught hours, relative attributes may be added to the relationship between COURSE and topics.

2) If course coordinators prefer to apply a maximum number of hour-coverage on the mapping between courses and their constituent KAs\KUs, relevant attributes may be added to the relationship between COURSE and KA and KU, respectively.

E. Articulation Matrix Design

A critical task of the accreditation process is to measure the performance of the courses and check if they achieve their predefined objectives. The starting point would be to use coursework design tools such as articulation matrices or course blueprints to map course outcomes to the different assessment tools (exams, quizzes, projects) in the course. In this paper, we refer to the articulation matrix, which is applied in the Accreditation Integration & Management System (AIMS) in the Faculty of Computing and Information Technology, King Abdulaziz University [8]. AIMS is a tool to support a sustainable multi-accreditation academic quality system (see Section VIII).

TABLE III.   A SIMPLIFIED ARTICULATION MATRIX. (AIMS)

NO	CLOs	ABET Student Outcomes	Assessment Tools									
			Quiz	Exam	Comprehensive Final Exam	Project (Individual)	Group Project	Homework Assignments	Graded Lab Work	Lab Exam	Student Work Portfolio	Formal Presentation
1	Apply concepts in ER modelling notation to model a real world problem	2		x								
2	Evaluate a proposed decomposition to determine the degree of its normal form.	1	x	x				x				

An articulation matrix is a tool used by course coordinators for structuring the detailed design of coursework. The structure includes elements of course design such as: assessment tools, in-class activities, and out of class activities. A simplified structure of the articulation matrix is illustrated in Table III. For linking course performance with ABET, CLOs are mapped to a set of 5 ABET Student Outcomes (SO) [2]. The mapping is used to compute the overall student attainment in each individual SO.

1) Data requirements

For each course, it is required to store the different assessment tools in the course and their percentage of marks out of 100%. For each CLO in a course syllabus, the CLO is associated with assessment tools such as exams, projects, etc. Each CLO and assessment tool pair is associated with the week it is assessed in, and the percentage of marks for the assessment. The same CLO can be evaluated in more than one assessment tool; hence different assumptions can be applied.

In this section, we only provide the requirements and DB design for the "assessment tools" in the matrix. However, a straight forward modification can be applied to add course activities or any other attributes as necessary.

2) DB design

The DB design of the articulation matrix is illustrated in Fig. 8, and its corresponding mapping in relational notation is represented in Table IV. All relations are normalized to third normal form.

F. Implementation Tips

The following implementation tips apply to the course design module (syllabus and articulation matrix):

1) Two different interfaces with relevant authorizations for both curriculum\accreditation committee members and course coordinators must be provided.

2) Features for editing course content.

3) Automation for the workflow in Fig. 6 is recommended.

4) Features for querying the knowledge base must be provided. The queries must cover the necessary statistics about the courses and curriculum, such as: number of courses covering a specific topic, the number of courses covering a specific SO, or the depth of coverage of specific SO.

5) Features for auto-generating course syllabi must be added.

It is essential to apply data entry features that automatically assist in preserving curriculum coherence. We provide here some examples.

1) To avoid redundancies in topics and CLOs, provide the course coordinator with up-to-date statistics about the courses already covering these topics and CLOs.

2) To avoid over coverage of topics, KAs, or KUs, provide course coordinators with up-to-date statistics

about the coverage hours of these attributes in the curriculum.

3) To avoid choosing topics that are not relevant to the course, design pull-down menus for selecting topics from the KUs and cross-referenced KUs specified for the course.

4) The use of database assertions and triggers will help maintain design policies and guidelines. For example: maintaining a percentage of "Familiarity" CLOs in a curriculum within a specific range or monitoring the depth of coverage at different granularity levels (KA, KU, or topic).
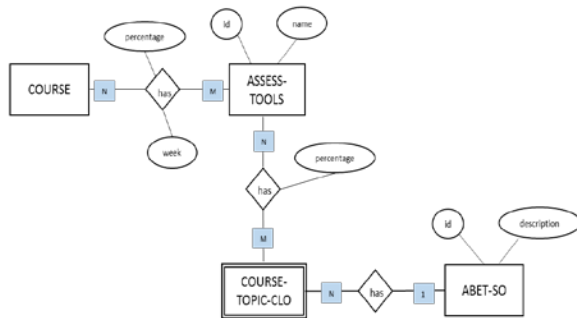
Fig. 8. The ER Diagram for the Articulation Matrix.

TABLE IV. DEISGN OF THE ARTICULATION MATRIX IN RELATIONAL NOTATION

Relation	Description	Foreign Key	Referenced relation
ABET-SO(id,description)	Enumeration relation for storing ABET SOs. Example:(1,"Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions")	–	–
ASSESSMENT-TOOLS(id,name)	Enumeration relation for storing different assessment tools. Example: (1,"Exam"),(2,"Quiz")	–	–
COURSE-ASSESSMENT(c-id,at-id,week,percentage)	Cross reference relation to store the mapping of courses to assessment tools with the percentage of each assessment out of 100	c-id	COURSE.id
		at-id	ASSESSMENT-TOOLS.id
COURSE-TOPIC-CLO-ASSESSMENT(c-id,t-id,clo-id,at-id,percentage)	Cross reference relation to store the percentage of each CLO assessment in a given assessment tool and the week its being conducted in	(c-id,t-id,clo-id)	COURSE-TOPIC-CLO(c-d,t-id,clo-id)
		at-id	ASSESSMENT-TOOLS.id
COURSE-TOPIC-CLO(c-id,t-id,clo-	Modify the relation in table 3 by mapping the	(c-id,t-id)	COURSE-TOPIC(c-id,t-id)
id,so-id)	CLO in each course to the corresponding ABET SO	clo-id	CLO.id
		so-id	ABET-SO.id

VIII. ACCREDITATION MANAGEMENT

Once the correct and complete design of courses is stored, different accreditation components can be added to the framework. These components will help accreditation committees in the evaluation process of their programs. These components are to be designed based on the evaluation and accreditation techniques deployed by colleges and institutions.

The proposed knowledge base in this paper will serve as a baseline for assessment and evaluation components that are necessary for the continuous improvement cycle. According to the additional add-on components, other data may be captured and stored. For example, after each course offering, it will be required to store student results in exams and course work and then used in assessment calculations. To avoid potential analysis paralysis problem, a careful selective approach must be followed to decide which data will be used -and consequently stored- for assessment, evaluation, and as evidence for accreditation.

Another category of tools can be used for curriculum analytics. For example, the Curriculum Analytics Tool (CAT) in [15] generates the competency scores for the entire curriculum across cognitive and progression levels.

The existence of a knowledge base will empower the curriculum and accreditation personnel with limitless capabilities, especially with data-driven decision making based on accurate and up-to-date statistics and information about the curriculum and its outcomes. This will allow the authorized personnel to apply necessary changes and improvements based on valid evidence.

A developed system of the proposed framework will significantly support the documentation needed for accreditation. We provide examples from the ABET SSR report in the following points:

- Criterion 3 (Student outcomes): in this criterion, the implemented system shall auto-generate reports of the curriculum to SO mappings, CLO to SO mappings, individual courses to SO mappings, and course syllabi.

- Criterion 4 (Continuous Improvement): in this section, the implementation of the improvement life cycle is documented. Reports about direct assessment attainment scores at SO, course, and program level should be auto-generated. Also, the data-driven decisions for curriculum tuning must be reported. Any tools developed in the framework must be documented as evidence.

- Criterion 5 (Curriculum): Different curriculum analysis reports should be auto-generated. For example, evidence of breadth and depth of advanced computing topics or depth of mastery levels can be auto-generated.

IX. RELATED SYSTEMS

Within academic institutions, different tools and systems have been developed to aid in the process of continuous improvement, accreditation, and curriculum-based analytics.

For example, AIMS provides an ABET-compliant web-based system for course articulation matrices and course assessment. The CLOs in the matrix are filled manually by course coordinators through a programmed Excel sheet. At the end of each semester, all student grades in selected assessment tools and selected outcomes are archived. The system then publishes statistics about student attainment in these outcomes. This will allow instructors to provide a detailed analysis of these outcomes and provide suggestions for improvements. The system also offers different accreditation reporting features. The absence of integration with BoK is a significant drawback in AIMS. The analysis that is based on the relationship of outcomes and curriculum content is done manually.

Curriculum analytical tools, on the other hand, are tools that adjust the alignment between program and course level competencies. As an example, a curriculum analytical tool was developed in the Department of Electrical & Computer Engineering University of New Mexico to improve graduation rates [17]. The tool models sequences of courses in the curriculum as directed graphs and allocate patterns where failing a course will delay students' graduation, then reform the curriculum accordingly.

As for accreditation support tools, expert comparison of commercially available tools such as EvalTools [9], CLOSO [10], and WEAVEonline [11] can be found in [20]. EvalTools provide a course management system, outcome-based assessment, and reporting features. CLOSO is an outcome-based assessment tool related to ABET accreditation. Both EvalTools and CLOSO do not support the continuous improvement life cycle. WEAVEonline provides an accreditation management system with partial support for the continuous improvement life cycle.

X. CONCLUSION

The accreditation and the continuous improvement of curriculums and are two different but parallel and interrelated processes. They both need 1) systemized and methodological approaches to achieve them, and 2) acquisition of comprehensive curriculum and accreditation related knowledge base. There is a need to propose analytical methods to measure the compatibility of existing curriculums against the underlying BoK. In this paper, we focus on the importance of integrating the BoK in the continuous improvement cycle and discuss its impacts on the coherence and consistency of curriculums. A general framework model for curriculum and accreditation management is proposed. The database design of the BoK repository, the curriculum, and the workflow for designing courses are provided. We also discuss the means of linking the proposed design with accreditation tasks.

ACKNOWLEDGMENT

REFERENCES

[1] Accreditation Council for Business Schools and Programs (ACBSP) Available: https://acbsp.org/

[2] Accreditation Board for Engineering and Technology (ABET). Available: https://www.abet.org/

[3] ACM\IEEE (CS 2013) Report, Available: https://www.acm.org/binaries/content/assets/education/cs2013_web_final.pdf or http://ai.stanford.edu/users/sahami/CS2013/.

[4] ABET Criteria. Available: https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-computing-programs-2019-2020/#GC3

[5] Common Body of Knowledge (CBK) Available: https://www.isc2.org/Certifications/CBK

[6] Business Analysis Body of Knowledge (BABOK) Available: https://www.iiba.org/standards-and-resources/babok/

[7] Landscape Architecture Body of Knowledge Study Report (LABOK), Available: https://www.asla.org/uploadedFiles/CMS/Education/Accreditation/LABOK_Report_with_Appendices.pdf.

[8] Accreditation Integration & Management System (AIMS), Available: https://fcit.kau.edu.sa/aims/home.php.

[9] EvalTools. Available: https://iu.myschooling.net/EvalTools6/EvalTools/

[10] CLOSO. Available: http://www.smart-accredit.com/

[11] WEAVEonline. Available: https://app.weaveonline.com/login.aspx

[12] A. S. Barb and N. Kilicay-Ergib, "Applications of Natural Language Techniques to Enhance Curricular Coherence, " Complex Adaptive Systems Conference with Theme: Leveraging AI and Machine Learning for Societal Challenges, vol. 168, pp.88-96, 2020.

[13] D. Bateman, S. Taylor, E. Janik, and A. Logan, "Curriculum coherence and student success," Champlain College CEGEP, vol. 22, no. 5, pp. 8-18, Dec 2007.

[14] R. M. Felder and R. Brent, "Designing and teaching courses to satisfy the ABET engineering criteria," Journal of Engineering Education, vol. 92, no. 1, pp. 7-25, Jan 2003.

[15] S. Gottipati and V. Shankararaman, "Competency analytics tool: Analyzing curriculum using course competencies," Education and Information Technologies, vol. 23, no. 1, pp. 41-60, Jan 2018.

[16] T. Hatzakis, M. Lycett, and A. Serrano, "A programme management approach for ensuring curriculum coherence in IS (higher) education," European Journal of Information Systems, vol. 16, no. 5, pp. 643-657, 2007.

[17] G. L. Heileman, M. Hickman, A. Slim, and C. Abdallah, "Characterizing the complexity of curricular patterns in engineering programs," in ASEE Annual Conference & Exposition, Columbus, 2017.

[18] G. Mayo, W. Wu, T. McCuen, R. R. Issa, and D. Smith, "Implementation of the BIM body of knowledge (BOK) framework for program planning in academia," in Proceedings of the 12th BIM Academic Symposium & Job Task Analysis Review, pp. 2-9, Mar 2018.

[19] K. Mohiudddin, A. Islam, S. Mohd, and M. Shariff, "Evaluation of an Academic Program: The Case of Computing Accreditation Commission Framework in Higher Education," International Journal of Emerging Technologies in Learning (iJET), vol. 14, no. 11, pp. 70-91, 2019.

[20] A. Namoun, A. Taleb, and M. Benaida, "An Expert Comparison of Accreditation Support Tools for the Undergraduate Computing Programs," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 9, no. 9, Aug 2018.

[21] K. E. Willcox and L. Huang, "Mapping the CDIO curriculum with network models," CDIO, 13th International CDIO Conference, pp. 78–88, June 2017.

[22] P. A. Quezada-Sarmiento et al., "Body of Knowledge Model and Linked Data Applied in Development of Higher Education Curriculum," in Science and Information Conference, vol. 943, pp. 758-773, April 2019.

[23] E. Rata, "Knowledge-rich teaching: A model of curriculum design coherence," British Educational Research Journal, vol. 45, no. 4, pp. 681-697, May 2019.

[24] W. H. Schmidt and R. S. Prawat, "Curriculum coherence and national control of education: issue or non-issue?," Journal of curriculum studies, vol. 38, no. 6, pp. 641-658, Dec 2006.

[25] W. Y. Wong and M. Lavrencic, "Using a risk management approach in analytics for curriculum and program quality improvement," 6th International Conference on Learning Analytics & Knowledge, , 1st Learning Analytics for Curriculum and Program Quality Improvement Workshop Edinburgh, APR 2016.

APPENDIX

Fig. 9.   Browsing the Body of Knowledge.

Fig. 10. Storing the Body of Knowledge.

Fig. 11. Designing a Course from the Body of Knowledge.

A Hybrid Deep Learning Model for Arabic Text Recognition

Mohammad Fasha[1]

Amman, Jordan

Bassam Hammo[2], Nadim Obeid[3]

King Abdullah II School for Information Technology
The University of Jordan
Amman, Jordan

Jabir AlWidian[4]

Princess Sumayah School for Technology
Amman, Jordan

Abstract—**Arabic text recognition is a challenging task because of the cursive nature of Arabic writing system, its joint writing scheme, the large number of ligatures and many other challenges. Deep Learning (DL) models achieved significant progress in numerous domains including computer vision and sequence modelling. This paper presents a model that can recognize Arabic text that was printed using multiple font types including fonts that mimic Arabic handwritten scripts. The proposed model employs a hybrid DL network that can recognize Arabic printed text without the need for character segmentation. The model was tested on a custom dataset comprised of over two million word samples that were generated using (18) different Arabic font types. The objective of the testing process was to assess the model's capability in recognizing a diverse set of Arabic fonts representing a varied cursive styles. The model achieved good results in recognizing characters and words and it also achieved promising results in recognizing characters when it was tested on unseen data. The prepared model, the custom datasets and the toolkit for generating similar datasets are made publically available, these tools can be used to prepare models for recognizing other font types as well as to further extend and enhance the performance of the proposed model.**

Keywords—*Arabic optical character recognition; deep learning; convolutional neural networks; recurrent neural networks*

I. INTRODUCTION

Optical Character Recognition (OCR) is the process of recognizing text in images and transforming it into a machine encoded text. OCR is an important research area and generally, it can be classified into two main groups, online OCR and offline OCR. Online OCR involves recognizing text while typing in real time such as recognizing digital stylus writing on mobile phones, while offline OCR involves the recognition of text in document images such as scanned documents archives, printed application forms, bank cheques, postal mail and many others. In addition, OCR addresses two main categories of text; machine printed text and handwritten text and each of these two areas has its own challenges. Printed text is faced with the challenge of the diverse font types and the various formatting styles as well as the quality of the printed and the scanned images, while handwritten text is considered more challenging because of the diverse writing styles of individuals.

Recognizing Arabic text in images has additional challenges that are mainly caused by the cursive nature of Arabic script. In addition, Arabic characters are connected in words, and the writing system has a large number of ligatures, which increase the challenge of recognizing text based on characters' boundaries. Further, the scarcity of labelled datasets for Arabic language increases the challenges of developing new solutions that depend on supervised learning models.

This work presents a model that employs a hybrid DL network to recognize multiple Arabic fonts types including fonts that mimic Arabic handwritten scripts. The hybrid DL model uses a Convolution Neural Network (CNN) and a Bi-Directional Short Long Term Memory network (BDLSTM) and it operates in an end-to-end fashion without the need for character segmentation.

To test the performance of the model, a number of datasets made of (18) different fonts types were compiled. The fonts of the datasets were collected from online sources and they were selected because they exhibit high cursive nature that mimics Arabic handwriting styles. The sample words for generating the custom datasets were extracted from Arabic Wikipedia Dump and they comprise over two million words samples.

Several experiments were performed to examine the model's performance and to assess its generalization capabilities. Despite being a moderate model in terms of its complexity (i.e. can be trained on a single CPU), the same single model was able to achieved (98.76%) in Character Recognition Rate (CRR) and (90.22%) in Words Recognition Rate (WRR) for all the tested font types. However, the model demonstrated degradation in its performance when it was tested on an unseen dataset or noisy images. In the case of unseen dataset, it achieved a CRR success rate of (85.15%), while in the case of noisy images it achieved a CRR success rate of (77.29%). However, resolving these issues require additional investigations, which are out of the scope of this work.

The remaining of the paper is organized as follows: In Section II the challenges related to Arabic text recognition are discussed. In Section III an overview of the related work is presented. Section IV presents the compiled dataset. The proposed model and its process flow is presented in Section V. Section VI presents the experiments and the obtained results and finally, Section VII concludes the paper and identify some future research avenues.

II. CHALLENGES RELATED TO ARABIC TEXT RECOGNITION

Arabic writing system is used by different nations around the world, this includes the (21) Arab countries as well as other nations such as Kurdish, Pashto, Persian, Sindhi, Uighur and Urdu. The base writing system is made of the base alphabets of Arabic language, which consist of (28) alphabets and (10) Hindi format numerals. The alphabets are written from right-to-left while the numbers are written from left-to-right. On the other hand, the shape of Arabic alphabets can change according to their position in the word. Fig. 1 below shows the based alphabets of Arabic language as well as their variations according to their position in the word.

A main challenge of Arabic writing system is related to its cursive nature knowing that alphabets are written in a joint flowing style. In this respect, characters in Arabic script, whether handwritten, typed or printed are connected within words and they might overlap within the same word or across words (i.e. inter and intra overlapping). In addition, spaces can occur within words and across them while various Arabic characters share the same main shape (e.g. ba, ta, tha as shown in Fig. 1). However, these characters are distinguished by the number of dots added under or above the base alphabet, which increase the challenge of identifying the correct alphabet.

Further, the shapes of Arabic characters are represented using different glyphs according to the characters' position in the word. Accordingly, different shapes are used when the character appears at the beginning, middle or at the end of a word.

pronounce	end	mid	begin	isolated	pronounce	end	mid	begin	isolated
ḍād	ض	ض	ض	ض	ʾalif	ـا			ا
ṭāʾ	ط	ط	ط	ط	bāʾ	ـب	ـبـ	بـ	ب
ẓāʾ	ظ	ظ	ظ	ظ	tāʾ	ـت	ـتـ	تـ	ت
ʿayn	ع	ع	ع	ع	thāʾ	ـث	ـثـ	ثـ	ث
ghayn	غ	غ	غ	غ	jīm	ـج	ـجـ	جـ	ج
fāʾ	ـف	ـفـ	فـ	ف	ḥāʾ	ـح	ـحـ	حـ	ح
qāf	ـق	ـقـ	قـ	ق	khāʾ	ـخ	ـخـ	خـ	خ
kāf	ـك	ـكـ	كـ	ك	dāl	ـد			د
lām	ـل	ـلـ	لـ	ل	dhāl	ـذ			ذ
mīm	ـم	ـمـ	مـ	م	rāʾ	ـر			ر
nūn	ـن	ـنـ	نـ	ن	zāy / zayn	ـز			ز
hāʾ	ـه	ـهـ	هـ	ه	sīn	ـس	ـسـ	سـ	س
wāw	ـو	و			shīn	ـش	ـشـ	شـ	ش
yāʾ	ـي	ـيـ	يـ	ي	ṣād	ـص	ـصـ	صـ	ص

Fig 1. Arabic base Alphabet.

Similarly, the shape of Arabic characters in printed text varies depending on the used font, as well as its formatting and printing style. Additionally, natural languages that use the Arabic writing system extends the base alphabets by adding special diacritics over some characters to better adapt the writing system to the phonemes of the designated language. A thorough discussing about these challenges can be found in [1]. All these characteristics make the recognition of Arabic text a challenging task, especially for the models that depend on segmenting characters prior to the recognition process [2].

The next section presents some of the related work that was introduced to address some of these challenges and the approaches that were followed.

III. RELATED WORK

The recognition of Arabic text is still a challenging task because of many intricate features related to the nature of Arabic writing system [3]. Work in this domain is an active research area where many models are continuously proposed for the problem of automatically recognizing printed or handwritten text. Each of these domains has its own challenges and requirements. The challenges of recognizing printed Arabic text are driven by the need for a ubiquitous model that can efficiently recognize Arabic text that is printed using multiple font types and using different formatting styles. On the other hand, the challenges facing the Arabic handwritten text are driven by its high variety due to the diversity of individuals writing styles. In this section, we present an overview about the related work in both domains and the methods that were employed to recognize text.

A recent model for recognizing printed Arabic characters in isolation mode was presented in [4] which applied K-Nearest Neighbor (KNN) and Random Forest Tree (RFT) algorithms to recognize Arabic text. That model used statistical methods to extract features from text images. These features included the dimensions of the text shape, the transition of pixels, the number of black vs white pixels and regional ratios of pixel values. The KNN classifier achieved a successful rate of (98%), while the RFT classifier achieved (87%). Similarly, the authors in [3] introduced a model for recognizing Arabic printed text using linear and nonlinear regression. In that work, text in images were initially thinned and segmented into sub words. Next, the relations between word segments were represented using a numerical coding scheme that represented characters as a sequence of points, lines, ellipses and curves. Using that scheme, a unique code was established for each character form and a unique list of codes were used to recognize each font type. Finally, linear regression technique was used to validate the representations against a ground truth table using distance measures. The model was evaluated using (14000) words samples and it has achieved an accuracy rate of (86%).

In [5], the authors proposed a model for segmenting Arabic printed text that can serve as a preliminary step in the text recognition process. The model that was presented in that work applied contours analysis and template matching techniques to recognize text. The contour segmentation was determined by the local minima values of the contour and the template-based technique involved scanning the positions of black pixels after

segmenting the text into lines and sub words using horizontal and vertical projections. The model was evaluated using a custom multi-font corpus and was also benchmarked against five other methods. The model achieved an enhanced accuracy over the other models with a score of (94.74%).

Arabic text recognition research was also influenced by the progress that was achieved in deep learning technology. Earliest work in implementing DL approaches to Arabic text recognition can be traced back to [6]. In that work, a Multi-Dimensional Long Short Memory (MDLSTM) network and Connectionist Temporal Loss (CTC) were used to recognize Arabic text in images. The model was tested on the IFN/ENIT dataset of Tunisian handwritten Town names [7] and an accuracy levels of (91.4%) were reported.

A more recent work in the field of Arabic text recognition using DL models was carried out by [8]. The domain of that work was the recognition of Arabic script in historical Islamic manuscripts. The presented model used various preprocessing techniques to enhance the quality of the scanned images and to segment the text prior to the recognition process. CNNs were used to recognize the preprocessed text and accuracy levels ranging from (74.29%) to (88.20%) were reported.

In [9], a model for recognizing Arabic handwritten text using neural networks was presented. Initially, the noise in images was reduced using multiple image preprocessing techniques. The characters in words were segmented into regions using a threshold-based method and these regions were used to construct feature vectors. The model was examined on a custom dataset collected from volunteered writers and a CRR of (83%) was reported. Similarly, the work of [10] presented a three-layers CNN model for recognizing Arabic handwritten characters in isolation mode. The model was examined on AIA9k [11] and AHCD [12] datasets and CRR of (97.6%) was reported. In [2], the authors presented a DL based model for recognizing Arabic handwritten text using a MDLSTM network and CTC loss function. The objective of that work was to assess the effects of extending the dataset using data augmentation techniques and to compare performance of the extended model against other similar models. The KHAT handwritten dataset [13] was used to train and evaluate the model and a CRR level of (80.02%) was reported. Finally, in [14] a hybrid DL model for detecting printed Urdu text in scanned documents was discussed. The model employed a hybrid combination of CNN and BDLSTM along with CTC loss and it was tested on URDU and APTI datasets [15]. The model was able to achieve CRR rates of (89.84%) and (98.80%), respectively.

Reviewing the related work revealed that there is a shortage in work that addresses Arabic printed text using DL models. Our work should present some footsteps in this research area and provide toolkits that can be utilized to further extend and enhance the achieved outcomes.

IV. THE COMPILED DATASET

During the last period, several Arabic printed datasets were introduced by the community including: DARPA, APTI , PATDB, APTID/MF, and RCATSS [5], [15]–[18]. Nevertheless, there is no consensus on a standard dataset that

can adopted by the community that can be used for benchmarking printed text recognition. Consequently, the available datasets vary in their content types, sizes, formatting styles and fonts types [19]. A more thorough listing of similar datasets can be found in [20].

As stated earlier, the main objective of this work was to examine the performance of DL based models in recognizing Arabic text that was printed using fonts that mimic Arabic hand writing styles. For that purpose, no suitable dataset was found and consequently a number of custom datasets were compiled to serve the purpose i.e. Arabic Multi-Fonts Dataset (AMFDS). These dataset were prepared using the (18) fonts depicted in Fig. 2 below.

In this respect, a custom toolkit for generating the datasets was prepared. This toolkit can be used to generate any number of text image samples using any required font type. It can also be configured to generate samples as separate image files or as a single binary repository for all the samples.

Table I next shows the main characteristics of each generated dataset. As presented in the table, the (ae-Nice) font was selected to generate the single-font datasets. This font type was selected because its printing style clearly exhibits cursive structures that mimic Arabic hand writing script. Similarly, the (ae-Nice) and the (K-Karman) font's types were selected to generate the two-font's datasets. Finally, datasets (4) and (5) in Table I were generated using the font types that are presented in Fig. 2.

Datasets (1, 2, 3) in Table I include duplicate samples because the same set of words was used to generate samples for each font type. In addition, these datasets have minor redundancies within the samples of each font because words were randomly sampled from Arabic Wikipedia Dump and no filtering was applied. Datasets (4 and 5) in Table I are unique (disjoint) datasets where no single word is replicated across the entire dataset. The current version of the datasets contains words samples that have a length of (7 to 10) characters and all the words were generated using font size (26) and bold formatting style.

Font Type	Font sample
ae_Nice	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
Aref_Merma	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
bader_lamaat	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
Baam Marah Beh Light	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
B Cheshmeh	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
B Pastch	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
BSetareh	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
B Ziba	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
Dast Nevis	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
Duma Font	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
DecoType Thuluth	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
Ghalam-1	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
Ghalam-2	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
K Kamran	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
K Tabassom	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
pHalls Khodkar	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
Tharwat Emara Modern	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.
ToyorAljanah	التعرف على النصوص العربية باستخدام تقنيات التعلم الآلي.

Fig 2.   The Set of the Selected Fonts.

TABLE I.   THE PREPARED DATASETS

Dataset #	Number of fonts	Number of samples	Duplicates in samples	Dataset size	Fonts
1	1	60,000	Exist	148MB	ae-Nice
2	1	120,000	Exist	295MB	ae-Nice
3	2	240,000	Exist	490MB	ae-Nice, K-Karman
4	18	2,160,000	Exist	6200 MB	The fonts shown in Fig. 2
5	18	450,000	Unique words	1300MB	The fonts shown in Fig. 2

Further, each dataset is comprised of two main data files: a labels file and binary file. The labels file is a normal text file that contains details about the word samples, this includes the Arabic word represented by the image, the font type, the font style, the font size and a value that represent the starting index of that image in the binary file. Hence, the byte stream of the designated image begins at the starting index and spans to a length equals to the image's size (in bytes). This binary file represents a single repository for all the images in the dataset. Unlike the common adopted approaches of using single image file for each word sample, the format presented in this work is more appropriate for addressing large data files with large number of samples and it is more scalable as it facilitates moving datasets around different execution environments i.e. cloud based environments, it also facilitates the processing of image data in terms of loading, preprocessing and training.

The datasets that were used in the experiments are made publically available at [1], similarly, the toolkit that can be used to generate different samples is made publically available at [2].

V.   PROPOSED MODEL

In general, text recognition systems implement a series of tasks before recognizing text in images. These tasks can be classified into five main categories which includes: the normalizing of document images to enhance their quality, the detection of text regions within a document and segmenting text accordingly, the extraction of useful features from text, implementing the recognition process and employing post processing techniques to enhance the accuracy of the achieved results. The focus of this work is on the recognition process; while the other tasks are out of the current scope and might be addressed in future research.

The design and the implementation of the proposed model was based on the work presented in [21]. In that work, a hybrid NN for recognizing handwritten text in scanned historical documents was presented. The model presented in this work employs the same intuition and adapts the model to recognize different styles of Arabic printed text.

The proposed model is comprised of two main components; a Convolutional NN and a Recurrent NN. These networks are stacked together in an end-to-end fashion that can perform word-level recognition without the need for character level segmentation.

1https://drive.google.com/drive/folders/1mRefmN4Yzy60Uh7z3B6cllyy OXaxQrgg?usp=sharing

2 https://github.com/JTCodeStore/TextIimagesToolkit

TABLE II.   MODEL DESIGN

CNN				
Layer	Filter size	# of filters	Pooling window	Output size
1	(5, 5)	32	(2, 2)	(64,16,32)
2	(5, 5)	64	(2, 2)	(32,8,64)
3	(3, 3)	128	(1, 2)	(32,4,128)
4	(3, 3)	128	(1, 2)	(32,2,128)
5	(3, 3)	256	(1, 2)	(32,1,256)
BDLSTM				
Layer	# of hidden units			
1	256 x 2 (forward and backward)			
2	256 x 2 (forward and backward)			

Table II presents the specifications of the proposed model.

As shown in the previous figure, the filter sizes in the initial two layers of the CNN employs filters of size (5, 5) units. This filter size is suitable for extending the receptive field of the early layers of the network. The three remaining convolution layers in the network employed filters sizes of (3, 3).

Further, the convolution process in the model employed zero padding so that it can preserve the size of the input image throughout the convolution process. The pooling process in the initial two layers used a sliding window of size (2x2) while the remaining three layers used a window of size (1, 2).

The CNN is stacked on top of a BDLSTM in an end-to-end manner. The BDLSTM had (2) LSTM layers, each layer had two LSTM cells that implements the forward and backward passes of inputs in the network, and each LSTM cell had (256) hidden units. Fig. 3 next shows the general architecture of the model.

The processing of the model starts with the CNN accepting input images of size (128×32). Therefore, prior to injecting the images into the model, these images were resized to a size of (128×32) units. The resizing process changes the shape of the input images by compressing it and shifting the location of the text. However, CNNs are shift invariant and they are tolerable to such variances. Fig. 4 below shows the effects of applying the resizing process on a sample word.
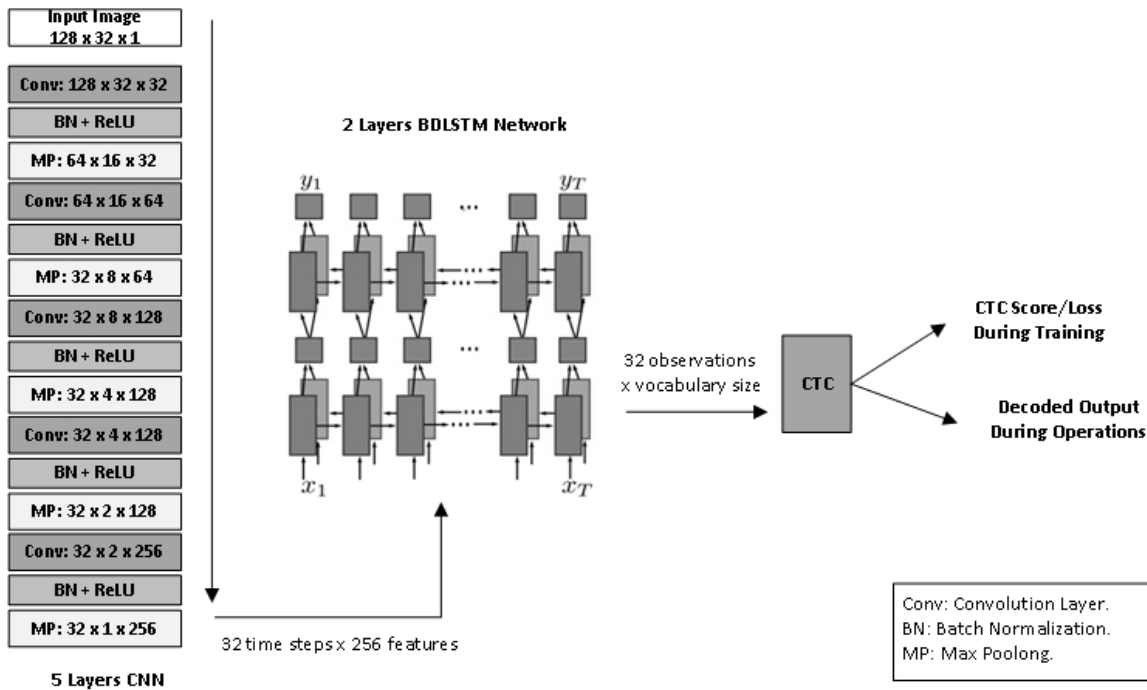
Fig 3.   General Architecture of the Proposed Model, which is Comprised of (5) Layers CNN, (2) Layers BDLSTM and CTC Loss Function.

Fig 4.   Input Images before and after Resizing.

Once an input image is received by the model, the CNN implements four main operations: Convolution, Batch Normalization, Activation and Max Pooling.

The convolution process extracts the relevant features from the input image through the filters and passes the values to a batch normalization process to mitigate the effects of covariance shift [22]. Next, ReLU activation function is implemented to eliminate negative values and to minimize the effects of vanishing gradients. The results of the activation function are passed to a max-pooling layer which performs sub sampling by selecting the most relevant features and gradually downsizes the input size into an array of (32 time-steps×256 features). Fig. 8(a-f) in Appendix (A) show the output of the convolution layers for the sample Arabic word (al rasmeya "الرسمية"). These figures demonstrate how the earlier layers in the CNN captures detailed features while the later layers capture more generalized features. In addition, Fig. 8(f) in the appendix shows the output of the final layer of the CNN which demonstrate how the features are concentrated in the lower

indexes (0-10 of 32) of the output array as the words length in the experiments ranged between (7-10) characters.

The final output of the CNN is passed to the BDLSTM network, which learns the sequence or the temporal dimension in the input image. The output of the BDLSTM network is an array of size (32-observations×vocabulary size) knowing that a single character in the input text might be represented by one or more observation sequence.

Sample outputs of the BDLSTM network for Arabic word (al rasmeya "الرسمية") are presented in Fig. 8(h-j) in Appendix (A). The figures demonstrate how the values for a specific output sequence increased at locations that corresponds to its index in the vocabulary that shown in Fig. 5. This vocabulary represents the unique characters that are represented in the custom dataset.

Next, the BDLSTM observations are passed to a CTC loss function, which performs a probabilistic based mapping between the observations and ground-truth labels. According the model's specifications, the CTC function is able to recognize words of size (32) characters while each time step sequence can represent one of the (38) different characters that are shown in Fig. 5.

Finally, the CTC function computes the loss value and back propagates it to the network to initiate the end-to-end learning process. RMSOptimizer [23] was used to implement the optimization process in the proposed model.
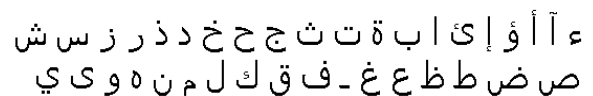
Fig 5.   List of Unique Characters in the Compiled Dataset.

VI. EXPERIMENTS AND RESULTS

To train and examine the proposed model, a TensorFlow based implementation was prepared using Python, this implementation is made publicly available at [3].

The model was trained and tested using Google Colab platform. That platform provides computing environments that includes hardware acceleration that can be used to train different DL models. The environment that were used for implementing the model in this work had the following specifications:

- Python version: 3.6.
- Tensor flow version: 2.0.
- Hardware Accelerator: 12GB NVIDIA Tesla K80 GPU.
- RAM: 12.72 GB.
- HDD: 64.4 GB.

To evaluate the model, CRR and Words Recognition Rate (WRR) measures were employed, the formulas of these measures are shown below:

$$CRR = \frac{\sum \textbf{Levenshtein EditDistance}(\textbf{Recognized Text}, \textbf{Ground Truth})}{\textbf{Total Number of Processed Characters}}$$

$$WRR = \frac{\textbf{Number of correctly identified words}}{\textbf{Total number of words}}$$

The model was examined through a number of testing scenarios using the five datasets that were presented in Table I. For this purpose, the datasets were split into a (train, validate, and test) segments using a ratio of (80%, 10%, and 10%) respectively. Several testing scenarios were implemented to examine the model's performance and the results of these testing scenarios are shown in Table III below.

Initially, the model was tested using a relatively small single-font dataset (i.e. dataset #1 in Table I). The results of this test were (97.5%) for CRR and (85.18%) for WRR. Next, the model was tested on a larger dataset with the same font setting (dataset #2 in Table I) and the CRR success rate enhanced to (99.04%) while the WRR achieved (94.29%). In general, the model achieved good results when it was examined on a single-font type dataset regardless of the type or the formatting style of the tested font.

To examine the model's performance on a more diverse dataset, it was tested using a two-fonts dataset (dataset #3 in Table I). Using this dataset set, the model achieved a CRR rate of (99.88%) and a WRR rate of (99.2%).

The high success rates that were achieved in the previous testes were further challenged by testing the model using an extended dataset that is comprised of two million word samples (i.e. dataset #4 in Table I). In this testing scenario, the model achieved good results in CRR (99.27%) but it demonstrated a minor degradation in WRR (94.32%). This behavior can be justified by emphasizing that the WRR is highly dependent on the accuracy of the CRR, and a minor flaw in the recognition of a single character shall affect the recognition of all the related words, especially that the dataset samples included relatively long words (a length of 7–10 characters).

To evaluate the model more accurately in terms of overfitting, it was trained on a new dataset that has no duplication (dataset #5 in Table I). Using this disjoint dataset, the performance of the model demonstrated a minor degradation in the reported accuracy, but it still achieved good results which were (98.76%) for CRR and (90.22%) for WRR.

To test the model's generalization capabilities, an experiment was performed using a new dataset with words length of (5-6) characters. Although the model was not trained on this length of words, it was able to achieve a CRR success rate of (98.71%) and a WRR success rate of (92.4%). Obviously, this was an indication that the model can be generalized beyond the samples that were used in the training process.

In the same aspect, a pilot testing for the model was implemented on the external APTI database [15]. In this experiment, an Arabic typesetting font of size (10) was selected. Although the model was not trained or fine-tuned on this database, it was able to achieve a CRR success rate of (85.15%), while the WRR success rate was degraded to (23.7%). Again, this performance degradation can be justified by the high correlation between the WRR and the accuracy levels of CRR.

Finally, the behavior of the model was tested on some noisy datasets. For this purpose, a selected set of test samples were induced with salt and pepper (S&P) noise and Speckle noise. Fig. 6 below shows a sample word before and after noise transformations.

For this last experiment, the model achieved an acceptable CRR success rate of (82.01%) for the induced S&P noise, while it achieved (77.29%) for both S&P and Speckle noises. However, for the WRR success rates, the model reported (21.48%) for S&P alone and (14.18%) for both noises; S&P and Speckle. Fig. 7 below summarizes the model's performance in all conducted experiments.

The conducted experiments show that the model was able to achieve high accuracy results when it was trained on a specific set of fonts. This might be an indication that the number of the supported fonts could be extended using the same intuitions in a more complex model i.e. larger number of parameters. In addition, several techniques can be incorporated to enhance the WRR accuracy level that has degraded under some testing scenarios. This might include employing post-processing techniques such as language models in order to enhance the overall accuracy of the WRR. In addition, various image-preprocessing techniques can be examined to mitigate the effects of noisy environments.

---

[3] https://github.com/JTCodeStore/ArabicDLOCR

TABLE III.   PERFORMANCE RESULTS OF THE MODE

#	Experiment Name	Dataset	Validation Accuracy %		Test Accuracy %	
			CRR	WRR	CRR	WRR
1	Single font model	1	98.37	90.28	97.5	85.18
2	Single font model – larger dataset	2	99.148	94.66	99.044	94.29
3	Two fonts model	3	99.93	99.5	99.88	99.2
4	(18) fonts, duplicate words across fonts types	4	99.38	94.84	99.27	94.32
5	(18) fonts, unique words across the dataset	5	98.81	90.53	98.76	90.22
6	Testing model generated in expriment 5 above on five character words	-	-	-	98.71	92.4
7	Testing model generated in expriment 5 above on APTI dataset – new font	-	-	-	85.15	23.7
8	Testing model generated in expriment 5 above with salt & pepper noise.	-	-	-	82.01	21.48
9	Testing model generated in expriment 5 above with salt & pepper and speckle noise.	-	-	-	77.29	14.18

Original image   Salt and Pepper noise   Salt and Pepper + Speckle noises

Fig 6.   Sample Images with Noise Adjustments.

Fig 7.   Model Performance for All Conducted Experiments.

VII. CONCLUSION

In this work, a hybrid DL model for recognizing Arabic text in images was presented. The objective of this work was to examine the models competency in recognizing Arabic text that was printed using multiple font types, including fonts that exhibit high cursive nature that mimic Arabic handwriting script. The proposed model demonstrated good CRR in most testing scenarios including the testing on a disjoint dataset, the testing on a pilot external database and the testing under noisy environments. The overall performance of the model is open for more enhancements through incorporating language models to enhance the overall WRR accuracy as well as using image processing techniques to mitigate the effects of noise in images. The same model can also be examined in recognizing Arabic handwritten text. Such measures might be investigated in a future work.

REFERENCES

[1] M. Alghamdi and W. Teahan, "Printed Arabic script recognition: A survey," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 9, pp. 415–428, 2018, doi: 10.14569/ijacsa.2018.090953.

[2] A. D. Riaz Ahmad, Saeeda Naz, Muhammad Afzal, Sheikh Rashid, Marcus Liwicki, "A Deep Learning based Arabic Script Recognition System," Int. Arab J. Inf. Technol., no. October, 2017.

[3] A. A. Shahin, "Printed Arabic Text Recognition using Linear and Nonlinear Regression," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 1, pp. 227–235, 2017, doi: 10.14569/ijacsa.2017.080129.

[4] A. E. Hassanien, M. F. Tolba, and A. T. Azar, "Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifier," Commun. Comput. Inf. Sci., vol. 488, no. November, pp. 10–17, 2014, doi: 10.1007/978-3-319-13461-1.

[5] A. Zoizou, A. Zarghili, and I. Chaker, "A new hybrid method for Arabic multi-font text segmentation, and a reference corpus construction," J. King Saud Univ. - Comput. Inf. Sci., 2018, doi: 10.1016/j.jksuci.2018.07.003.

[6] A. Graves, "Offline Arabic Handwriting Recognition with Multidimensional Recurrent Neural with Multidimensional Recurrent Neural Networks," Adv. Neural Inf. Process. Syst., no. January 2008, 2008, doi: 10.1007/978-1-4471-4072-6.

[7] M. Pechwitz, S. Snoussi, and N. Ellouze, "IFN / ENIT-database of handwritten Arabic words," no. May 2014, 2002.

[8] H. A. Bodour Alreha,Najla Alsaedi, "Historical Arabic Manuscripts Text Recognition Using Convolutional Neural Network," Conf. Data Sci. Mach. Learn. Appl. Hist., 2020.

[9] A. Mohsin and M. Sadoon, "Developing an Arabic Handwritten Recognition System by Means of Artificial Neural Network," J. Eng. Appl. Sci., vol. 15, no. 1, pp. 1–3, 2019, doi: 10.36478/jeasci.2020.1.3.

[10] K. Younis and A. Khateeb, "Arabic Hand-Written Character Recognition Based on Deep Convolutional Neural Networks," Jordanian J. Comput. Inf. Technol., vol. 3, no. 3, p. 186, 2017, doi: 10.5455/jjcit.71-1498142206.

[11] M. Torki, M. E. Hussein, A. Elsallamy, M. Fayyaz, and S. Yaser, Window-Based Descriptors for Arabic Handwritten Alphabet Recognition: A Comparative Study on a Novel Dataset. 2014.

[12] H. E.-B. Ahmed El-Sawy, Mohamed Loey, "Arabic Handwritten Characters Recognition using Convolutional Neural Network," WSEAS Trans. Comput. Res., no. 5, pp. 11–19, 2017, doi: 10.1109/IACS.2019.8809122.

[13] S. A. Mahmoud et al., "KHATT: Arabic offline Handwritten Text Database," Proc. - Int. Work. Front. Handwrit. Recognition, IWFHR, pp. 449–454, 2012, doi: 10.1109/ICFHR.2012.224.

[14] M. Jain, M. Mathew, and C. V. Jawahar, "Unconstrained OCR for Urdu using deep CNN-RNN hybrid networks," Proc. - 4th Asian Conf. Pattern

Recognition, ACPR 2017, pp. 753–758, 2018, doi: 10.1109/ACPR.2017.5.

[15] F. Slimane, S. Kanoun, H. El Abed, A. M. Alimi, R. Ingold, and J. Hennebert, "ICDAR2013 competition on multi-font and multi-size digitally represented arabic text," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, no. August, pp. 1433–1437, 2013, doi: 10.1109/ICDAR.2013.289.

[16] F. K. Jaiem, S. Kanoun, M. Khemakhem, H. El Abed, and J. Kardoun, "Database for Arabic printed text recognition research," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8156 LNCS, no. PART 1, pp. 251–259, 2013, doi: 10.1007/978-3-642-41181-6_26.

[17] A. G. Al-Hashim and S. A. Mahmoud, "Benchmark database and GUI environment for printed arabic text recognition research," WSEAS Trans. Inf. Sci. Appl., vol. 7, no. 4, pp. 587–597, 2010.

[18] R. Davidson, R., Hopely, "Arabic and Persian Training and Test Data Sets," 1997.

[19] L. Alhomed and K. Jambi, "A Survey on the Existing Arabic Optical Character Recognition and Future Trends," Int. J. Adv. Res. Comput. Commun. Eng., vol. 7, no. 3, pp. 78–88, 2018, doi: 10.17148/IJARCCE.2018.7213.

[20] I. Ahmed, S. A. Mahmoud, and M. T. Parvez, "Guide to OCR for Arabic Scripts," Guid. to OCR Arab. Scripts, no. November 2017, 2012, doi: 10.1007/978-1-4471-4072-6.

[21] H. Scheidl, "Handwritten Text Recognition in Historical Documents," 2011.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 32nd Int. Conf. Mach. Learn. ICML 2015, vol. 1, pp. 448–456, 2015.

[23] G. Hinton, "Neural Networks for Machine Learning - Lecture 6a - Overview of mini-batch gradient descent.," 2012, doi: 10.1017/9781139051699.031.

APPENDIX A

A. CNN and RNN Layers Sample Outputs
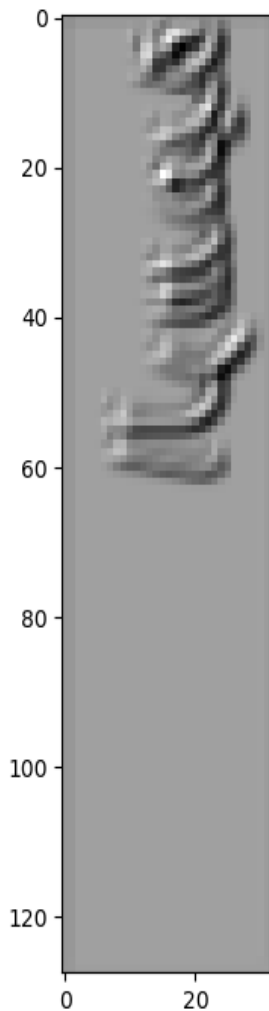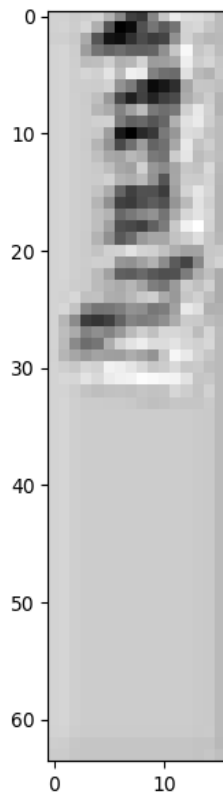
(a) CNN Layer [1,32].   (b) CNN Layer [2,32].   (c) CNN Layer [3,32].   (d) CCN Layer [4,32].   (e) CCN Layer [5,32].
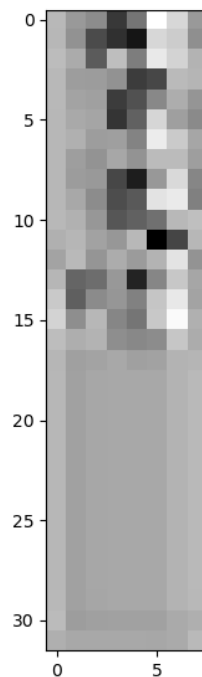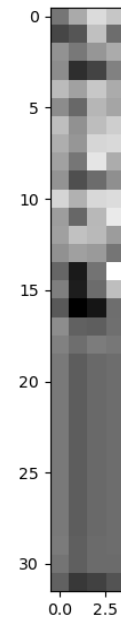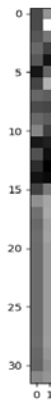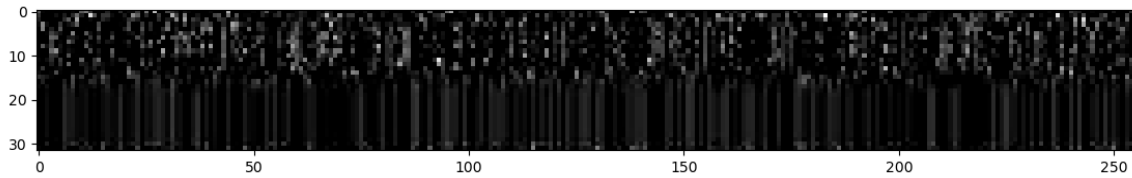
f: Final Output of the CNN (32 Time Steps × 256 Features).
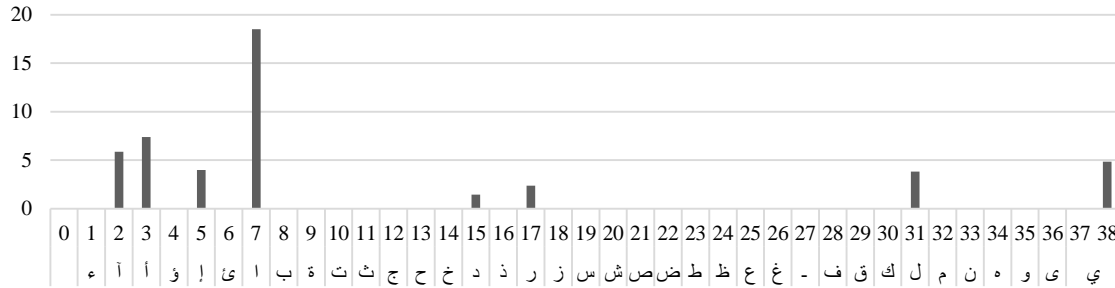
g: Time step [1 of 32] – Highest Observation at Vocabulary Position [7] = Character Alef.
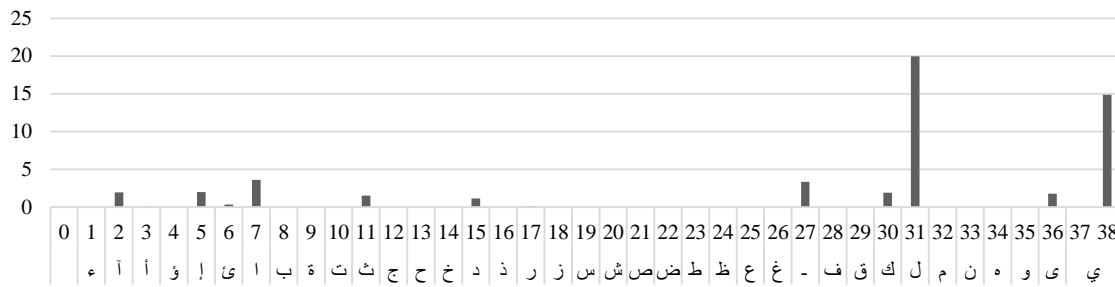
h: Time step [2 of 32] – Highest Observation at Vocabulary Position [31] = Character Laam.
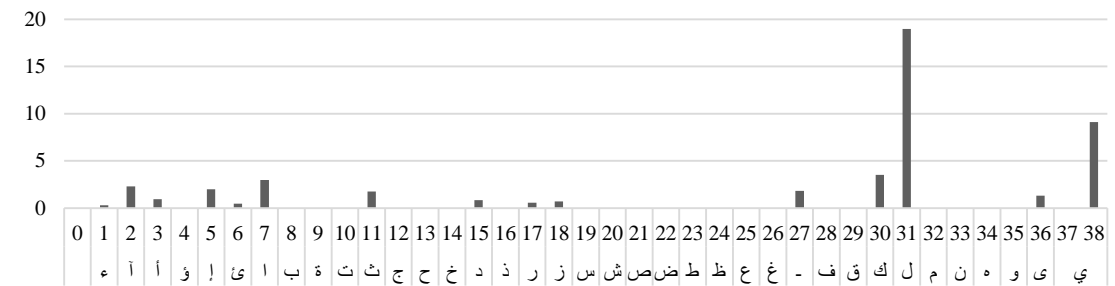
i: Time Step [3 of 32] – Highest Observation at Vocabulary Position [31] = Also Character Laam.
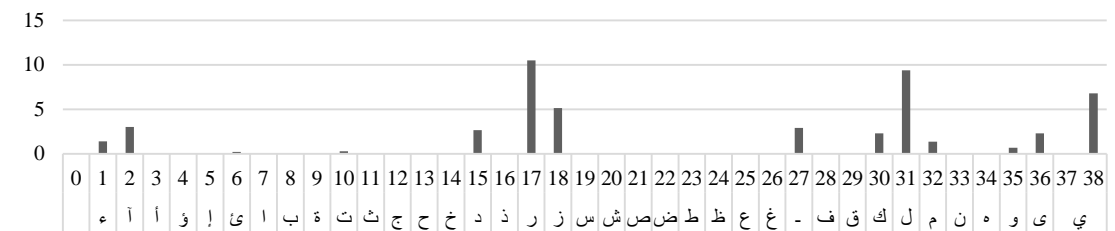
j: Time Step [4 of 32] – Highest Observation at Vocabulary Position [17] = Alphabet Raa, also High Value at Position [31] Character Laam.

Fig 8.   (a-j) CNN and RNN Layers Sample Outputs.

Cyber Security Defence Policies: A Proposed Guidelines for Organisations Cyber Security Practices

Julius Olusegun Oyelami[1], Azleena Mohd Kassim[2]

School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia

Abstract—**Many organisations have been struggling to defend their cyberspace without a specific direction or guidelines to follow and they have described and identified cyber attack as a devastating potential on business operation in a broader perspective. Since then, researchers in cyber security have come out with numerous reports on threats and attack on organisations. This study is conducted to develop and propose a Cyber Security Defence Policies (CSDP) by harmonising and synthesizing the existing practices identified from the literature review. Observation and questionnaire were adopted to evaluate, review and collect data under ethical agreement from 10 organisations. The validation is based on the principal components for the proposed CSDP and the proposed CSDP, using SPSS as the statistical tool. The result shows that, the validation of the proposed CSDP by 20 experts reveals a standard deviation of 0.607, 0.759, 0.801, 0.754, 0.513, 0.587 and 0.510 on each of the principal components without a missing value respectively. While the correlation matrix and the reproduced correlation matrix for the proposed CSDP indicated 61% and the percentage of acceptance on the principal components for the proposed CSDP are higher than 50%. Therefore, from the outcome, it has shown that the acceptance responds towards the proposed CSDP and the result from the principal components analysis (eigenvalue analysis) are significant enough for implementation and can be adopted by organisations as a guidelines for organisation cyber security practices.**

Keywords—Cyber security; cyber defence policy; organisation; cyber security practices

I.   INTRODUCTION

Many organisations have described and identified cyber attack as potentially having some devastating implications on business operation in a broader perspective. Thus, researchers in security have come out with numerous reports on threats and attack on organisations [1]. Cyber security weaknesses have been widespread, taking place on organisation and pose risk on their assets [2]. Many organisations have been struggling to defend their cyberspace without a specific direction or guidelines to follow. Recent studies have shown that, implementation of cloud security (CSe) is a measure that could be used as a security policy [3], [4]. This reasoning is supported by [1], where they stated that in implementing CSe, organisation need to evaluate IT resource with its management portfolio, aligned with the organisation's strategies and objectives to establish logical security system control to verify the confidentiality, integrity, and availability of information.

Deficiencies in organisation's cyber security implementation and policy are becoming a global issue and more challenging where it has raised growing concern among professional [6], [7]. Over the years there has been series of cyber-attack reports on organizations asset which has been noted as high-risk area to many organisations as most attacks are from the network system [4], [5], [11]. The emergence of the networking security standard like(i.e., IEEE 802.11, ISO27001 has therefore contributed to the network and cyber security but its deployment in many organisations are still few and organisation need more harmonise standards to combat cyber attacks [9], [10]. Many organisations are going into cloud without observing the threats involved [3]. While wireless network provides opportunity for greater mobility and flexibility for organisation operations, it also poses security risks to the organisation at large [2], [5], [9]. It is recommended for organisations to combine management and technical controls [7], [11], [14]. A report by [12] submitted that more cyber-attacks are expected by the fall of 2020. In line with that, a fully-owned subsidiary of Intel Corporation have also indicated that cyber-crimes are causing the global economy to lose more than $400 billion every year according to report from [4] and [13]. Cyber security involves something more than just a passive compliance with security practices. Cyber-crime and cyber-attack today is becoming inevitable to many organisation, both in government and private sector, and the best way to maintain cyber security is to implement a proactive approach towards virtual security at all times [5], [7], [11].

In this paper, we intend to identify policies which organisations can implement as security policies for cyberspace. Therefore, in this paper, it is postulated that these identified policies can be used as a blueprint for organisation cyber security practices. Our objectives in this paper are to identify several security program practice by these organisations and synchronise the identified security practices into one holistic framework for cyber security defence policies.

The gap analysis from the literature review had suggested that many organisations are still struggling to identify effective cyber security policies for information and data protection without knowing the right security policies to implement. Thus, the proposed cyber security defense policies can be useful for cyber security application.

The next section in this paper will present the related works that have been done in cyber security disaster recovery (CSDR), cyber security governance (CSG), cyber security risk management (CSRM), cyber security incident monitoring and

auditing (CSIMA), cyber security intelligent objectives (CSIO), could security (CSe) and cyber security management program (CSMP). Next, the methodology of this study is presented, followed by validating the proposed Cyber Security Defence Policies (CSDP). The proposed blueprint is presented followed by discussions and the future directions as the continuation from this paper are also presented.

II. RELATED WORKS

In [7], the cyber security is used as one of the measures that an organisation could take to mitigate cyber-attack, and this argument is further supported by [5]. Mittal [5] stated that an organisation needs to implement cyber security disaster recovery (CSDR) in respect to cyber security, which might enable the organisation to evaluate abilities to continue business operations in the case of cyber incidents. Studies have also shown that cyber security governance (CSG) play an important role in security policy making [1], [7], [8]. In a similar fashion, [4] has shown that organisations need cyber security intelligent objectives (CSIO) to withstand the challenges of security breaches and this statement is credited to the report made by National Institute of Standard and Technology [4]. They further argue that, for organisation to implement CSIO, they need periodical review of cyber security systems and information technology architecture and also perform technical security testing to identify potential threats and vulnerabilities. Mittal [5] agreed with the statement put forward by [6] that organisation's information technology practices need to be aligned with other security practices to ensure effective cyber security intelligence.

According to [6] and [8], in order to identify potential threats, organisations need to evaluate their existing IT policies, practices and data governance to ensure compliance with regulatory and legal requirements as to improve the quality and control of the organisation cyberspace. Studies by [3] and [4] presented cyber security risk management (CSRM) which was identified as one the practices required in organisation to manage risk associated with cyber attacks. They further argued that, organisations need to evaluate data classification practices for adequate alignment with the organization's policies and evaluate the information security and privacy policies practices to ensure that the physical and environmental controls of information assets are adequately safeguarded. As indicated in [2] and [5], the use of cyber security management program (CSMP) is important to outline cyber security policies and practices which will relate to the organisation's knowledge management life-cycle and cyber security life-cycle management in order to select supplier, manage project policies and practices that involve system review (software projects).

In [7], it was stated that, to determine whether project requirements are met before implementation and also to check every phases of project for bugs and other malicious elements in the system development life-cycle, a security management program is required. This could be a move to mitigate security breaches. In a similar fashion, studies conducted by [6] and [9]

expressed a concern that, organisation need to perform cyber security audit and monitor incidents in accordance with audit standards and a risk-based audit strategy and it will enable organisation to establish whether systems are protected to provide value to the organisation.

In [10], the cyber security incident monitoring and auditing (CSIMA) was claimed to be able to report and document all activities in the context of cyber security decision making, evaluate whether the risks have been sufficiently addressed and create confidence report to stakeholders. If security practices are harmonised, it could create a strong defensive mechanism in the cyberspace [9], [10], [11]. It was asserted that, people around the world will communicate and exchange information and ideas irrespective of physical location or geographical distances [5], [9], [11]. Today, the cyber world has transformed our lives and makes that possible [7], [14]. As our lives are revolving around the internet of thing (IoT), all our activities also seem to depend on it. Organisation should be aware of their current situation and know how to build a comprehensive cyber security policy. Upon weighing the organisation's susceptibility to risk, along with mitigation and assessment of their current practices, if the organisation is able to observe a gap that could be vulnerable to attacks; then, urgent action is highly recommended to consolidate, integrate and build a formidable mitigation plan.

The works by [15], [16], [17] and [18] agreed that a single security standard, security policy and security practice is not enough to combat cyber terrorism and cyber criminals due to the advancement in technology. These studies agreed that cyber crimes are transnational crimes that are affecting both organisations and humanity. Therefore, in this era, organisations need to harmonise security frameworks security standards and security practices to stand the new challenges emanating from cyber crime.

III. VALIDATING THE PROPOSED CYBER SECURITY DEFENCE POLICIES (CSDP)

In this study, a preliminary investigation using survey questionnaires was deployed to identify security practices of these 10 organisations both in Africa and Asia. The preliminary investigation was based on cyber security practices of these organisations as this study intend not to reveal the names and identity of these organisations base on the agreed ethic for data collection.

As shown in Table I, organisations labelled as 1, 2, 4, 5, and 7 observe security practices CSe, CSIO CSDR, CSG, CSRM and CSMP respectively. It was also identified that, organisation 3, 6, 8, 9 and 10 practice combination of at least two different practices. CSIMA was practiced in organisations labelled as 8 and 9, both combined with two different practices; the first combined with CSe and the latter combined with CSG. These practices (i.e., CSe, CSIO, CSDR, CSG, CSRM, CSMP, and CSIMA) are harmonized as a component to formulate and propose the Cyber Security Defence Policies (CSDP); in order to accommodate the difference practices.

TABLE I.   SUMMARY OF THE ORGANISATION IDENTIFIED SECURITY PRACTICES

Organisation Number	Identified Practices	Abbreviation
1	Cloud Security	CSe
2	Cyber Security Intelligent Objectives	CSIO
3	Cyber Security Intelligent Objectives & Cyber Security Governance	CSIO, CSG
4	Cyber Security Disaster Recovery	CSDR
5	Cyber Security Governance	CSG
6	Cyber Security Risk Management & Cyber Security Disaster Recovery	CSRM, CSDR
7	Cyber Security Management Program	CSMP
8	Cyber Security Incident Monitoring and Auditing & Cloud Security	CSIMA, CSe
9	Cyber Security Incident Monitoring And Auditing & Cyber Security Governance	CSIMA, CSG
10	Cyber Security Risk Management and Cyber Security Governance & Cyber Security Governance	CSRM, CSG

A. About the Organisations

These ten organisations are both private and government organisations. The government organisations were established more than a decade ago while the private organisations were established not less than six years. These organisations have been identified as having good reputations for cyber security programs based on the result from the investigation and they have successfully configured applicable security management policies in-place. Each of these policies outlined specific examples on techniques and control used by these organisations to increase their security program's effectiveness.

B. Validating Procedure

In this work, the seven security practices are benchmarked and organised under several technical practices identified from the preliminary investigation conducted on 10 organisations from two continents, specifically in Malaysia and Nigeria. We validated these practices using 20 experts from the cyber security area and information security discipline. The experts are in a positions of Information Security Officer's (ISO), Information Technology Manager's (ITM) and Chief Information Officer's (CIO). We selected our independent and dependent variables and the scale of measurements was adjusted to 0.05, where value that is less than 0.05 will indicate rejection and value more than 0.05 will indicate acceptance.

The variables and scaling sampling takes account of exact values from variables such as age, gender, job title and years of work experience. Some variables are assigned with weights to indicate the different level of importance of the respective variables' values, such as; Level of certification in Information Technology i.e. Basic=5, Intermediate=10 and Expert=15, Level of education i.e. Diploma=5, B.Sc=10, MSc=15 and PhD=20, Area of Expertise i.e. Information security=1, Information security management=2, and Cyber security =3.

The validation is based on (1) the proposed Cyber Security Defence Policy (CSDP) and (2) the principal components (i.e. content analysis) observed from the 10 organisations. Data was collected using questionnaire and the design of the questionnaires were based on the security practices identified from both the literature review and the identified security practices during preliminary investigation using survey questionnaire as a component in formulating the proposed CSDP.

C. Statistical Procedure

Initially, this study carry out most of the descriptive statistics concerning the variables assessed using Likert scale. Our test focus on standard multiple regression and linear regression using Statistical Package for Social Science (SPSS). Subsequently, this study proceeds to determine each of the components within the Cyber Security Defence Policies (CSDP) using factor analysis and eigenvalue analysis. This eigenvalue analysis is a linear operation that will provide the properties of the CSDP structure. Reliability on the security practices was determined using Cronbach's alpha coefficient. The questionnaire focused on the identified practices for the content analysis for the proposed CSDP. Likert scaling was deployed to anchor the responses and to enable the respondents to read meaning to the questionnaires (Strongly agree=5, Agree=4, Neutral=3, Disagree=2, and Strongly disagree=1). The ANOVA experiment was considered by using SPSS statistical analytical tool using descriptive methods and to established correlation between the total score using factor analysis and the results are presented in tables and graph for visualization. In our test for regression analysis, scale of measurement between 0.35 and 0.50 was used as an average value.

D. Data Analysis

In the data analysis, as shown in Table II and Table III, where they depict the descriptive analysis on the propose Cyber Security Defence Policies (CSDP) and the total variance of the CSDP on the seven components for the proposed CSDP. Where as Table IV and Table V illustrate the correlation matrix and the reproduced correlation matrix on the seven CSDP components for the propose CSDP. Based on the validation of the proposed CSDP by 20 experts, the descriptive analysis reveals a standard deviation of 0.607, 0.759, 0.801, 0.754, 0.513, 0.587 and 0.510 on each of the seven principal component without a missing value respectively, as illustrated in Table II.

This indicates that, the seven principal component for the proposed CSDP are reliable to be implemented. In calculating the matrix of correlation between items, we excluded value lower than 0.25. For example, the value between 0.09 and 0.24 are considered as a small value, while value between 0.35 and 0.50 can be considered as an average value in this study. While conducting the analysis to explore the setting of our questionnaires based on the seven principal components, it allows us to check each of our parameters on a set of variables obtained. The results point to the extraction of minimum of 25% and maximum of 50% with a total percentile of 75% as shown in Table II. Fig. 1 depicts the scree plot for the eigenvalue analysis, which shows the line plot of the

eigenvalue of factors or the principal components in an analysis. Since we retained seven factors as components for the proposed CSDP, this eigenvalue analysis provides the properties of the CSDP as a structure and represent the value for each components. This shows that the proposed CSDP based on its seven components, increased in value with 2.4. This indicate that the proposed CSDP can be brought forward as a guideline in cyber security practices in organisation.

Fig. 2 illustrates the percentage of acceptance based on the seven components. Fig. 2 explains the result from the validation of the proposed CSDP by security experts, and it also illustrates the acceptance rate of each of the principal components for the proposed CSDP (CSe 46%, CSIO 60%, CSDR 65%, CSG 60%, CSRM 60%, CSMP 62% and CSIMA 69%) The correlation matrix and the reproduced correlation matrix are shown in Table IV and Table V for the proposed CSDP, indicate 61% with an absolute values greater than 0.05. This means that from the factors analysis based on the ANOVA test results on the seven components for the proposed CSDP by the experts has shown that, the acceptance respond towards the proposed CSDP are significant enough for implementation and can be adopted by organisations as a blueprint for organisation cyber security policies.

TABLE II. THE DESCRIPTIVE ANALYSIS ON CYBER SECURITY DEFENCE POLICY (CSDP)

		CSe	CSIO	CSDRP	CSG	CSRM	CSMP	CSIMA
N	*Valid*	20	20	20	20	20	20	20
	Missing	0	0	0	0	0	0	0
Std. Deviation		.607	.759	.801	.754	.513	.587	.510
Variance		.368	.576	.642	.568	.263	.345	.261
Minimum		3	3	3	3	4	3	4
Maximum		5	5	5	5	5	5	5
Percentiles	*25*	4.00	4.00	4.00	4.00	4.00	4.00	4.00
	50	5.00	5.00	4.50	5.00	4.50	4.00	5.00
	75	5.00	5.00	5.00	5.00	5.00	5.00	5.00

TABLE III. THE TOTAL VARIANCE OF THE CSDP PRINCIPAL COMPONENTS

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	2.308	32.970	32.970	2.308	32.970	32.970
2	1.623	23.185	56.155	1.623	23.185	56.155
3	1.286	18.376	74.531	1.286	18.376	74.531
4	.716	10.231	84.762			
5	.520	7.423	92.185			
6	.358	5.110	97.296			
7	.189	2.704	100.000			

Extraction Method: Principal Component Analysis.

TABLE IV. THE CORRELATION MATRIX ON THE CSDP COMPONENTS

		CSe	CSIO	CSDRP	CSG	CSRM	CSMP	CSIMA
Correlation	CSe	1.000	.057	.216	.230	.169	.074	.425
	CSIO	.057	1.000	.459	.221	-.203	.455	.143
	CSDRP	.216	.459	1.000	.401	-.256	.324	.090
	CSG	.230	.221	.401	1.000	.272	-.095	.082
	CSRM	.169	-.203	-.256	.272	1.000	-.437	.101
	CSMP	.074	.455	.324	-.095	-.437	1.000	.553
	CSIMA	.425	.143	.090	.082	.101	.553	1.000
Sig. (1-tailed)	CSe		.405	.180	.165	.238	.379	.031
	CSIO	.405		.021	.175	.196	.022	.274
	CSDRP	.180	.021		.040	.138	.081	.353
	CSG	.165	.175	.040		.123	.345	.365
	CSRM	.238	.196	.138	.123		.027	.337
	CSMP	.379	.022	.081	.345	.027		.006
	CSIMA	.031	.274	.353	.365	.337	.006	

TABLE V. THE REPRODUCED CORRELATION MATRIX ON THE CSDP COMPONENTS

		C.Se	CSIO	CSDRP	CSG	CSRM	CSMP	CSIMA
Reproduced Correlation[a]	**C.Se**	.615[a]	.100	.165	.360	.362	.179	.600
	CSIO	.100	.617[a]	.651	.304	-.373	.488	.152
	CSDRP	.165	.651	.750[a]	.517	-.243	.361	.091
	CSG	.360	.304	.517	.784[a]	.343	-.173	.013
	CSRM	.362	-.373	-.243	.343	.723[a]	-.527	.079
	CSMP	.179	.488	.361	-.173	-.527	.860[a]	.572
	CSIMA	.600	.152	.091	.013	.079	.572	.867[a]
Residual[b]	**C.Se**		-.043	.052	-.130	-.193	-.105	-.176
	CSIO	-.043		-.192	-.084	.170	-.033	-.010
	CSDRP	.052	-.192		-.117	-.013	-.037	-.001
	CSG	-.130	-.084	-.117		-.071	.077	.069
	CSRM	-.193	.170	-.013	-.071		.090	.022
	CSMP	-.105	-.033	-.037	.077	.090		-.019
	CSIMA	-.176	-.010	-.001	.069	.022	-.019	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 13 (61.0%) non redundant residuals with absolute values greater than 0.05.
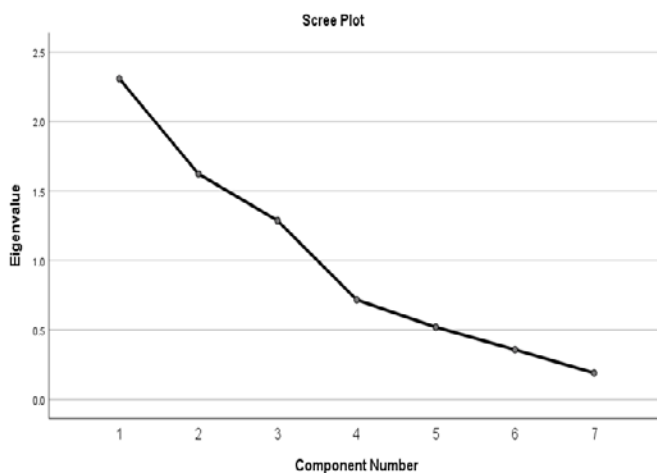
Fig. 1. Scree Plot for the Eigenvalue Analysis (Principal Components Analysis) for the Proposed CSDP.
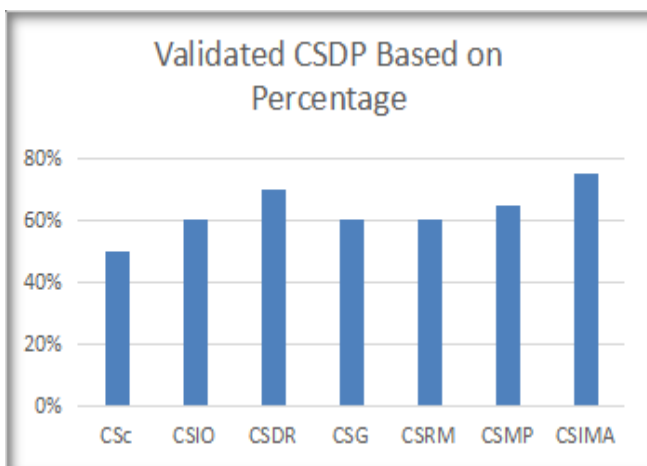
Fig. 2. The Validated CSDP based on Percentage of Acceptance.

IV. RESULT

Based on the data analysis and the validation of the proposed Cyber Security Defence Policies (CSDP) by 20 experts, the descriptive analysis reveals a standard deviation of 0.607, 0.759, 0.801, 0.754, 0.513, 0.587 and 0.510 on each of the principal component without a missing value, respectively, as illustrated in Table I. It indicates that the seven principal component for the (CSDP) are reliable for implementation. In order to calculate the matrix of correlation between items, value lower than 0.25 were excluded. We used the value of 0.09 and 0.24 can be considered as a small value, while value between 0.35 and 0.50 can be considered as an average and the value between the 0.50 and 2 can be interpreted to be a big or large value, therefore, we eliminate all indicators with a correlation less than 0.35.

While conducting analysis to explore the setting of our questionnaires based on the seven principal components analysis, we are able to check each of our parameters on a set of variables obtained. The results point to the extraction of three factors with about 75% of the total variance as shown in Table I. Fig. 1 depicts the screen plot that corroborates this analysis and Fig. 2 illustrates the percentage of acceptance. The correlation matrix and the reproduced correlation matrix are shown in Table III and Table IV.

For the proposed CSDP, it indicates 61% with an absolute values greater than 0.05. Based on the factors analysis, the ANOVA test results on the proposed CSDP and the principal components (eigenvalue analysis) of the CSDP by the experts, the acceptance responds towards the proposed CSDP and its principal components are significant enough for implementation and can be adopted by organisations as a blueprint for organisation cyber security practices. In this study, finding shows that 1) not all organisations are practicing most of the identified security practices, 2) these organisations implemented specific practices according to their need while neglecting other aspect of security measures, 3) the formulated

CSDP might be more effective than the existing security practices due to its completeness with other security practices identified from other organisations, 4) the proposed CSDP might be a complete and harmonised security practices that can support the entire security operations and defend the organisations from cyber criminals and cyber attacks against valuable assets (i.e., information, data and intellectual properties) and 5) implementing CSDP will cover a wide range of security management and create insight on how organisations should create an effective security policies to guide security practices.

V. DISCUSSION

This paper presents the proposed Cyber Security Defence Policies (CSDP) that could be used by any organisation as a blueprint to follow and to guide to secure data and information. This CSDP is recommended for: 1) organisations that intend to plan, develop, implement cyber security policies, 2) individuals or information technology professional i.e. computer system and network engineer, who design, deploy, administer and maintain organisation network security systems, 3) individuals/IT professional who are information security or network security personnel with information system, monitoring responsibilities on organisational information security management, 4) management personnel who might require a technical basis for supporting a decision-making process and 5) those wishing to increase their knowledge on cyber security policies as well.

Fig. 3 illustrates the proposed CSDP as a guideline for organisation to follow when implementing cyber security practices. Organisation that intends to implement Cyber Security Defence Policies (CSDP) should identify the seven basic components within the proposed CSDP. This involves the implementation of Cloud Security (CSe). Cloud security enables organisation to check on IT vendors' products properly before implementing the product to avoid security risk, and it also allows the organisation to secure their data in the cloud as well as create opportunity for organisation to identify risk of mistaken identity during and after implementing authentication platform for data and information users. Data in the cloud need encryption standards and procedures mechanism. With the implementation of CSDP, the organisation will have the advantage to detect the threats that posed a great danger to organisation infrastructure and enhance identity and access management control.
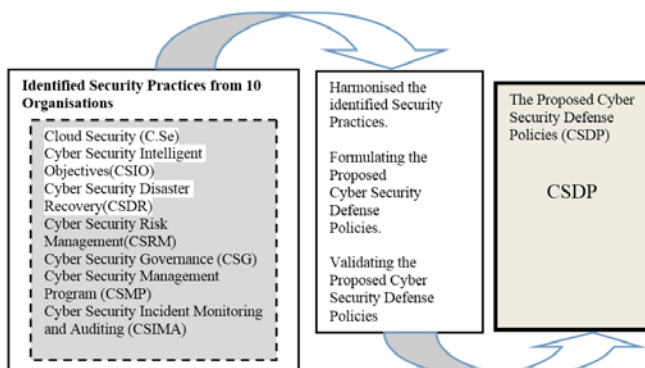
Fig. 3. The Proposed Cyber Security Defence Policies (CSDP).

The implementation of CSDP will enhance Cyber Security Intelligent Objectives (CSIO) for the organisation by improving the role base access control. This enables the organisation to access role and monitor duty that are performed by computer system. Besides, it enables the organisation to have technical control on software development and operation towards business continuity.

CSDP provides information security governance, risk management and cryptography as part of the security intelligent objectives and enhance security frameworks, legal regulation, cybercrime investigation and compliance by setting objectives for security operation, physical and environmental security. These enable the organisation to have both technical and management control on cyber security architecture, telecommunication, network security and the management control that focus on technology, people and leadership.

Besides that, the implementation of CSDP creates a platform for organisation Cyber Security Disaster Recovery (CSDR) which enables the organisation to resume business operation after cyber incident. It also enables the organisation to be resilience by creating cyber security events management team, network and digital forensic team, cyber security endpoint forensic team, cyber log management team, cyber security monitoring and response team. These teams enable organisation to focus on vulnerability intelligence analysis, cyber threats intelligence analysis, compliance and risk analysis, insider and outside threat analysis, advance persistence threat (APT) analysis, IT vendor analysis for effective incident recovery and mitigation.

Implementing CSDP means operating at the edge of security policies which will improve the organisation economic value due to Cyber Security Governance (CSG) provided. Governance involve the identification of both internal and external influences to the organisation (i.e. emerging technologies, social media, business environment, risk tolerance, regulatory requirements, third-party considerations, threat landscape) to ensure that these factors are continually addressed. Governance is all about IT leadership that focus on maintaining information security governance, to ensure organisational goals and objectives are supported by the information security program, to ensure security strategy are aligned with organisational goals and objectives to guide the establishment and improve the management of information security program which further create confidence to clients and stakeholders.

The implementation of CSDP will also improve the overall management of security practices because, it will enable the organisation to use the Cyber Security Risk Management (CSRM) to a better advantage by improving technical report on non-compliance and other changes in information risk. This is to facilitate the risk management decision-making process through appropriate and effective management of risk to an acceptable level and standards, monitor the internal and external factors using risk indicators, threat landscape, geopolitical and regulatory change. This may require reassessment of risk to ensure that changes to existing, or new risk scenarios are identified and managed appropriately. In order to manage risk to acceptable levels based on

organisational risk appetite, appropriate risk treatment options can be implemented.

Cyber Security Management Program (CSMP) is another valuable component of the CSDP. It involves a program which focuses on security issues. It aligns security program with the operational objectives of other business functions (i.e. human resources, accounting, procurement and IT) to ensure that the security program adds value to and protects the business operations. It also promotes a program for Cyber Security Awareness and Training (CAT) to foster an effective security culture that includes people and technologies. This is done by compiling and present reports to key stakeholders on the activities, trends and overall effectiveness of the security management program and the underlying business processes in order to communicate security performance.

Cyber Security Incident Monitoring and Auditing (CSIMA) is another component within the CSDP. It complements other components as it relates to security monitoring and audit. Cyber Security Incident Monitoring and Auditing (CSIMA) is also one of the components within the CSDP. It works as a complement to other components as it relates to security monitoring and audit. Its benefits includes: 1) maintaining an organisational definition of severity hierarchy security incidents to allow accurate classification, 2) categorisation of response to incidents by creating opportunity for organisation to organise the level of risk i.e., low risk, medium risk and high risk, 3) training and equipping incident response teams to respond to security incidents in an effective and timely manner, 4) support to the organisation to conduct post-incident reviews to determine the root cause of incidents in other to develop corrective actions, reassess risk, evaluate response effectiveness, maintain communication plans and processes, 5) support to the organisation to manage communication with internal and external entities and maintain processes to investigate and document related to security incidents in order to determine the appropriate response.

Implementing CSDP, organisation will be able to test, review and revise incident response plan periodically to ensure an effective response to security incidents and to improve response capabilities, this enable the organisation to build security capability towards cybercrimes and cyber criminals. With the CSDP, organisation will be able to test, review and revise (as applicable) the incident response plan periodically to ensure an effective response to security incidents. Response capabilities are expected to improve and CSDP enables the organisation to build security capability towards cybercrimes and cyber criminals.

The CSDP is expected to provide basic set of security practices that are in conjunction with other relevant security practices. Therefore, from this study, it is posited that the proposed SCDP can be a tool for security controls and when adequately implemented can be effective in reducing cyber security risks on organisation asset.

VI. CONCLUSION

In achieving the objectives and established findings in the study, the following are applied: 1) SPSS data analysis that focus on descriptive analysis and standard multiple regression are applied to achieved the stated results, 2) eigenvalue analysis also known as factor analysis was applied to determine reliability on the seven identified principal components for the proposed CSDP, 3) survey questionnaire are set to determine the level of security practices and to identity the security practices from the selected organisations, 4) validation questionnaire are set based on the identified security practices from the selected organisations to be validated by 20 experts from security discipline.

The summary of findings in this study had revealed that implementing CSDP will cover a wide range of security management and create insight on how organisations should create an effective security policies to guide security practices. In comparison with prior studies, which shows that most organisations limit security practices according to their needs without considering other security factors or measures. The implementation of the proposed CSDP will create a wider security measure for organisation security program due to its seven principal components that are selected from a wide range of security program from reputable security organisations.

Limitations to this study are also identified in the area of data collections and the validation of the proposed CSDP. During data collections, not all the respondent respond to all questions, few questions are left out. These few unanswered questions are excluded from the data entry. Aside from these limitations, the 10 organisation for data collection and 20 experts for the validation of the proposed CSDP responded enough to suggest and make conclusion in this study that the proposed CSDP can be implemented as a cybersecurity policies and create a guidelines for organisations security practices.

From the data analysis outcome and validation of the proposed CSDP and the result from the eigenvalue analysis of the principal components for the proposed CSDP, it has shown that positive acceptance responds towards the proposed CSDP and that the proposed CSDP are significant enough for implementation, thus can be adopted by organisations as a guideline for organisation cyber security practices to mitigate cyber crimes and attacks.

In summary, from the identified cyber security practices from these 10 organisations, the harmonisation of the identified security practices to formulate cyber security defense polices (CSDP) and the validation of the proposed CSDP have been the achievements in this work.

The finding suggests that implementing CSDP cover a wide range of security management and create insight on how organisation should effectively manage and create security policies within security program. CSDP can also support organisation in testing, reviewing and revising incident response plan periodically to ensure an effective response to security incidents and to improve response capabilities towards cybercrimes, cyber criminals and cyber attacks simultaneously. This indicate a scientific value to cybersecurity program that is achieved in this paper.

In this case, this study recommends a new CSDP as a guideline for organisations cyber security practices and as it has been validated by experts from reputable organisations that

have succeeded in cyber security policy implementation. In conclusion, it is highly recommended for organisations to implement a set of guidelines like the Cyber Security Defence Policies (CSDP) for managing and mitigating threats and put in place a governance that will govern how a system must be used.

VII. FUTURE WORK AND RECOMMENDATION

Paying attention to cyber security is the most effective way to help the next generation understand the importance of virtual safety. Today, youngsters are at risk of numerous cyber threats ranging from hackers trying to encourage inadvertent disclosure of private information to individuals resorting to cyber bullying, harassment, and social embarrassment. A proactive approach towards cyber security will help the next generation imbibe lessons that will help them deal with such challenges in a confident and effective manner.

The ability to consistently classify information at all points in its life cycle and across the entire IT infrastructure is critical. If the information cannot be classified correctly, then it will not be able to be managed appropriately. Static classification of information by the information owner is not workable in today's global environment and so consistent automation is also required. In the future, Artificial Intelligent (AI) can be applied with some enhancement towards cyber security that will take account of machine learning and deep leaning, using python for large cyber security incident data-set to improve decision-making, knowledge sharing and to establish effective cyber security policies in organisations.

ACKNOWLEDGMENT

REFERENCES

[1] E. J. Murray and A. Durcikova, "Integrating IS security with knowledge management: Are we doing enough to thwart the persistent threat?" 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, 2014, pp. 3452-3459.

[2] N. Perlroth, M. Scott, and S.Frenkel, "Cyberattack hits Ukraine then spreads internationally," The New York Times, June 27, 2017, Accessed: October 5, 2019. [Online]. Available: https://www.nytimes.com/2017/06/27/technology/ransomware-hackers.html

[3] C. Skouloudi and M. Fernandez, "Towards secure convergence of cloud and IoT," The European Union Agency for Cybersecurity (ENISA), September 17, 2018, Accessed: October 5, 2019. [Online]. Available: https://www.enisa.europa.eu/news/enisa-news/towards-secure-convergence-of-cloud-and-iot

[4] M. P. Barrett, "Framework for Improving Critical Infrastructure Cybersecurity Version 1. 1," April 16, 2018, Accessed: October 24, 2019. [Online]. Available: https://www.nist.gov/publications/framework-improving-critical-infrastructure-cybersecurity-version-11

[5] S. Mittal, "Knowledge for cyber-threat intelligence," Ph.D. dissertation, Univ. of Maryland, MD, USA, 2019.

[6] E. O. Yeboah-Boateng and E. Akwa-Bonsu, "Cyber-security intelligence gathering: Issues with knowledge management," in Cyber Security and Threats: Concepts, Methodologies, Tools, and Applications, IGI Global, 2018, pp. 1454-1479.

[7] T. Poppensieker and R. Riemenschnitter "A new posture for cybersecurity in a networked world," in Perspectives on transforming cybersecurity, Digital McKinsey and Global Risk Practice, 2019, pp. 18-26.

[8] K. Min, C. S. Chai, and M. Han, "An international comparative study on cyber security strategy," International Journal of Security and its Applications, vol. 9(2), 2015, pp. 13-20.

[9] D. Manky, "Threats in the information age," in Cyber Security Threats, Challenges and Opportunities, Australian Computer Society, 2018, pp. 14-15.

[10] P. S. Seemma, S. Nandhini, and M. Sowmiya, "Overview of cyber security," International Journal of Advanced Research in Computer and Communication Engineering, vol. 7(11), 2018, pp.125-128.

[11] B. Mat, S. D. M. Pero, R. Wahid, and B. Sule, "Cybersecurity and digital economy in Malaysia: Trusted law for customer and enterprise protection," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8(8S3), 2019, pp. 214-220.

[12] CSIS, "Net losses: Estimating the global cost of cybercrime. Economic impact of cybercrime II", CSIS (Centre for Strategic and International Studies), June 5, 2014, Accessed: [Online]. Available: https://www.csis.org/analysis/net-losses-estimating-global-cost-cybercrime

[13] S. McKune, "An Analysis of the International Code of Conduct for Information Security", Citizen Lab, September 28, 2015, Accessed: June 26, 2018 [Online]. Available: https://citizenlab.ca/2015/09/international-code-of-conduct/

[14] E. Minei and J. Matusitz, "Cyberspace as a new arena for terroristic propaganda: an update examination," International Journal of Ethics of Science and Technology Assessment, vol. 9(1), 2012, pp. 163–176.

[15] L.M. Rhode, "Human traficking as cybercrime," AGORA International Journal of Administrative Science, vol. 1(1), 2017, pp.23-29.

[16] M. N. Katsantonis, I. Kotini, P. Fouliras and I. Mavridis, "Conceptual framework for developing cyber security serious games," In Global Engineering Education Conference (EDUCON) 2019 IEEE, pp. 872-881, Dubai, United Arab Emirate, 2019.

[17] M. Mahmud, M. S. Kaiser, M. M. Rahman, M. A. Rahman, A. Shabut, S. Al-Mamun and A. Hussain, "A brain-inspired trust management model to assure security in a cloud based IoT framework for neuroscience applications" Journal of Cognitive Computation. vol.10, 2018, pp. 864-873.

[18] T. Poppensieker and R. Riemenschnitter, "A new posture for cybersecurity in a networked world," in Perspectives on transforming cybersecurity, D.Chinn, J. M. Kaplan, and T. Poppensieker, Digital McKinsey and Global Risk Practice, 2019, pp. 18-26.

New RTP Packet Payload Shrinking Method to Enhance Bandwidth Exploitation Over RTP Protocol

AbdelRahman H. Hussein[1], Mwaffaq Abu-Alhaija[2]

Department of Networks and Information Security
Al-Ahliyya Amman University (AAU)
Amman, Jordan

Kholoud Nairoukh[3]

Department of Computer Science
Al-Ahliyya Amman University (AAU)
Amman, Jordan

Abstract—One of the pillars to run businesses is the telecommunications. Most of the institutions are migrating, if not already migrated, to Voice over Internet Protocol (VoIP) technology. However, VoIP still need some improvements, in terms of networks bandwidth exploitation and VoIP call quality, to meet the businesses expectations. Networks bandwidth exploitation, which is our concern in this paper, has been enhanced using different approaches and methods. This paper suggests a new method to enhance networks bandwidth exploitation Packet's payload shrinking (compression) approach. The suggested method works with the RTP protocol and called RTP Payload Shrinking (RPS) method. As the name implies, the RPS method will reduce the size of the RTP packet payload, through shrinking it based on specific algorithm, which enhances the networks bandwidth exploitation. The RPS method utilizes the RTP fields to store the values that are needed to apply the shrinking algorithm at the sender and receiver sides. The effectiveness of the proposed RPS method has been examined in comparison to conventional RTP protocol without shrinking. The deployment result showed that the saved bandwidth ratio has reached up to nearly 17% in the tested scenarios. Therefore, enhancing the network bandwidth exploitation.

Keywords—VoIP; RTP protocol; payload compression; bandwidth exploitation

I. Introduction

Telecommunication systems are critical for businesses nowadays. In the last decade, the current traditional telecommunication systems (public switched telephone network) has witnessed very high steep usage curve. This because they do not provide business with the needed features. These features include group calls, video calls, voice to text call conversion, and … etc. As an alternative, Voice over IP (VoIP) has emerged and propagated in all businesses because it supports most, if not all, the needed features [1,2]. As the name implies, unlike traditional telecommunication systems, VoIP works over IP networks and it does not need a separate installed network. This is definitely saves huge efforts of the IT departments in businesses because they are managing and maintaining one network, instead of a separate network for telecommunications and IP data [3,4]. Nevertheless, traditional telecommunication systems provide a better voice call quality than VoIP technology. This is because traditional telecommunication systems use a dedicated channel for the voice call while VoIP use shared channel over packet based networks. Therefore, VoIP packet might be delayed or lost which degrades the call quality. Apart from call quality, VoIP

encounter a bandwidth efficiency utilization dilemma that needs to be dealt with and overcome [5,6].

There are three main strategies that have been followed by VoIP developer to handle this dilemma. The first one is VoIP packets concatenations, which is concatenate a number in one header instead of one header to each VoIP packet [7]. The second strategy is VoIP packet header compression, which is reduces the 40 bytes typical VoIP packet header to two or four bytes [8]. The third strategy is VoIP packet payload compression, which uses several tactics to reduce the payload size. One of the most used tactics is using a device or program, called codec, to compress the voice digital data based on some compression mechanisms [9]. Another tactic is using voice activity detection to unsend the silence of the call [10]. This paper proposes a new tactic to compress the VoIP packet payload and save the VoIP network bandwidth. The proposed tactic finds out the difference between the consecutive VoIP packets payload using certain algorithm. The proposed method does not cause any additional header to be added to the VoIP packet because it employs the existing unused VoIP packet header fields whenever needed.

There are several types of protocols that are used with VoIP technology. The IP, UDP, and RTP are typically forms the VoIP packet header. RTP is used with different media applications such as voice call, video call, multimedia conferencing, etc. Fig. 1 shows the 12-byte RTP protocol format. In case of voice call, many of RTP protocol header fields, such as the four bits contributing source (CSRC) count (CC) field, are unused and their values are zeros during the entire call. The proposed method in this paper uses this field (CC field) in its internal algorithm to achieve the payload compression and save the bandwidth. Therefore, as mentioned earlier, no additional header is added by the proposed method to the typical VoIP packet header [11, 12].

0-1	2	3	4-7	8	9-15	16-31
V	P	X	CC	M	PT	Sequence Number
Timestamp						
Synchronization source (SSRC) Identifier						
Contributing Source (CSRC)						
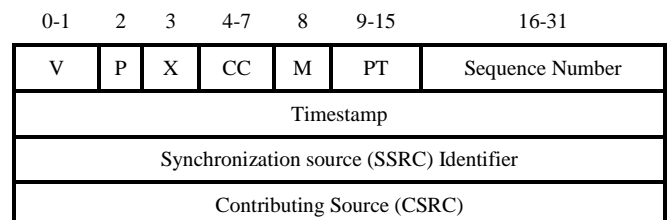
Fig 1.   RTP Protocol Header Format.

The rest of this paper is organized as follows. In Section 2, the existing VoIP bandwidth enhancement methods, which related to this work, will be discussed. In Section 3, the entities and procedures of the proposed method will be discussed. In Section 4, the proposed method will be implemented and evaluated. Finally, Section 5 concludes the paper.

II. RELATED WORKS

The dilemma of wasting the network bandwidth when using has addressed by great efforts of network researchers. To realize these efforts, this section explains some of the key method that promote the bandwidth usage by the VoIP applications.

The typical VoIP packet header contains 12 bytes RTP, 8 bytes UDP, and 20 bytes IP protocols, in total 40 bytes. Sandlund, K. et al realizes that several fields of these protocols are predetermined when start the call and not changing in all the same call packets. Therefore, these fields can be send at the beginning of the call and removed from all the call's packets. In addition, Sandlund, K. et al realizes that most of the other fields in these protocols are changing based on certain pattern. Therefore, this pattern can be calculated using certain mathematical equations. Thus, again, these fields can be removed from all the call's packets and retrieved at the receiver side using these equations. Based on these two facts, Sandlund, K. et al have managed to optimize the RTP/UDP/IP header from 40 bytes to two bytes. This method is known as RObust Header Compression (ROHC). Accordingly, when using ROHC method, the wasted bandwidth resulting from the 40 RTP/UDP/IP header has reduced by 95%, thus, very high bandwidth saving ratio [13].

Another strategy to address the dilemma of wasting the network bandwidth when using the VoIP applications is VoIP packets concatenations. Majeed, A. et al have suggested a VoIP frame concatenation method that works at the data link layer of the wireless network. The suggested method concatenates the VoIP frames in one RTP/UDP/IP/Ethernet header, rather than a standalone header to each packet, over multi-rate wireless LAN. To suit the multi-rate wireless LAN, the Majeed, A. et al based their method on two heuristics. First, the incoming VoIP frames are divided based into groups based on their transmission rate. Then, the VoIP frames within the same group only are concatenated. Second, permit the VoIP frames with a small margin of cross data rate between the VoIP frames groups to be concatenated. These heuristics have improved the number of synchronous VoIP calls by up to 200%, when using a single 802.11g Access Point, and lessens WLAN delays. Therefore, the suggested method has highly improved the bandwidth saving ratio [14].

The previous two strategies, VoIP packets concatenation and VoIP packets header compression, has been combined together in one method by Sze et al. At first, the only RTP protocol within the VoIP packet header is optimized from 12 bytes to two bytes. The new RTP protocol header is attached to a VoIP packet payload and forms a runt packet. Then, these runt packets are concatenated in one UDP/IP header at the transport layer. The proposed method works with the H.323 VoIP signaling protocol. Therefore, the process and steps that

performed by H.323 to initiate the VoIP call have been modified, in order to inform the receiver about the new format of the arrived VoIP packets. The combination of the VoIP packets concatenation and VoIP packets header compression has highly improved the bandwidth saving ratio, where the results showed that the bandwidth utilization has been improved by up to 300% in the tested scenarios. [15]

Similar to previous work, Abualhaj et al. have, also, combined two strategies to improve bandwidth saving ratio of VoIP applications. However, the proposed method on this work combined VoIP packets concatenation and VoIP packet payload compression. At the sender side, the incoming packet are separated into VoIP packet payload and VoIP packet header (RTP/UDP/IP). VoIP packet payload is compressed based on delta algorithm. The core idea of the delta algorithm is to find the difference of the successive VoIP packet payload to produce a small payload. After that, a new small header along with the RTP header are attached to the small payload to produce a small packet. Finally, several small packets are combined together in one UDP/IP header and the resulted packet is transmitted to its destination. At the receiver side, the opposite steps and processes are occurred to restore the original packet size and format. The evaluation result of the Delta-Multiplexing method showed that the bandwidth utilization was improved by up to 72% when using LPC Codec [16].

As mentioned earlier, one of the key strategies to lessen the network wasted bandwidth ratio caused by the VoIP applications is VoIP packet payload compression. In 2020, one of the key VoIP packet payload compression methods, called voice frame pruning (VFP), was proposed. The main idea of the proposed VFP method is to eradicate group of bits from the head or tail of the VoIP packet payload based on specific pruning algorithm. The pruning algorithm counts the bits (ones or zeros) at the head or tail of the VoIP packet payload and eradicate the higher number of bits. The VFP method make use of the unused fields of the RTP header to indicate the location and number of the eradicated bits. The receiver side restore the VoIP packet payload to its original size based on the value in these fields. Though it's simple, the proposed VFP method achieved a considerable bandwidth saving ratio by up to eight percent in the tested cases [11]. This paper proposes a novel payload compression method to improve bandwidth utilization over RTP protocol. Unlike the discussed payload compression methods, the proposed method focuses only on compact the voice frame without aggregation. In addition, no additional header is needed to restore the original size of the voice frame, because the proposed method reemploys and utilizes the current CC field of the RTP protocol header. The proposed method is called RTP Payload Shrinking (RPS). The details of the RPS method is discussed in the next section.

III. THE PROPOSED RPS METHOD

This section shows the detailed process of the proposed RPS method. RPS method is designed to handle the poor network bandwidth exploitation of the VoIP technology. As the name implies, the RPS method will achieve this by reducing the size of the RTP packet payload through

shrinking. In general, the RPS method treats the data of the RTP payload as a numerical value. Doing so allows making mathematical calculation on the RTP payload. Based on that, the RPS method will calculate the difference of the sequences RTP payload through subtraction. Knowingly, the result of the subtraction operation produces smaller value that the minuend value. Therefore, performing subtraction on the sequences RTP payload produces a smaller payload and, thus, achieves better bandwidth exploitation. To demonstrate, denote to one RTP payload as i and the other RTP payload as j. The size of payload i is fifteen bits "111001100110101" and the size of payload j is fifteen bits "110011101001001". Then, the difference between payload i and payload j (i - j) is thirteen bits 101111101100, which is smaller than the normal payload size (110011101001001), thus improves the bandwidth exploitation. Please note that the value of payload i and j are not real and only for demonstration purpose. The real value of an RTP packet payload could reach 240 bits or even more deepening on the use Codec. Therefore, the bandwidth saving resulting from the subtraction operation is considerable. Section 3.1 and section 3.2 demonstrate the RPS method process at the sender and receiver side, respectively.

A. RPS Method Processes at Sender Side

At the sender side, the RPS method shrinks the size of the RTP packet payload by performing several processes. First, the RTP payload is dissociated from the header. Second, the difference between the RTP payloads that are following each other is found. For example, suppose the packet n payload is i, the packet n+1 payload is j, and the result of the subtraction is k. Then, k is equal to i – j shall j less than i, otherwise, k is equal to j – i. If k is less than j, then j= k, otherwise, j= j. Third, modify the value of the Contributing Source (CSRC) Count (CC) field in the RTP header of the RTP packet header. The CC value will be used to recover the RTP payload initial format as generated by the codec at the sender side. Modifying the CC field will be discussed below. Finally, associate the new payload (new value of j) to the RTP packet header and dispatch the resulted new RTP packet to the destination. Fig. 2 demystify the internal process of the proposed RPS method at sender side.

As discussed in point three above, the CC field of the RTP header will be modified by the RPS method. The CC field of the RTP filed is only used for certain situations such as group call or video conferencing. It's unused on other situations such VoIP calls which are the concern in this paper. Therefore, the CC field can be reemployed to perform other functions [11]. The proposed RPS method will be using it to recover the initial format of RTP payload at the receiver side. The RPS method requires i) a value to indicate which of the successive RTP payloads is greater (payload i or payload j in the example above) and ii) another value to indicate whether the RTP payload has been replaced with new value (subtraction result) or remain unchanged. Accordingly, each of these two indicators have got two possibilities and, thus, two bits is required. The RPS method will employ only the leftmost two bits of the 4 bits CC field. The first leftmost bit indicates which of the successive RTP payloads is greater (payload i or payload j). The value of one indicates i is greater than j while the value of zero indicates j is greater or equal to i. The second

leftmost bit indicates whether the RTP payload has been replaced with the subtraction result or remain unchanged. The value of one indicates that the RTP payload has been replaced with the subtraction result while the value of zero indicates that the RTP payload remain unchanged. Fig. 3 shows the subtraction algorithm (Second process) and modifying the CC fields of the RTP header (Third process).
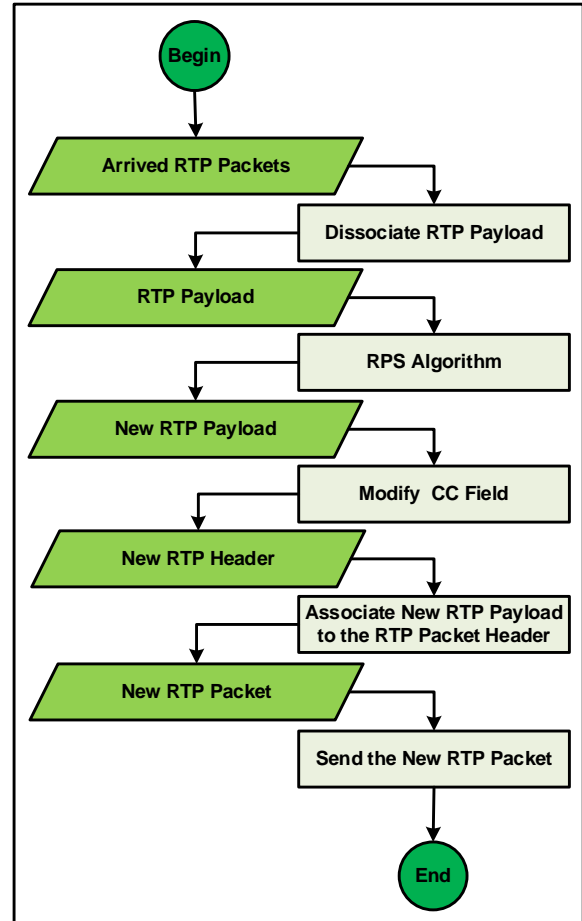
Fig 2.   RPS Method Process (Sender Side).

RPS Method Algorithm → Sender Side

```
1   // I is packet n payload
2   // J is packet n+1 payload
3   // FLC First leftmost CC Field
4   // SLC Second leftmost CC Field

5   if (I > J) {
6       result= I – J
7       FLC = 1
8   else
9       result= J – I
10      FLC = 0
11  }
12  if (result < J) {
13      J= result
14      SLC = 0
15  else
16      J= J
17      SLC = 1
18  }
```

Fig 3.   RPS Method Algorithm (Sender Side).

B. RPS Method Processes at Receiver Side

The receiver of the VoIP call must recover the initial format of RTP payload as it produced by the codec at the sender side. This is will be achieved by the help of the CC fields values, which have been set by the RPS method algorithm at sender side. Initially, the RTP payload and the RTP packet header are split up. Then, the values of the CC field in the RTP packet header are inspected to decide the operation to be performed on the RTP payload. If the second leftmost bit of the CC field is equal to one, then no operation is performed on the RTP payload and it kept as it is. However, if the second leftmost bit of the CC field is equal to zero, then the first leftmost bit of the CC field need to be inspected. The value of zero of the leftmost bit of the CC field causes the two RTP payloads, of the current RTP packet and the previous RTP packet, to be added. The value of zero of the leftmost bit of the CC field causes the RTP payload of the current RTP packet to be subtracted from the RTP payload of the previous RTP packet. Performing any of the last two operations will produce the initial format of RTP payload of the current RTP packet. Before the final step, the CC field values must be erased (set to zero) to avoid any misinterpretation by the receiver. Fig. 4 shows the restoring operation of the RTP payload. Finally, reattach the resulted RTP payload to the RTP packet header. Fig. 5 demystify the internal process of the proposed RPS method at receiver side.

```
RPS Method Algorithm → Receiver Side
 1   // I is packet n payload
 2   // J is packet n+1 payload
 3   // FLC First leftmost CC Field
 4   // SLC Second leftmost CC Field

 5   if (SLC is equal to 1) {
 6       Payload-Size= J
 7   else if (SLC is equal to 0) {
 8     if (FLC is equal to 0) {
 9         Payload-Size= J + I
10     else if (FLC is equal to 1)
11         Payload-Size= I - J
12   }
```

Fig 4.   RPS Method Algorithm (Receiver Side).
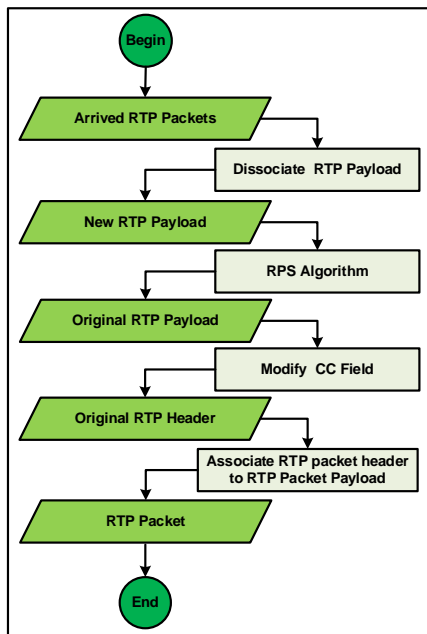
Fig 5.   RPS Method Process (Receiver Side).

IV. RPS METHOD RESULTS AND PERFORMANCE EVALUATION

In this section, the effectiveness of the proposed RPS method will be estimated. The proposed RPS method intended to provide a better consuming of the network bandwidth when the VoIP applications are running. The three measures that are going to be used to estimate the proposed RPS method are number of eliminated bits, saved bandwidth, and the capacity of the RTP flows running at the same time in a specific bandwidth.

A. Eliminated Bits Per RTP Payload

This subsection will show the effectiveness of the proposed RPS method by investigating the number of bits that have been removed from the RTP payload. The number of removed bits have been investigated for 5 different flows. The investigation was performed on a small sample of 7 RTP packets of each flow. The size of the RTP payload after removing the bits (P_New) was compared with the size of the RTP payload before removing the bits (P-Old). Fig. 6 shows the removed bits of RTP payload when using the proposed RPS method. Clearly, the more the number of removed bits, the better the effectiveness of the proposed RPS method. Please note that the first packet of each flow was not affected, and remain with the same size, because there are no packets preceded it. However, the size of the other RTP payloads were reduced by some (different) values. These values are different from RTP payload to another depends on the bits distribution pattern within the RTP payload.

B. Saved Bandwidth

This subsection will show the effectiveness of the proposed RSP method by examining the saved bandwidth ratio. The saved bandwidth ratio was examined for seven different flows. Fig. 7 shows the saved bandwidth ratio of the proposed RPS method. The saved bandwidth ratio has been enhanced by up nearly between 12.1% and 17%. This is, to some extent, a considerable saved bandwidth ratio by the proposed RPS method, which enhances the bandwidth exploitation. The values of the saved bandwidth ratio are different from RTP flow to another because of the different in the bits distribution pattern within the RTP payload.
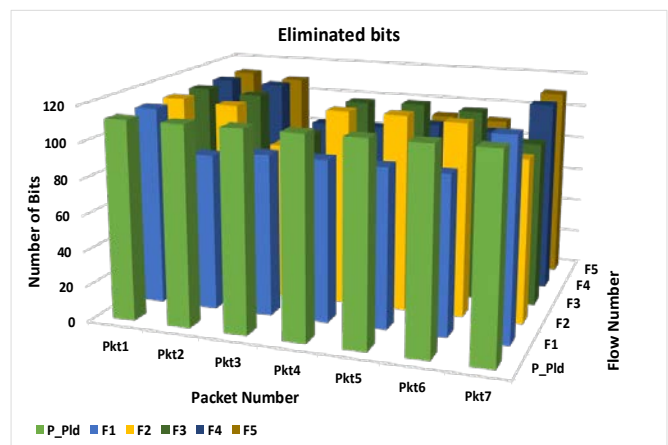
Fig 6.   RTP Packet Payload Eliminated Bits.
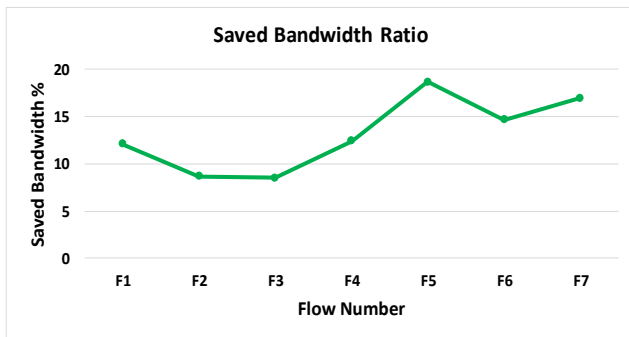
Fig 7.   Saved Bandwidth Ratio.

C. Capacity Improvement

This subsection will show the effectiveness of the proposed RPS method by examining capacity concurrent RTP flows. The capacity improvement ratio was examined with ten different flows. Fig. 8 shows the capacity improvement ratio of the proposed RPS method. The capacity improvement ratio has been enhanced by up nearly between 2.2% and 4.9%. Thus, the bandwidth exploitation is enhanced when using the proposed RPS method. Again, the values of capacity improvement ratio are different from RTP flow to another because of the different in the bits distribution pattern within the RTP payload.
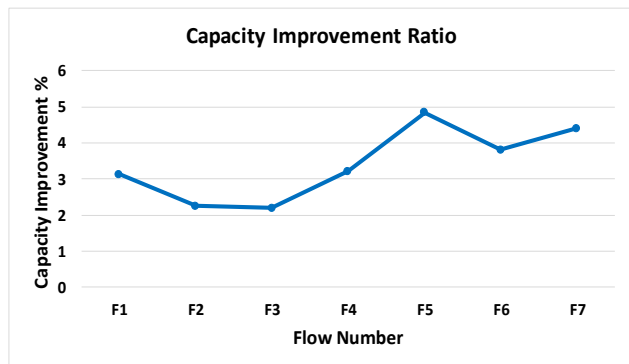
Fig 8.   Capacity Improvement Ratio.

V. CONCLUSION

Telecommunication system is one of the key pillars for businesses evolution. Due to its various features, VoIP technology is dominating the business telecommunication nowadays. The developers are exerting high efforts to enhance network bandwidth exploitation when using VoIP technology. This paper has contributed to these efforts by proposing a new RTP packet's payload shrinking method, called RPS that reduces the RTP packet's payload size. The RPS method's algorithm reduces the RTP packet's payload size by calculating and transmitting the different between the successive RTP packet's payloads, instead of the full RTP packet's payload. The RPS method's algorithm utilizes the RTP fields to store the values that are needed to apply the shrinking algorithm at the sender and receiver sides. The effectiveness of the proposed RPS method has been examined in comparison to conventional RTP protocol without shrinking. The deployment result showed that the saved bandwidth ratio has reached up to nearly 17% in the tested scenarios. Therefore, enhancing the network bandwidth exploitation. The future work will deploy the proposed RSP method with other network bandwidth exploitation methods, such as VoIP packets multiplexing methods.

REFERENCES

[1] D. Haryono, W. Sudiharto and A. G. Putrada, "QoS Improvement Analysis of VoIP Service Which Uses Overlay Network. Case Study: Calling AWS VoIP Gateway From Bandung, Indonesia," 2018 International Seminar on Application for Technology of Information and Communication, Semarang,, pp. 381-387, doi: 10.1109/ISEMANTIC.2018.8549807.

[2] H.M. Mahdi, A.M. Suhail, A. Maaruf, H. H. AbdelRahman, and A.G. Muzafar, "Simulation and Analysis of Quality of Service (QoS) Parameters of Voice over IP (VoIP) Traffic through Heterogeneous Networks", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 8, no. 7, 2017.

[3] A. Gupta, and A. Chaudhary, "A metaheuristic method to hide MP3 sound in JPEG image," Neural Computing and Applications, vol. 30(5), Septemper 2018, pp. 1611-16181.

[4] M. Abualhaj, M. Al-tahrawi, and S. Al-khatib. "Performance Evaluation of VoIP Systems In Cloud Computing." Journal of Engineering Science and Technology 14.3 (2019): 1398-1405.

[5] M. Abualhaj, S. Al-Khatib, Q. Shambour, " PS-PC: An Effective Method to Improve VoIP Technology Bandwidth Utilization over ITTP Protocol", Cybernetics And Information Technologies, 20.3 (2020): 147-158.

[6] M. M. Abualhaj, M. M. Al-Tahrawi, and S. N. Al-Khatib, "A New Method to Improve Voice over IP (VoIP) Bandwidth Utilization over Internet Telephony Transport Protocol (ITTP)," Proceedings of the 2019 8th International Conference on Software and Information Engineering, pp. 192-195, April 2019.

[7] Aji, R. F., et al. "Throughput Fairness Using Packet Aggregation on 802.11 g Networks." (2016).

[8] M. Tomoskozi, P. Seeling, P. Ekler, & F. Fitzek,. "Performance evaluation and implementation of IP and robust header compression schemes for TCP and UDP traffic in static and dynamic wireless contexts". (2017) Comput. Sci. Inf. Syst., 14, 283-308.

[9] M. Abualhaj, "CA-ITTP: An Efficient Method To Aggregate VoIP Packets Over ITTP Protocol.". International Journal of Innovative Computing Information and Control 15.3 (2019): 1067–1077.

[10] Q. Shambour, et al, "Effective Voice Frame Shrinking Method to Enhance VoIP Bandwidth Exploitation" International Journal of Advanced Computer Science and Applications(IJACSA), 11(7), 2020.

[11] Abualhaj, Mosleh M., et al. "Effective Voice Frame Pruning Method to Increase VoIP Call Capacity." TEM Journal 9.1 (2020): 48-54.

[12] Hartpence, B. (2013). Packet Guide to Voice over IP: A system administrator's guide to VoIP technologies. O'Reilly Media, Inc.

[13] K. Sandlund, G. Pelletier, and LE. Jonsson, "The robust header compression (rohc) framework," RFC 5795, March 2010.

[14] A. Majeed, and N. Abu-Ghazaleh. "Packet aggregation in multi-rate wireless LANs." 2012 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON). IEEE, 2012.

[15] Sze, H. P., Liew, S. C., Lee, J. Y., & Yip, D. C. (2002). A multiplexing scheme for H.323 voice-over-IP applications. IEEE Journal on Selected Areas in Communications, 20(7), 1360-1368.

[16] M. M. Abu-Alhaj, M. S. Kolhar, L. V. Chandra, O. Abouabdalla, and A. M. Manasrah, "Delta-multiplexing: A novel technique to improve VoIP bandwidth utilization between VoIP gateways," 2010 10th IEEE International Conference on Computer and Information Technology, pp. 329-335, June 2010.

Arabic Handwritten Character Recognition based on Convolution Neural Networks and Support Vector Machine

Mahmoud Shams[1]

Department of Machine Learning
and Information Retrieval, Faculty of
Artificial Intelligence, Kafrelsheikh
University
Kafrelsheikh, Egypt, 33511

Amira. A. Elsonbaty[2]

Department of electronic and
Communication
Higher Institute of Engineering and
Technology in New Damietta, Egypt.

Wael. Z. ElSawy[3]

Department of Business Information
System
Faculty of Management Technology
& Information System, PortSaid
University, Egypt, 42511

Abstract—**Recognition of Arabic characters is essential for natural language processing and computer vision fields. The need to recognize and classify the handwritten Arabic letters and characters are essentially required. In this paper, we present an algorithm for recognizing Arabic letters and characters based on using deep convolution neural networks (DCNN) and support vector machine (SVM). This paper addresses the problem of recognizing the Arabic handwritten characters by determining the similarity between the input templates and the pre-stored templates using both fully connected DCNN and dropout SVM. Furthermore, this paper determines the correct classification rate (CRR) depends on the accuracy of the corrected classified templates, of the recognized handwritten Arabic characters. Moreover, we determine the error classification rate (ECR). The experimental results of this work indicate the ability of the proposed algorithm to recognize, identify, and verify the input handwritten Arabic characters. Furthermore, the proposed system determines similar Arabic characters using a clustering algorithm based on the K-means clustering approach to handle the problem of multi-stroke in Arabic characters. The comparative evaluation is stated and the system accuracy reached 95.07% CRR with 4.93% ECR compared with the state of the art.**

Keywords—Handwritten Arabic recognition; convolutional neural networks; support vector machine

I. INTRODUCTION

In recent times, Arabic letter recognition has become required in many applications especially in this epidemic, as traditional education is being transferred to digital education that may need distance learning via the Internet for children or people who need to learn Arabic. Moreover, it is very difficult to recognize Arabic handwritten language especially to evaluate the correct way for writing Arabic characters for undergraduate or postgraduate students, making scientific research is demanded. Therefore, in this paper, we present an Arabic handwritten character recognition system to be one of the effective keys for Arabic handwritten recognition.

Deep learning techniques are effective tools for feature extraction and classification in computer vision and pattern recognition fields. In this paper, DCNN is proposed to classify and to extract the features of the input images represent the handwritten Arabic characters [1-3].

SVM is an efficient supervised statistical classifier used to classify the input patterns separating the nearest data points using hyperplanes [4-6].

The main contribution of this paper can be summarized as follows:

1) Building an architecture based on DCNN to classify and recognize the handwritten Arabic characters using a fully connected layer.

2) Using Drop out SVM with the softmax layer to proceed with the input features in the existence of DCNN to boost and compare the result with the fully connected DCNN.

3) Handling the problem of multi-stroke similar Arabic handwritten characters.

4) A comparison study is performed to evaluate the proposed algorithm with the state of the most recent approaches.

The rest of this paper is organized as follows: Section II illustrates the literature survey of the related work; the proposed algorithm is investigated in Section III; the results and discussion is demonstrated in Section IV; finally, the conclusion and future work are presented in Section V.

II. RELATED WORK

Recognition of letters and characters are essential in text mining and information retrieval system especially search engine. In the present, there are many attempts to recognize, classify, verify, and identify Arabic handwritten characters using machine learning (ML) approaches.

Arabic handwritten recognition based on SVM and radial basis function (RBF) is presented by Elleuch et al. [7], the error classification rate (ECR) of their proposed algorithm was 11.23%. Moreover in [8] a deep learning approach based on SVM they named as DSVM is composed of three SVMs after using dropout and they achieved ECR 5.86%.

A CNN approach based on the SVM classifier is further utilized by Elleuch [9] to boost the extracted features from Arabic handwritten and the recognition accuracy reached to 97.35% and 93.41% using 24 and 66 classes of HACDB database respectively. Furthermore, a comparative study of the methods depends on CNN and SVM are listed in [10] that indicated that the ECR achieve results 5.83%, and 2.09% using the HACDB database of 66 and 24 classes respectively.

El-Sawy [11] et al presented an algorithm based on CNN for handwritten Arabic digits. They used 10 classes based on fully connected CNN and the accuracy obtained reached 88% with ECR=12%. Moreover in [12], they used CNN for Arabic handwritten character recognition. Their ECR is 5.1% using the ReLU activation function that calculates the maximum values of non-saturated feature maps of the input images. They used fully connected 28 classes CNN with classification accuracy 94.4%.

Younis [13] present DCNN to recognize Arabic handwritten characters and they achieve 94.8% and 97.6% for AIA9k and the AHCD databases, respectively.

Hassan et al. [14] presented an algorithm based on scale-invariant feature transform (SIFT) and SVM for offline handwritten Arabic recognition applied to the AHDB database with a promising recognition rate reached to 99.08%.

For isolated Arabic character recognition, Huque et al. [15] presented a comparative study of using K nearest neighbors (KNN), SVM, and sparse representation classifier (SRC) to recognize Arabic handwritten characters base on feature fusion.

The major problem in Arabic handwritten recognition is the multi-stroke problem that stated there is a great confusion with similar character stroke i.e. there are some may be difficult to recognize such as "ط" and "ظ" the point "." Is the major

difference as presented in [16]. Moreover, a hierarchal clustering of the most common stroke characters is presented by Boudelaa [17] et al. A detailed comparative study of using DCNN in Arabic handwritten character recognition is presented by Ghanim et al [18]. They compared their approach using Alex net DCNN with a recognition rate of 95.6% with a dynamic Bayesian network (DBN) and SVM. Moreover, Mustafa et al [19] a deep learning approach is presented to classify and recognize Arabic names, and the accuracy reached 99.14% in the SUST-ARG-names dataset.

The major limitations of the stated woks are the inability to recognize multi-stroke Arabic characters. As well as clustering the similar characters for unsupervised learning methodologies.

In this paper, we present an algorithm based on using DCNN with a fully connected layer and softmax and dropout the input features to be classified using SVM. We can use a k-means clustering algorithm to handle the multi-stroke Arabic characters recognition.

III. PROPOSED METHODOLOGY

A. Deep Convolutional Neural Network

CNN is one of the most common and wide range of tools in computer vision and pattern recognition. In this paper, we present DCNN to extract the features of the input image that represent the Arabic handwritten characters. Furthermore, a fully connected layer consists of 28 classes represent the Arabic characters is applied. We used 3 convolutional layers called C1, C2, C3. The applied input images are normalized to 64×64. Then the mask of CNN is 7×7 is applied to the 64 × 64 images to produce a convoluted layer that is pooling in a maximum manner that is called max-pooling. Every convolution layer has 32 mini-batch based on a mini-batch gradient descent optimizer.
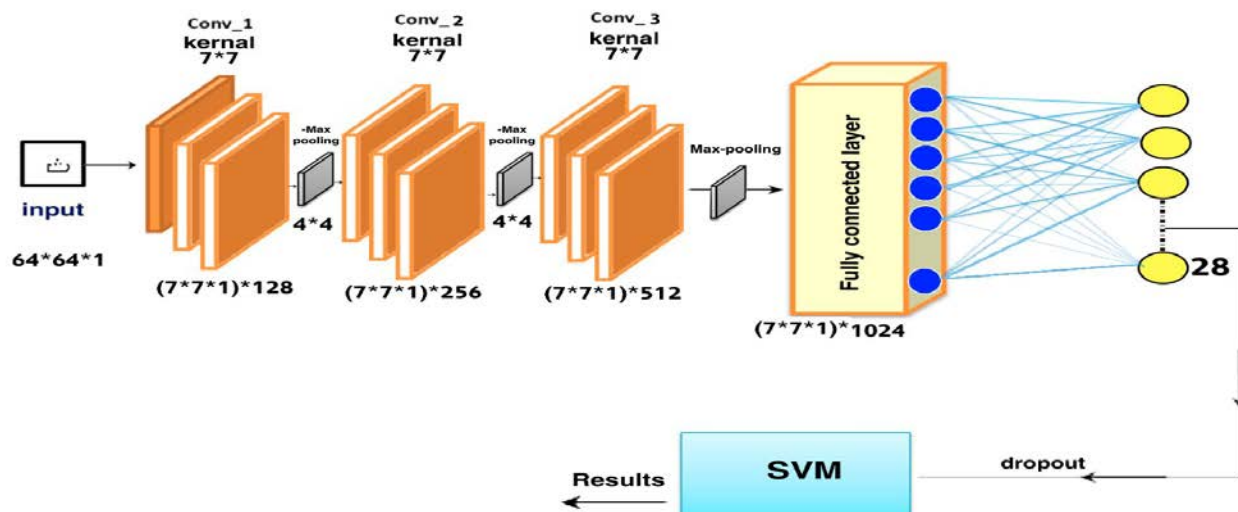
Fig 1. The Proposed DCNN-SVM Architecture.

To ensure the ability of the proposed DCNN to extract more features details, we present three convolutional layers and apply the ReLU activation function. The max-pooling size 4 × 4 is applied after each convoluted layer. Afterward, a fully connected layer that represents 28 classes that are used to classify the input pattern represents the Arabic character.

We present SVM as a statistical kernel classifier after the soft-max drop-out to boost the results of the fully connected layer. Fig. 1 illustrates the proposed DCNN-SVM architecture that investigates all the steps required to recognize and classify the Arabic handwritten characters.

B. Support Vector Machine

In this paper, to boost the results obtained from DCNN to classify the input patterns with 28 classes which represent the Arabic handwritten characters we present SVM with multi-class labels. SVM can classify the dropout fully connected DCNN to 28 classes that represent all cases of Arabic characters that will be illustrated in detail in the results and discussion section. Table I discuss in detail the architecture of the proposed DCNN-SVM architecture.

This paper proposed an algorithm used to handle the problem of multi-stroke of Arabic handwritten characteristics by using the same protocol investigated in [16]. Moreover, the clustering of similar Arabic characters is performed using the K-means clustering algorithm by quantizing the vectors to 13 clusters that represent the masterstroke and similar characters. The clustering procedure based on k-means is based on the same protocol presented by Ntalianis et al [20]. Table II investigates the 13 clusters that represent Arabic character masterstroke, similar characters, handwritten characters, and English pronunciation for each character.

TABLE I. THE CHARACTERISTICS OF THE PROPOSED DCNN-SVM ARCHITECTURE

Layer Name.	DCNN-SVM characteristics	
	Input Size	*Kernel Size*
Input	64×64×1	7×7
Conv-1	64×64×128	7×7
Max-Pooling	4×4	-
Conv-2	32×32×256	7×7
Max-Pooling	4×4	-
Conv-3	16×16×512	7×7
Max-Pooling	4×4	-
Fully-Connected	1×1×1024	-

TABLE II. THE ARABIC CHARACTERS MASTERSTROKE AND THEIR SIMILAR HANDWRITTEN CHARACTERS

Arabic Character Master Stroke	Similar Characters	Handwritten Characters	English Pronunciation
أ	ا-أ-إ		Alef
ب	ب ت ت ث ث ن ي-ى		Baa Taa Thaa Noon Yaa
ج	ج ح خ		Gem Haa Khaa
د	د – ذ		Dal Zal
ر	ر-ز-و		Raa Zeen Waw

Arabic Character Master Stroke	Similar Characters	Handwritten Characters	English Pronunciation
س	س		Seen
	ش		Sheen
ص	ص		Saad
	ض		Daad
ط	ط		Taaa
	ظ		Zaaa
ع	ع		Aeen
	غ		Gheen
ف	ف		Faa
	ق		Kaaaf
ك	كـ		Kaf
	ل		Lam
م	م		Mem
ه	هــه		Haa

IV. Results And Discussion

In this section, the experimental results were performed in core i7 NVidia 4G-GT 740m GPU environment based on MATLAB 2020 a. The dataset used is founded in [12] with the following characteristics listed in Table III. Furthermore, we applied 840 tested images from 3 persons each person write each character 10 times. A sample of database collected and compared in this paper is shown in Fig. 2.

TABLE III.   DATABASE CHARACTERISTICS

No. of Characters	16800
No. of participants	60
No. of writing each character	10 per participant from "Alef" to "Yaa"
No. of the training set (80%)	13440 (480 images per class)
No. of the testing set (20%)	3360 (120 images per class)
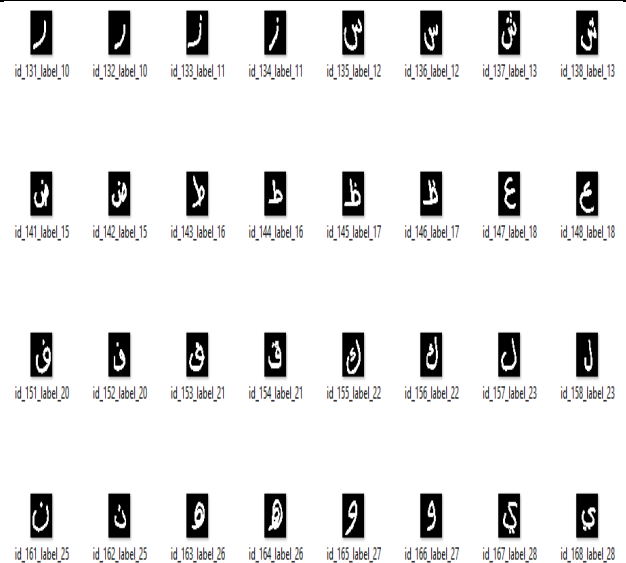
Fig 2.   Samples of the Arabic Handwritten Characters [12].

TABLE IV. THE HYPERPARAMETER VALUES OF THE PROPOSED DCNN-SVM ARCHITECTURE

Parameter	Value
Maximum No. of Iterations	400
Learning Rate	0.02
Weight decay	0.001
Momentum	0.8
Batch Size	32

The experimental results were executed based on the following fixed hyper-parameter values listed in Table IV as follows:

The above table investigated the maximum number of iteration of the DCNN is 400 with learning rare 0.02 and weighted decay 0.001 and momentum 0.8 with mini-batch gradient descent optimize with batch size 32.

C. Experiment 1

Based on the hyper-parameter values listed in Table IV, in this experiment we train and test the DCNN-SVM architecture for Arabic handwritten characters using the database in [12] to determine the classification accuracy and the ECR of the proposed system. After 400 iterations the final accuracy of the trained images was 98.08% and 95.07% for the tested images as shown in Fig. 3.

The ECR is determined by as (100 – corrected classified patterns). In our case, the ECR is 1.92% and 4.93% in training and testing, respectively.
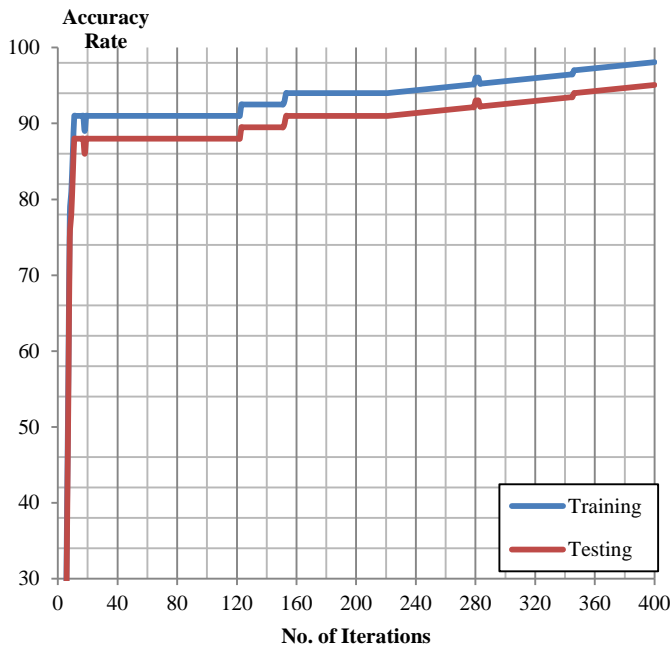
Fig 3. The Accuracy Rate of the Proposed DCNN-SVM Architecture.

TABLE V. AVERAGE CORRECT AND ERROR CLASSIFICATION RATE OF THE TESTED HANDWRITTEN ARABIC CHARACTERS

Handwritten Characters	Correct Classification Rate (%)	Error Classification Rate (%)
أ	95.25	4.75
ب	95.21	3.79
ج	97.20	2.8
د	97.23	2.77
ر	93.21	6.79
س	92.54	7.46
ص	94.02	4.98
ط	90.25	9.75
ع	95.21	4.79
ف	96.12	3.88
ك	94.25	5.75
م	98.24	1.76
Average	95.07	4.93

To prove the ability of the proposed system to recognize Arabic handwritten characters, we tested 840 images out of three persons i.e. 3×28×10 that are collected, cropped, and segmented and the evaluation results are shown in Table V of the tested images.

D. Experiment 2: Comparitive Evaluation

The comparative study of the proposed architecture is performed using the same database in [12]. Table VI shows the comparative evaluation of the proposed system compared with [12] as follows:

TABLE VI. COMPARATIVE EVALUATION OF THE CRR AND ECR OF THE PROPOSED DCNN-SVM AND LOEY ET AL. [12]

Author	CRR	ECR
Loey et al [12]	94.90%	5.10%
Proposed DCNN-SVM	95.07%	4.93%

A rough comparison between the proposed architecture with the state-of-the-art is investigated in Table VII by which we compare the CRR, and ECR concerning the database used, and the trained, tested data.

TABLE VII.   COMPARATIVE EVALUATION OF THE PROPOSED DCNN-SVM WITH THE STATE OF THE ART SYSTEM

Author	Database used	No. of Trained /Tested images	CCR (%)	ECR (%)
Elleuch [9-10]	HACDB images	6600 images	97.35% 24 class 93.41% 66 class	5.83% 24 class, 2.09% 66 class
Loey et al. [12]	Arabic Handwritten Characters Dataset	16800 images 13440 Training 3360 Testing images	94.90%	5.10%
Proposed DCNN-SVM	[12] and 840 additional images	16800 images 13440 Training 3360 Testing images 840 tested images	95.07%	4.93%

V. CONCLUSION AND FUTURE WORK

This paper presents an effective deep convolutional neural network architecture used to extract and classify the Arabic handwritten characters. To ensure the reliability and efficiency of the proposed architecture, we apply a dropout support vector machine to classify and recognize the missing features that are not correctly classified by DCNN. Moreover, the proposed system divided the multi-stroke with similar Arabic characters to 13 clusters depends on K-means clustering. The proposed system achieves 95.07% correct classification accuracy with a minimum error classification rate of 4.93% compared with recent approaches. In the future, we plan to use full Arabic sentences that is maybe helpful for recognize and classify Arabic sentences as it considered one of the most challenges in the computer vision field.

REFERENCES

[1] Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson. "How transferable are features in deep neural networks?." In *Advances in neural information processing systems*, pp. 3320-3328. 2014.

[2] Ciregan, Dan, Ueli Meier, and Jürgen Schmidhuber. "Multi-column deep neural networks for image classification." In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3642-3649. IEEE, 2012.

[3] Szegedy, Christian, Alexander Toshev, and Dumitru Erhan. "Deep neural networks for object detection." In *Advances in neural information processing systems*, pp. 2553-2561. 2013.

[4] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." *Neural processing letters* 9, no. 3, 293-300, 1999.

[5] Cauwenberghs, Gert, and Tomaso Poggio. "Incremental and decremental support vector machine learning." In *Advances in neural information processing systems*, pp. 409-415. 2001.

[6] Noble, William S. "What is a support vector machine?." *Nature biotechnology* 24, no. 12, 1565-1567, 2006..

[7] Elleuch, Mohamed, Houssem Lahiani, and Monji Kherallah. "Recognizing Arabic handwritten script using support vector machine classifier." In *2015 15th International conference on intelligent systems design and applications (ISDA)*, pp. 551-556. IEEE, 2015.

[8] Elleuch, Mohamed, Raouia Mokni, and Monji Kherallah. "Offline Arabic Handwritten recognition system with dropout applied in Deep networks based-SVMs." In *2016 International joint conference on neural networks (IJCNN)*, pp. 3241-3248. IEEE, 2016.

[9] Elleuch, Mohamed, Najiba Tagougui, and Monji Kherallah. "A novel architecture of CNN based on SVM classifier for recognising Arabic handwritten script." *International Journal of Intelligent Systems Technologies and Applications* 15, no. 4, 323-340, 2016.

[10] Elleuch, Mohamed, Rania Maalej, and Monji Kherallah. "A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition." *Procedia Computer Science* 80, 1712-1723, 2016.

[11] El-Sawy, Ahmed, EL-Bakry Hazem, and Mohamed Loey. "CNN for handwritten arabic digits recognition based on LeNet-5." In *International conference on advanced intelligent systems and informatics*, pp. 566-575. Springer, Cham, 2016.

[12] El-Sawy, Ahmed, Mohamed Loey, and Hazem El-Bakry. "Arabic handwritten characters recognition using convolutional neural network." *WSEAS Transactions on Computer Research* 5, 11-19, 2017.

[13] Younis, Khaled S. "Arabic handwritten character recognition based on deep convolutional neural networks." *Jordanian Journal of Computers and Information Technology (JJCIT)* 3, no. 3, 186-200, 2017.

[14] Hassan, Alia Karim Abdul, Bashar Saadoon Mahdi, and Asmaa Abdullah Mohammed. "Arabic handwriting word recognition based on scale invariant feature transform and support vector machine." Iraqi Journal of Science, 381-387, 2019.

[15] Huque, Abu Sayeed Ahsanul, et al. "Comparative Study of KNN, SVM and SR Classifiers in Recognizing Arabic Handwritten Characters Employing Feature Fusion." Signal and Image Processing Letters 1.2, 1-9, 2019.

[16] Valikhani, Sara, Fardin Abdali-Mohammadi, and Abdolhossein Fathi. "Online continuous multi-stroke Persian/Arabic character recognition by novel spatio-temporal features for digitizer pen devices." *Neural Computing and Applications*, 1-20, 2019.

[17] Boudelaa, Sami, Manuel Perea, and Manuel Carreiras. "Matrices of the frequency and similarity of Arabic letters and allographs." *Behavior Research Methods,* 1-13, 2020.

[18] Ghanim, Taraggy M., Mahmoud I. Khalil, and Hazem M. Abbas. "Comparative Study on Deep Convolution Neural Networks DCNN-Based Offline Arabic Handwriting Recognition." *IEEE Access* 8 95465-95482, 2020.

[19] Mustafa, Mohamed Elhafiz, and Murtada Khalafallah Elbashir. "A Deep Learning Approach for Handwritten Arabic Names Recognition.", International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020.

[20] Ntalianis, Vaggelis, Nikos Dimitris Fakotakis, Stavros Nousias, Aris S. Lalos, Michael Birbas, Evangelia I. Zacharaki, and Konstantinos Moustakas. "Deep CNN Sparse Coding for Real Time Inhaler Sounds Classification." *Sensors* 20, no. 8, 2363, 2020.

Noise Reduction on Bracketed Images for High Dynamic Range Imaging

Seong-O Shim

Department of Computer & Network Engineering
College of Computer Science and Engineering
University of Jeddah
Jeddah, Saudi Arabia

Abstract—The quality of a high dynamic range (HDR) image produced from bracketed images taken at different camera exposure times can be degraded by noise contained in bracketed images. In this paper, we propose a noise reduction method on bracketed images based on exposure time ratio. First, for each pixel pair of a same scene point lying on two different images, the ratio of their intensity values is compared with the ratio of exposure times of the images on which the pixels are lying. If the compared ratios are close, these two pixels are included in noise-free pixels set. The complement of noise-free pixels set is defined as noisy pixels set. Then, the intensity value of each pixel in noisy pixels set is corrected by its expected value computed from noise-free pixel of the same scene point lying on another image. Experimental results show that all the noisy intensity values can be correctly restored from noise-free pixels except the saturated pixels. Denoising results by PSNR show that the proposed method outperforms the other recent denoising methods such as based-on pixel density filter (BPDF), noise adaptive fuzzy switching median filter (NAFSMF), and adaptive Riesz mean filter (ARmF).

Keywords—Image denoising; high dynamic range imaging; noise detection; noise removal; image restoration

I. Introduction

The human visual system (HVS) can perceive higher dynamic range (i.e., the ratio between the maximum and minimum intensity values in an image) than most of the cameras. This causes the gap between the visual quality of image captured from standard camera and the quality of actual scene perceived by human eye. One way to reduce this gap is using exposure bracketing technique [1]-[3] which is the process to achieve high dynamic range (HDR) image from merging multiple low dynamic range (LDR) images, which is called bracketed images, acquired by gradually increasing the camera exposure settings.

A typical problem in constructing HDR image is that if there are moving objects while capturing a series of bracketed images, they appear blended in merged HDR image which is called as ghosting artefacts. Thus, a lot of research has been conducted to remove ghosting artefacts in HDR imaging [4]-[9]. Like the standard images, the images acquired in bracketing process suffer from noise that usually occurs during acquisition. And this noise can degrade the quality of the final constructed HDR image. Therefore, the denoising of bracketed

images is one of the most important processes to obtain high quality HDR image. Since each bracketed image can be denoised using existing image denoising methods [10]-[14], more attention of research in HDR imaging has been focused on removing ghosting artefacts than image denoising. In this article, we propose a method to exploit information during exposure bracketing process which can be used to reduce noise in bracketed images.

Most of the exposure bracketing techniques use three images: slightly underexposed image, correctly exposed image and slightly overexposed image. Thus, we used three bracketed images in this work. It is found that the exposure time ratio between two differently exposed images is constant and the value is equivalent to the ratio between intensity values of the two corresponding pixels [15]. Therefore, if the ratio of the two corresponding pixels' intensity values is close to the exposure time ratio, these two pixels can be assumed to be noise-free. In this way, all the pixels in the bracketed images are checked whether they are noise-free, and the classified noise-free pixels are grouped into the noise-free pixels set. The noisy pixels set is made by taking complement of the noise-free pixels set. Then, the intensity values of the pixels in noisy pixels set are corrected by their expected values computed from noise-free pixels of the same scene point lying on another image. Experimental results on several datasets illustrate that the proposed method can correctly restore intensity value of any noisy pixel when there exists corresponding noisy-free pixel of the same scene point in another image. However, noisy intensity values of the saturated pixels that have maximum intensity values cannot be correctly restored because once an intensity value reaches a maximum value, it does not increase as the sensor exposure of a pixel increases.

The rest of this paper is organized as follows. Section II presents the proposed algorithm of based-on gamma-corrected exposure time ratio filter (BGEF). In Section III, we present experimental analysis on different datasets. Conclusions are given in Section IV.

II. Proposed Algorithm

Through the article, we let $L = \left[l_{ij}\right]_{r \times c}$ be an image where l_{ij} is the measured pixel intensity value at location (i, j), r is the number of pixels in a row of L, c is the number of pixels in a column of L. The total quantity of light accumulated at

camera sensor location (i, j) for Δt units of time produces a sensor exposure $e_{ij} \cdot \Delta t$, where e_{ij} is sensor irradiation at (i, j).

The measured pixel intensity value l_{ij} at location (i, j) is proportional to the sensor exposure $e_{ij} \cdot \Delta t$ scaled by the camera response function f as $f(e_{ij} \cdot \Delta t)$. Among various methods to estimate f [16], [17], gamma curve approximation is widely used by many camera manufacturers [18]. Assuming the camera response function to be a gamma curve, the pixel intensity at position (i, j) can be expressed as

$$l_{ij} = f(e_{ij} \cdot \Delta t) \approx (e_{ij} \cdot \Delta t)^\gamma \qquad (1)$$

For each pixel location (i, j), the gamma-corrected exposure time ratio $\alpha_s^t(i, j)$ between differently exposed images L_s and L_t is defined in [15] as

$$\alpha_s^t(i, j) = \frac{l_{ij}^t}{l_{ij}^s} \approx \left(\frac{e_{ij} \cdot \Delta t_t}{e_{ij} \cdot \Delta t_s}\right)^\gamma = \left(\frac{\Delta t_t}{\Delta t_s}\right)^\gamma, \qquad (2)$$

where l_{ii}^s and l_{ii}^t are pixel intensity values at (i, j) in image L_s and L_t respectively.

It can be observed in (2) that the gamma-corrected exposure time ratio α_s^t is constant for all pixel location (i, j). From (2), we can estimate the pixel intensity value $\hat{l}_{ii}^t$ at (i, j) in image L_t from the observed intensity value l_{ii}^s in image L_s as $\hat{l}_{ii}^t = \alpha_s^t \cdot l_{ii}^s$. Thus, if the estimated pixel intensity value $\hat{l}_{ii}^t$ is different from the observed value l_{ii}^t, one or both of l_{ij}^s and l_{ij}^t can be thought to be corrupted by noise.

A. Selection of Noisy Pixels

We obtain three images L_1, L_2, L_3 of the same scene by gradually increasing the camera exposure time $(\Delta t_1 < \Delta t_2 < \Delta t_3)$ and compute α_1^2, α_2^3 and α_3^1 from (2). For each pixel location (i, j) in each image frame, we compute the estimated pixel intensity value from another image frame as

$$\hat{l}_{ij}^1 = \alpha_3^1 \cdot l_{ij}^3$$

$$\hat{l}_{ij}^2 = \alpha_1^2 \cdot l_{ij}^1$$

$$\hat{l}_{ij}^3 = \alpha_2^3 \cdot l_{ij}^2 \qquad (3)$$

Let p_{ij}^k be pixel at location (i, j) in image L_k, P_n is noisy pixels set, P_f is noise-free pixels set and P_u is collection of all the pixels. P_n and P_f are initialized as empty set. Gamma-corrected exposure time ratios α_1^2, α_2^3 and α_3^1 cannot be corrupted values since they are computed from pre-determined camera parameters (i.e., exposure time and gamma value) as shown in (2). However, pixel intensity values l_{ii}^1, l_{ii}^2 and l_{ii}^3 can be corrupted by various types of noises. Thus, for each pixel p_{ij}^k, we guess from (3) whether the pixel belongs to P_n or P_f as

$$\left|l_{ij}^2 - \hat{l}_{ij}^2\right| < \epsilon \cdot l_{ij}^2 \rightarrow P_f = P_f \cup \{p_{ij}^1, p_{ij}^2\},$$

$$\left|l_{ij}^3 - \hat{l}_{ij}^3\right| < \epsilon \cdot l_{ij}^3 \rightarrow P_f = P_f \cup \{p_{ij}^2, p_{ij}^3\},$$

$$\left|l_{ij}^1 - \hat{l}_{ij}^1\right| < \epsilon \cdot l_{ij}^1 \rightarrow P_f = P_f \cup \{p_{ij}^3, p_{ij}^1\},$$

$$P_n = P_u - P_f, \qquad (4)$$

where ϵ is predefined tolerance and it is set to 0.1 in this work. When the difference between the observed pixel intensity value and the estimated pixel intensity value is small, both of the two pixels (i.e., the pixel whose intensity value is estimated and the pixel from which the other pixel's intensity value is estimated) can be assumed to be noise-free pixels. Finally, the noisy pixels set is constructed by subtracting noise-free pixels from the collection of all pixels.

B. Correction of Noisy Pixels

The intensity values of pixels classified as noisy in previous section can be corrected from the intensity values of noise-free pixels in another image lying on the same location. Equation (4) indicates that each pixel location (i, j), there are three cases: all three pixels are noise-free, two pixels are noise-free and the remaining one pixel is noisy, all three pixels are noisy. The case that two pixels are noisy and the remaining one pixel is noise-free cannot occur in (4). In this method, the noisy pixel in the second case can be corrected. For each pixel location (i, j), if only one value among l_{ii}^1, l_{ij}^2 and l_{ij}^3 is noisy, the noisy intensity value can be corrected as

$$(p_{ij}^1 \in P_n) \wedge (p_{ij}^2 \in P_f) \wedge (p_{ij}^3 \in P_f) \rightarrow l_{ij}^1 = \hat{l}_{ij}^1,$$

$$(p_{ij}^1 \in P_f) \wedge (p_{ij}^2 \in P_n) \wedge (p_{ij}^3 \in P_f) \rightarrow l_{ij}^2 = \hat{l}_{ij}^2,$$

$$(p_{ij}^1 \in P_f) \wedge (p_{ij}^2 \in P_f) \wedge (p_{ij}^3 \in P_n) \rightarrow l_{ij}^3 = \hat{l}_{ij}^3 \qquad (5)$$

In (5), the noisy pixel intensity value is corrected by its estimated value in (3) since the pixel, from which the value is estimated, is noise-free. However, if all three pixels lying on same location are noisy, the noisy intensity values cannot be corrected in this method.

III. EXPERIMENTAL RESULTS

The proposed method was applied on three sample datasets in [6] – 'Yard', 'Shop' and 'Cars' scenes. Each dataset has three images of a scene captured by doubling the camera exposure time. The sample datasets are shown in Fig. 1.

Out of several types of image noise that occurs during image acquisition and transmission, the impulse noise is one of common types of noise which includes salt-and-pepper noise (SPN) [19]-[21] and random-valued impulse noise [22], [23]. SPN is caused by sudden disturbances in the signal and the pixels affected by SPN hold a maximum or a minimum gray value. In experiments, we imposed SPN noises on the bracketed images with various ratios.

(a)

(b)

(c)
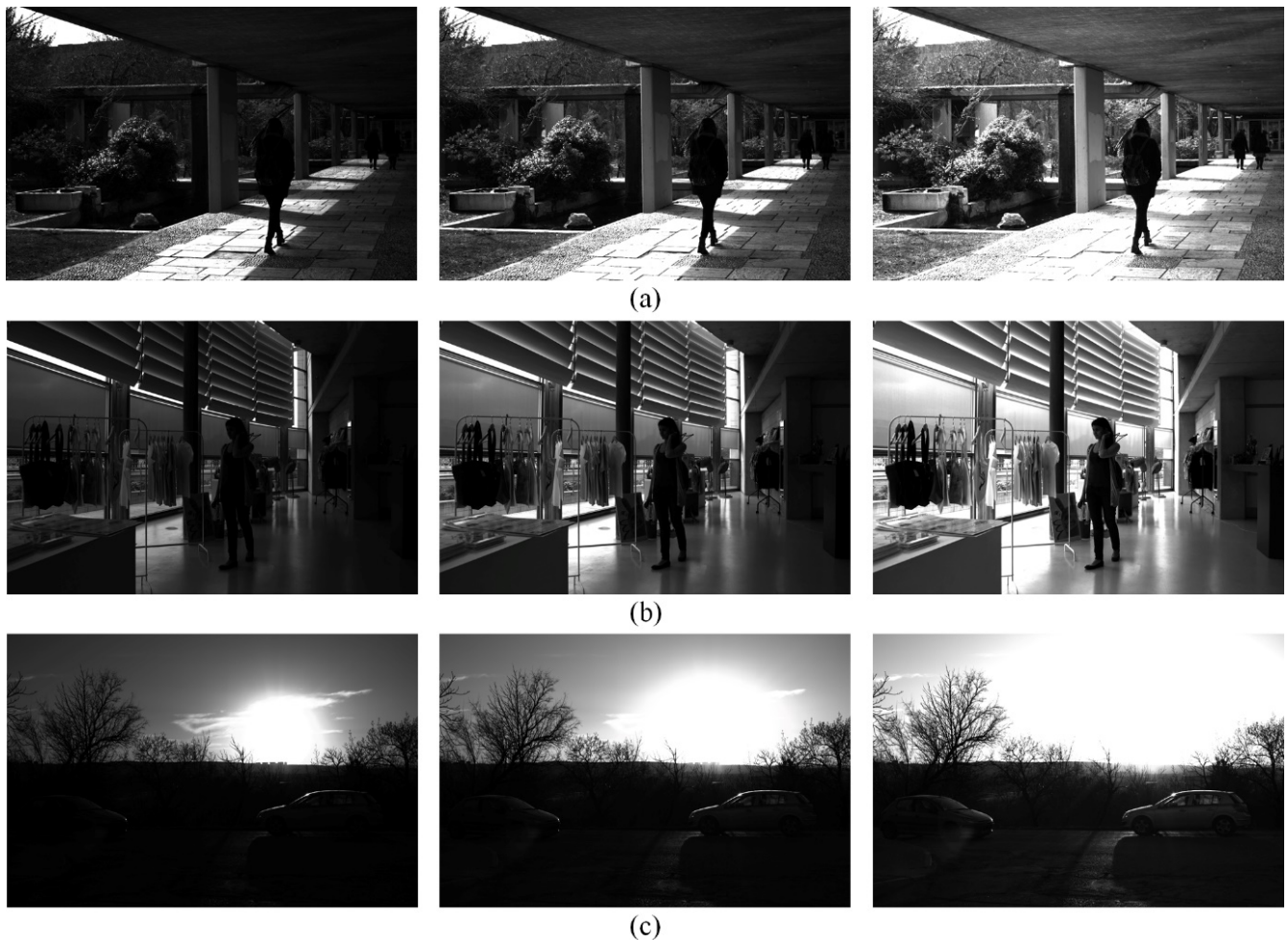
Fig. 1. Three Sample Scenes: (a) Bracketed Images of 'Yard' Scene, (b) Bracketed Images of 'Shop' Scene, (c) Bracketed Images of 'Cars' Scene.

Fig. 2 shows the denoising results of the middle image by the proposed method for 'Yard' scene where the first column shows the corrupted middle image by SPN, the second column shows the denoising results in case only the middle image is corrupted by SPN and the third column shows the denoising results in case all the three images are corrupted by SPN. Denoising for only one image is corrupted shows good results and denoising quality does not decrease much as the noise level increases. The floor parts and the top-left corner of the scene are saturated (full white) in all three images in Fig. 1(a). In these saturated regions, the proposed filter does not work well since the method rely on intensity ratio between images and the intensity ratio on saturated region does not reflect correct exposure time ratio. However, for all three images are corrupted, the denoising quality decreases as the noise level increases. For a certain location (i, j), the proposed method can correct the noisy intensity value when only one pixel is noisy

out of three pixels as in (5). It means, when only one image is corrupted by noise, theoretically 100% of noisy pixels except saturated pixels can be corrected by corresponding two other noise-free pixels lying on the same location in other two images. However, when all the three images are corrupted by α % noise level, the probability of a noisy pixel can be corrected is equivalent to the probability that the corresponding two other pixels are noise-free which is $(1 - \alpha/100)^2 \times 100$ %; thus, only some portion of noisy pixels can be corrected. The same experiments were conducted on 'Shop' and 'Cars' scenes and their denoising results are shown in Fig. 3 and Fig. 4, respectively. In both experiments, we can see the expected results as in Fig. 2. In 'Cars' scene in Fig. 1, the saturated parts around sun in the sky occupy large areas in the image. Thus, the denoising failure around sun is notable in this scene.

Fig. 2. Denoising Results of the Middle Image by the Proposed Method for 'Yard' Scene: (a) Corrupted Images with Different SPN Ratios, (b) Denoising Results in Case Only the Middle Image is Corrupted by SPN, (c) Denoising Results in Case All the Three Images are Corrupted by SPN.

Fig. 3. Denoising Results of the Middle Image by the Proposed Method for 'Shop' Scene: (a) Corrupted Images with Different SPN Ratios, (b) Denoising Results in Case Only the Middle Image is Corrupted by SPN, (c) Denoising Results in Case All the Three Images are Corrupted by SPN.

Fig. 4. Denoising Results of the Middle Image by the Proposed Method for 'Cars' Scene: (a) Corrupted Images with Different SPN Ratios, (b) Denoising Results in Case Only the Middle Image is Corrupted by SPN, (c) Denoising Results in Case All the Three Images are Corrupted by SPN.

For quantitative assessment of the effectiveness of image denoising, we use the common error metric peak signal-to-noise ratio (PSNR) which is defined as

$$PSNR(U,V) = 10\log_{10}\left(\frac{peakval^2}{MSE(U,V)}\right), \quad (6)$$

where *peakval* is the maximum intensity value in image datatype and *MSE* stands for mean square error defined as

$$MSE(U,V) = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(u_{ij} - v_{ij}\right)^2, \quad (7)$$

where $U = \lceil u_{ij} \rceil$ is a ground truth and $V = \lceil v_{ij} \rceil$ is an evaluated (denoised) image. The higher value of PSNR indicates a better image quality.

In Table I, PSNR results of the proposed BGEF method on middle images are compared with recent SPN filtering techniques − based-on pixel density filter (BPDF) [24], noise adaptive fuzzy switching median filter (NAFSMF) [25], adaptive Riesz mean filter (ARmF) [26]. PSNR results show that BGEF performs better than the other filtering methods in all three scenes. Since BGEF reduces noise in bracketed

images using exposure time ratio computed from a priori knowledge, it can be considered as preprocessing technique before applying existing denoising filters. Thus, we also tried to apply BPDF, NAFSMF and ARmF after applying the proposed BGEF, each of which is termed as BGEF-BPDF, BGEF-NAFMSMF and BGEF-ARmF respectively. We can see these combined filters have further improved the results when only one image is corrupted. However, in some cases when all three images are corrupted, the combined filters showed worse performance than single BGEF filter as the noise level increases.

TABLE I. PSNR RESULTS OF THE METHODS FOR THREE SCENES ('YARD', 'SHOP' AND 'CARS') WITH DIFFERENT SPN RATIOS

Scene	Filter	10%	20%	30%	40%	50%	60%	70%	80%	90%
Yard	BPDF	14.7948	11.4381	6.2123	5.6429	5.1229	4.6501	4.1760	3.7424	3.2312
(1 image	NAFSMF	15.0718	12.0522	7.1495	6.0000	4.7246	3.6800	3.0632	2.6904	2.3894
corrupted)	ARmF	15.0571	11.9665	6.8119	6.1399	5.4926	4.8384	4.1750	3.5102	2.8162
Yard	**BGEF**	**21.8402**	**18.7996**	**16.1532**	**13.9080**	**13.4803**	**13.0882**	**12.7175**	**12.4245**	**12.1324**
(1 image	BGEF-BPDF	29.6673	25.5969	22.2504	19.0001	18.2520	17.6507	16.7814	16.2212	15.5347
corrupted)	BGEF-NAFSMF	30.4498	27.2424	24.6925	22.2148	21.7592	21.3514	20.9294	20.4732	20.1004
	BGEF-ARmF	30.9520	27.7618	25.2913	22.9088	22.4732	22.0864	21.6867	21.2919	20.9479
Yard	**BGEF**	**18.9944**	**14.4602**	**8.4810**	**6.7813**	**6.1835**	**5.6645**	**5.2017**	**4.8173**	**4.4624**
(3 images	BGEF-BPDF	21.5826	15.9649	8.3388	6.4079	5.6515	4.9834	4.3708	3.8375	3.2594
corrupted)	BGEF-NAFSMF	21.7394	16.2873	9.3411	7.4919	6.0903	4.4046	3.3017	2.7642	2.4070
	BGEF-ARmF	21.8109	16.3091	9.0471	7.0728	6.1896	5.3225	4.4799	3.6633	2.8587
Shop	BPDF	14.7087	11.3417	9.3041	7.8520	6.7130	5.7847	4.9741	4.2544	3.5631
(1 image	NAFSMF	15.0598	12.0510	10.2880	9.0314	7.9707	6.5051	4.4296	3.1085	2.5226
corrupted)	ARmF	15.0192	11.9364	10.0583	8.6448	7.4472	6.3683	5.3174	4.2662	3.1910
Shop	**BGEF**	**24.9141**	**21.9940**	**20.1777**	**18.9365**	**17.9786**	**17.1791**	**16.5227**	**15.9288**	**15.4317**
(1 image	BGEF-BPDF	31.4590	27.0027	24.3549	22.8316	21.7053	20.8822	20.1632	19.4482	18.8491
corrupted)	BGEF-NAFSMF	35.7117	32.6009	30.8208	29.4552	28.3801	27.2541	25.9238	24.8079	23.9785
	BGEF-ARmF	35.9673	32.8380	31.0408	29.6687	28.5375	27.5648	26.5999	25.7053	24.7991
Shop	**BGEF**	**20.2299**	**15.2058**	**12.1941**	**10.1033**	**8.5269**	**7.2946**	**6.3060**	**5.5085**	**4.8720**
(3 images	BGEF-BPDF	21.8178	15.9985	12.5090	10.0352	8.1471	6.6743	5.4721	4.4840	3.6301
corrupted)	BGEF-NAFSMF	22.1633	16.4435	13.1993	10.9770	9.3180	7.9352	5.7493	3.4285	2.5703
	BGEF-ARmF	22.1693	16.4216	13.1239	10.8032	8.9757	7.4191	5.9865	4.6163	3.2977
Cars	BPDF	14.8187	11.3948	9.2533	7.8058	6.7074	5.7750	5.0027	4.2856	3.2390
(1 image	NAFSMF	15.3301	12.3603	10.5488	9.2893	8.2503	6.7610	4.7317	3.4027	2.5738
corrupted)	ARmF	15.2982	12.2514	10.3253	8.9084	7.7372	6.6436	5.6071	4.5534	2.9151
Cars	**BGEF**	**19.6201**	**16.6098**	**14.7801**	**13.6059**	**12.6231**	**11.8479**	**11.1817**	**10.5849**	**9.8555**
(1 image	BGEF-BPDF	25.2672	20.6474	18.0166	16.6609	15.6628	14.8775	14.3970	13.8015	13.2473
corrupted)	BGEF-NAFSMF	29.8784	26.8546	25.1235	23.8845	22.8178	21.5618	19.9973	18.8440	17.9175
	BGEF-ARmF	30.0175	26.9541	25.1497	23.8159	22.6637	21.6471	20.7404	19.8288	18.4150
Cars	**BGEF**	**17.8410**	**13.6806**	**11.0502**	**9.2383**	**7.8255**	**6.6908**	**5.7810**	**5.0148**	**4.1730**
(3 images	BGEF-BPDF	20.7992	15.3088	11.9637	9.7250	7.9892	6.5736	5.4645	4.4877	3.3172
corrupted)	BGEF-NAFSMF	21.9909	16.5503	13.3413	11.1763	9.5627	8.1487	6.0204	3.7153	2.6201
	BGEF-ARmF	22.0124	16.5366	13.2700	11.0044	9.2214	7.6584	6.2648	4.8930	3.0291

IV. CONCLUSIONS

A noise reduction method based on exposure time ratio on three bracketed images has been proposed. For each location (i, j), the ratio of intensity values of the corresponding pixels between two images is equivalent to gamma-corrected exposure time ratio which can be computed from a priori knowledge of camera parameters. Thus, when a ratio of two corresponding pixels' intensity values is different from gamma-corrected exposure time ratio, one or both of them are classified as noisy pixels. Then, the classified noisy pixels are corrected from noise-free pixels on the same location lying in another image using gamma-corrected exposure time ratio as a cue. Experimental results show that if only one image out of three is corrupted by noise, most of the noisy pixels' intensity values except saturated pixels can be correctly restored by the proposed method regardless of noise level imposed on the image. However, if all the three images are corrupted by noise, the denoising performance decreases as the noise level increases. Experimental results on several datasets using PSNR show that the denoising performance of the proposed method is better than the recent denoising filters BPDF, NAFSMF and ARmF. When the proposed method was used as a pre-processing filter and the existing noise specific denoising filters were applied on the processed images, the denoising performance has further improved.

REFERENCES

[1] S. Im, H. G. Jeon, and I. S. Kweon, "Robust Depth Estimation Using Auto-Exposure Bracketing," IEEE Trans. Image Process., vol. 28, no. 5, pp. 2451–2464, 2018.

[2] A. Leong-Hoi, P. C. Montgomery, B. Serio, P. Twardowski, and W. Uhring, "High-dynamic-range microscope imaging based on exposure bracketing in full-field optical coherence tomography," Opt. Lett., vol. 41, no. 7, pp. 1313–1316, 2016.

[3] T. Liu, H. Liu, Y. Wu, B. Yin, and Z. Wei, "Exposure Bracketing Techniques for Camera Document Image Enhancement," Appl. Sci., vol. 9, no. 21, pp. 4529, 2019.

[4] Q. Yan et al., "Deep hdr imaging via a non-local network," IEEE Trans. Image Process., vol. 29, pp. 4308–4322, 2020.

[5] P. Sen et al., "Robust patch-based HDR reconstruction of dynamic scenes," ACM Trans. Graph., vol. 31, no. 6, pp. 203–1, 2012.

[6] O. T. Tursun, A. O. Akyüz, A. Erdem, and E. Erdem, "The state of the art in HDR deghosting: a survey and evaluation," Comput. Graph. Forum, vol. 34, pp. 683–707, 2015.

[7] S. O. Shim and I. R. Khan, "Removal of ghosting artefacts in HDRI using intensity scaling cue," SIGGRAPH Asia 2017 Technical Briefs, pp. 1–4, 2017.

[8] H. Guo, B. Sheng, P. Li, and C. P. Chen, "Multiview High Dynamic Range Image Synthesis Using Fuzzy Broad Learning System," IEEE Trans. Cybern., pp. 1–13, 2019.

[9] Q. Yan, Y. Zhu, and Y. Zhang, "Robust artifact-free high dynamic range imaging of dynamic scenes," Multimed. Tools. Appl. vol. 78, no. 9, pp. 11487–11505, 2019.

[10] M. Haider, M. U. Riaz, I. Touqir, and A. M. Siddiqui, "Denoising in wavelet domain using probabilistic graphical models," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 11, pp. 317–321, 2016.

[11] X. Yang, Y. Xu, Y. Quan, and H. Ji, "Image Denoising via Sequential Ensemble Learning," IEEE Trans. Image Process. vol. 29, pp. 5038–5049, 2020.

[12] A. M. Elmogy, E. Mahmoud, and F. A. Turki, "Image noise detection and removal based on enhanced gridLOF algorithm," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 12, pp. 454–462, 2017.

[13] C. Tian, Y. Xu, and W. Zuo, "Image denoising using deep CNN with batch renormalization," Neural Netw., vol. 121, pp. 461–473, 2020.

[14] Y. Hou et al., "NLH: A blind pixel-level non-local method for real-world image denoising," IEEE Trans. Image Process., vol. 29, pp. 5121–5135, 2020.

[15] S. O. Shim, "Estimation of gamma-corrected exposure time ratio in multi-exposure images for removal of moving objects," Appl. Opt., vol. 59, no. 13, pp. 4076–4080, 2020.

[16] H. Farid, "Blind Inverse Gamma Correction," IEEE Trans. Image Process., vol. 10, no. 10, pp. 1428–1433, 2001.

[17] M. D. Grossberg and S. K. Nayar, "Determining the Camera Response from Images: What is Knowable?" IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 11, pp. 1455–1467, 2003.

[18] M. D. Grossberg and S. K. Nayar, "What is the Space of Camera Response Functions," IEEE Conf. Computer Vision and Pattern Recognition, USA, vol. 2, pp. II-602, 2003.

[19] L. Chun, S. Bishen, L. Shaohui, T. Kun, and M. Yingrui, "Iterative Removing Salt and Pepper Noise based on Neighbourhood Information," Int. J. Adv. Comput. Sci. Appl., vol. 9(1), pp.261–265, 2018.

[20] M. S. Christo, K. Vasanth, and R. Varatharajan, "A decision based asymmetrically trimmed modified winsorized median filter for the removal of salt and pepper noise in images and videos," Multimed. Tools. Appl., vol. 79, no. 1, pp. 415–432, 2020.

[21] Y. Xing, J. Xu, J. Tan, D. Li, and W. Zha, "Deep CNN for removal of salt and pepper noise," IET Image Process., vol. 13, no. 9, pp.1550–1560, 2019.

[22] A. Roy, L. Manam, and R. H. Laskar, "Region adaptive fuzzy filter: an approach for removal of random-valued impulse noise," IEEE Trans. Ind. Electron., vol. 65, no. 9, pp.7268–7278, 2018.

[23] J. Chen, G. Zhang, S. Xu, and H. Yu, "A blind CNN denoising model for random-valued impulse noise," IEEE Access, vol. 7, pp.124647–124661, 2019.

[24] U. Erkan and L. Gökrem, "A new method based on pixel density in salt and pepper noise removal," Turkish J. Elect. Eng. Comput. Sci., vol. 26, no. 1, pp. 162–171, 2018.

[25] K. K. V. Toh and N. A. M. Isa, "Noise adaptive fuzzy switching median filter for salt-and-pepper noise reduction," IEEE Signal Process. Lett., vol. 17, no. 3, pp. 281–284, 2010.

[26] S. Enginoğlu, U. Erkan, and S. Memiş, "Pixel similarity-based adaptive Riesz mean filter for salt-and-pepper noise removal," Multimed. Tools. Appl., vol. 78, no. 24, pp. 35401–35418, 2019.

Quality in Use of an Android-based Mobile Application for Calculation of Bone Mineral Density with the Standard ISO/IEC 25022

Jose Sulla-Torres[1], Andrea Gutierrez-Quintanilla[2]
Henry Pinto-Rodriguez[3]
Systems Engineering Program
Universidad Católica de Santa María, Arequipa, Perú[1, 2, 3]

Rossana Gómez-Campos[4], Marco Cossio-Bolaños[5]
Universidad Católica del Maule
Talca, Chile[4, 5]

Abstract—One of the most critical bone diseases is osteoporosis, which can be evaluated through measurements of bone mineral density. Though there is a lot of commercialized portable about health in general, few are oriented towards bone health and with a lack of user-friendly interface and data management system. This paper presents a mobile application development for the calculation of the Bone Mineral density, which integrates with Google technology. The Mobile-D methodology for the development of mobile applications, due to the sequentially in the processes or stages is used. Bone mineral density (BMD) was calculated using anthropometric regression equations, and an Android-based Mobile Application with Google technology was developed. By using Firebase Authentication and Firebase Storage provided by Google technology, it allows admin to have full control over database management. In short, this mobile application allows the calculation of the BMD of the students and data storage and data uploading to cloud storage for post-processing, online data management system with user authentication. In addition, the Internacional Organization for Standardization / International Electrotechnical Commission (ISO/IEC) 25022 standard was used to evaluate the quality in use of the mobile app, resulting in 93% of quality in use, this app being able to be used by health professionals for better decision-making.

Keywords—Android mobile application; bone mineral density; firebase; software quality; ISO/IEC 25022

I. INTRODUCTION

One of the most important diseases related to bones is osteoporosis and is characterized by decreased bone mass density. Thus, the bones become more fragile, and they break more easily [1]. The diagnosis of osteoporosis is made by measuring Bone Mineral Density (BMD), recommended by the World Health Organization [2].

In this context, the BMD assessment is decisive to prevent future bone accidents. BMD measurement is generally performed using double energy X-ray absorptiometry (DXA) scanning equipment, but other proposals allow the BMD to be calculated using anthropometric regression equations [3].

Furthermore, mobile devices are intensifying worldwide and have become an indispensable part of people's daily lives. Mobile applications have proven to be very useful for developing automated solutions to health problems [4] and

others associated with proper nutrition to control health conditions [5].

However, of the articles that have been reviewed to date, there are very few works related to bone health and its measurement of BMD, especially in Latin America and oriented towards school children. For this reason, the objective of this article is to develop an Android-based Mobile Application that calculates BMD utilizing anthropometric regression equations of children and adolescents.

For this, the Mobile-D methodology will be followed for the construction of the app; besides, Google's Firebase technology will be used for scalable storage and User Authentication. The validation of the Mobile App will be done using the ISO/IEC 25022 standard to evaluate the quality in use of the mobile app.

The rest of this paper is organized as follows: Section 2 is the background and related work. Section 3 describes the research material and method. Section 4 presents the results and discussion of this research. Finally, this paper is concluded with the future works of this study.

II. RELATED WORK

In this section, we discuss previous work in the development of health-oriented mobile applications, then those that are close to the BMD. We also review works that use Firebase services for all of this based on the use of the Mobile-D Methodology, finally, works that use the ISO / IEC 25022 for the validation of the quality in use.

In the study by Birkhoff and Smeltzer [6], a review was made of the perceptions of health applications focused on the user of mobile devices. The results obtained provided information on the perceptions, experiences, and motivations of users to incorporate health applications for mobile devices in their daily lives when they live with chronic diseases. Likewise, in the study of Baysari and Westbrook [7], they reviewed the human factors that are applied in the design, development, and evaluation of mobile applications developed to favor the issues of coordination of patient-centered care.

Recently Liu et al. [8] present us a mobile application developed on android that allows patient self-diagnosis through the use of a decision tree classification algorithm, this

application provides medical treatments and information on medications, including opinions of various doctors and related symptoms or causes. These functionalities were tested and evaluated, generating consistent results, and proposing the development of a future version for specialists.

About BMD, a study tested a new approach to help women engage in healthy behaviors related to osteoporosis [9]. This new approach includes beliefs, mechanical skills and abilities, and social facilitation delivered via a cell phone app. The percentage of bone density lost over 12 months was lower than expected. In [10] a mobile health application developed to meet the needs of newly diagnosed women with asymptomatic osteoporosis was tested where the use of an mHealth application is experienced after the diagnosis of osteoporosis and whether the application can help them prepare for decision-making about treatment and support them in self-management of osteoporosis. The application's usability was also tested to find out if any adjustments were necessary before deployment.

About using Mobile-D Methodology, Alnanih et al. [11], they propose a design that is based on the Mobile-D approach by dividing the functionalities of the interactive social, academic application into three versions. The output of the mapping is a set of guidelines that guide the designer through the designing of a mobile application in a distracting environment. In the article by Guevara-Vega et al. [12], they automated the inspection process in the economic activities of a municipality through the development of mobile applications using the Mobile-D methodology, they used the MVC (Model, View, Controller) software architecture and the ScriptCase and Android Studio web service tools. The quality of use of the mobile application was evaluated using the ISO / IEC 25022 standard and the USE questionnaire (usability, satisfaction, ease of use). In the case of Mobile-D, its most significant contribution is the model used for its elaboration, which incorporated developers in the methodological research work itself.

Recently a study presented mobile applications for an in-house designed home-based heart screening device that integrates with Google technology [13]. By using Firebase Authentication and Firebase Storage provided b Google technology, it allows admin to have full control over database management. From this article, we have taken the idea of using Firebase Database and Firebase Auth to take advantage of data scalability and security.

About quality standard ISO/IEC 25022 standard [14] defines quality in use measures for the characteristics defined in ISO/IEC 25010. These characteristics are then decomposed into sub characteristics and, finally, into measures. In order to be widely adopted, the ISO/IEC 25010 standard is defined generically, applies to any kind of software system. Scenarios demanding quality evaluation of specific systems, or systems with specific features or user needs, require customization [15].

According to the article of Farinango et al. [16], previous studies have been conducted that have demonstrated the effectiveness of information and communication technologies in supporting healthy lifestyle interventions. In particular, personal health record systems (PHR-S) that enhance self-care,

essential to support changes in lifestyle. The attributes of quality of efficiency, effectiveness, and user satisfaction were evaluated using metrics defined in the ISO / IEC 25022 standard. From this article, we have considered the use of the standard for the evaluation of the Mobile App.

In the article of Idri and Férnandez-Aleman [17], They evaluate, based on the ISO / IEC 25010 quality standard, the quality of the Mobile Personal Health Records (mPHRs) software, which are mobile applications that allow the organization and storage of patient health data to facilitate administration and access. For both patients and doctors. The results were that functional suitability and reliability are the most used quality characteristics by the apps evaluated [18]. Therefore, these indicators are considered essential for evaluating the quality of the software based on ISO / IEC 25022 standards [19].

As described, most jobs vary on Personal Health Mobile app issues. From the studies studied, we saw that few are oriented to the topic of bone health in children and adolescents. Given the importance of bone health in school children, it is for this reason that in this work, we intend to use the development of the Mobile App together with Google technology validated with ISO / IEC 25022 aimed at school bone health that is little discussed in the literature.

III. MATERIAL AND METHOD

To calculate the BMD, data were collected from school children corresponding to anthropometric evaluations carried out on male and female school children between the ages of 6 and 18 from four educational centers in the city of Arequipa-Peru. The study was approved by the local ethics committee (UCSM, 2018). It was developed according to Helsinki for humans.

The methodology used to build the Mobile App was Mobile-D in order to achieve rapid development cycles with a small team to achieve functional products in a short time. The methodology consists of the following Phases: Exploration, Initialization, Production, Stabilization and Testing, which are shown in Fig. 1.
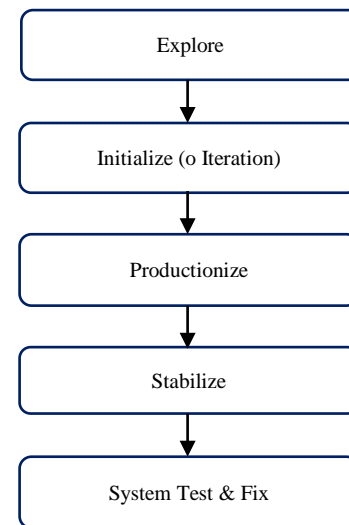
Fig. 1. Phases of the Mobile-D Methodology [20].

A. Explore

The purpose of the exploration phase is to plan and establish the characteristics and concepts of the project. This process is carried out in three stages, shown in Fig. 2.
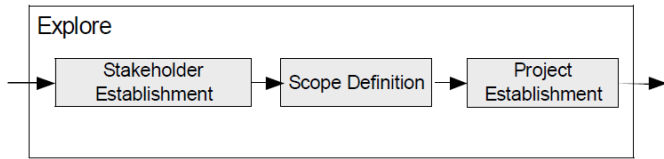
Fig. 2.   Explore Phase [20].

- *Establishment stakeholder*: School children are the objects of direct study, the results of the evaluation are of interest to them, as well as to their parents, just as the teachers want to take control of the evaluations. Health professionals are also interested in the results of the BMD Mobile App.

- *Scope Definition*: The application must allow User authentication to then enter the anthropometric attributes of the students into the application, then they will be processed to calculate the BMD through anthropometric regression equations that allow determining the degree of osteoporosis that a student has or not and to be displayed on a percentile graph, this must be scalable.

- *Project Establishment*. The development of the proposal occurs with the execution between the project team of the University and the Educational Centers for the evaluation of BMD in school children using a mobile application.

B. Initialize

The purpose of this phase is to prepare and identify all the necessary resources. In this phase, it is divided into the following stages, as shown in Fig. 3.
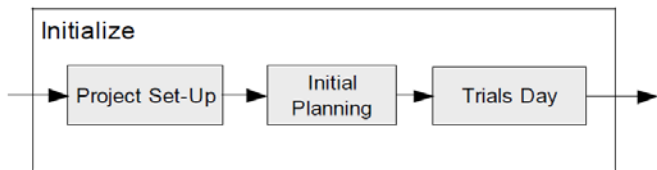
Fig. 3.   Initialize Phase [20].

- *Proyecto set-up*: For the configuration of the environment, we worked with the Android Studio 4.0 development environment, with Firebase Authentication and for data storage with the Firebase Realtime Database of Google technology.

The input and output attributes for the Mobile App were collected from school children and displayed in Table I, Data Dictionary.

- *Initial Planning*: The initial requirements were specified according to the IEEE 830 standard. The software architecture integrated into the system is based on the layered architecture; It has three main components: mobile interface, application server, and database

server. The recommendations of the ISO / IEEE 42010 standard on the Architecture description were used as a guide.

- *Trials day*: Here, plans were made to verify the readiness of critical development problems, such as the development environment, critical architectural elements, and other external entities of the software to be developed, as well as the communication between the elements and the entities. The output was the results obtained from the execution of the BMD calculation application for school children.

TABLE I.   DATA DICTIONARY

Attribute	Size	Data Type	Description
sSchool	30	String	School name
sName	30	String	Student name
iAge	2	Int	Student age
bGender	1	Bool	Gender of the student
fWeight	3.2	Float	Student Weight
fHeight	1.2	Float	Student Height
fSittingHeight	1.2	Float	Sitting Height
fForearmLength	1.2	Float	Forearm Length
fAbdominalCirc	3.2	Float	Abdominal Circumference
fFemurLength	1.2	Float	Femur Length
fBmd	2.2	Float	Bone mineral density
fBmi	2.2	Float	Body Mass Indice

C. Productionize

The production phase aims to implement the required functionality in the product by applying an incremental and iterative development cycle following the good practices of agile methodology [21]. The stages of the Productionize phase are presented in the following Fig. 4.

- *Planning*: The planning of the requirements analysis, Acceptance tests, planning of iterations, post-iteration tasks were carried out.

  The primary function of the Mobile App is to calculate the BMD bone mineral density that was estimated from the equations proposed by Gómez-Campos et al. [22], for both men and women. For each equation, the APVC value, forearm length, and knee diameter are required. Here is the formula:

  BMD (men) = 0.605+0.056*APHV + 0.008* Forearm Length + 0.022* Knee diameter.

  BMD (women) = 0.469+0.027* APHV + 0.007* Forearm Length + 0.019* Knee diameter.

  Note: To calculate APHV it is necessary to consult Mirwald et al [23]. This equation requires the following data: chronological age, weight, standing height, sitting height, and LP leg length (LP = standing height - sitting height).

- *Working day*: It was done following the recommendations of the agile frameworks, choosing the requirements with the highest priority, and implementing them through sprints.

- *Release*: The functional version was launched incrementally for testing and verification of results, system integration, pre-launch testing, and acceptance testing was performed.

Fig. 5 displays the user authentication screen interface and anthropometric data input to the mobile app for BMD calculation.

D. Stabilize

The last integration actions are carried out to ensure that the complete system works correctly. The stabilization phase pattern can be observed in Fig. 6 through the following stages.

Each new addition must work correctly with the rest of the software, and it is done for each new function developed. The use of Firebase Realtime Database ensures the scalability of the data, as well as its adequate access and query operations.

The refactoring of the requirements for defining the type of Inspector, interface refinements, and execution of acceptance tests was carried out.
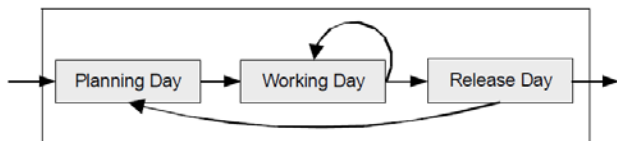
Fig. 4. Productionize Phase [20].
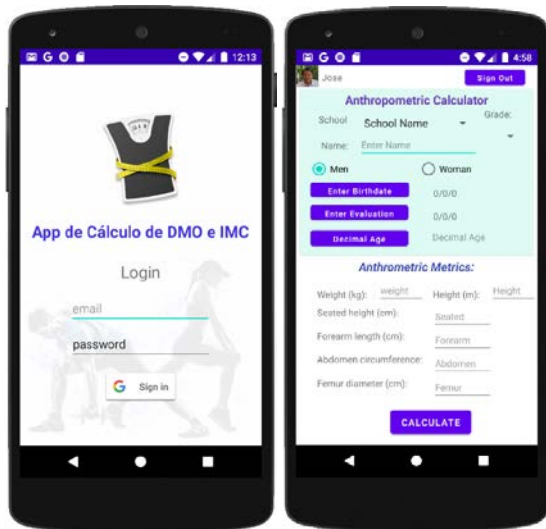
Fig. 5. Interfaces of Authentication and Data Entry of the Mobile App.
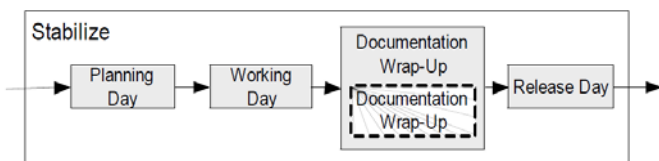
Fig. 6. Stabilize Phase [20].

Fig. 7 shows the storage that is made in the Firebase Realtime Database, which organizes it according to the entered data of the students.

According to the data and the calculation utilizing anthropometric regression equations, the Mobile App obtains the BMD. Fig. 8 shows the result of the BMD calculation that indicates the degree of osteoporosis if the student has or not, and the result is also shown in a percentile graph.

E. System Test

In this last phase, the complete software is tested to see if the product implements the required functionalities correctly, and to correct the errors found.
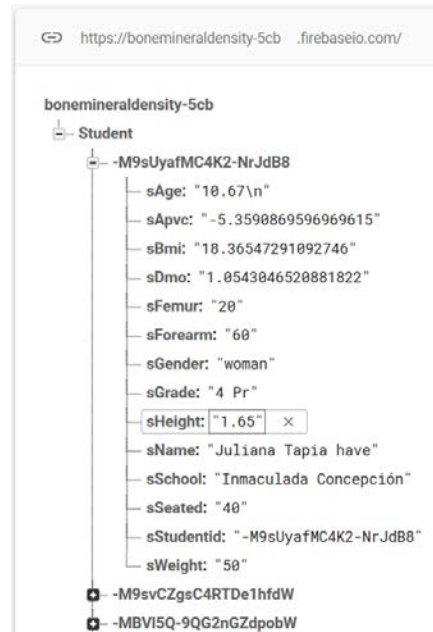
Fig. 7. Data Storage in Firebase Realtime Database.

Fig. 8. Result of BMD Calculation in the Mobile App.

Fig. 9 illustrates the stages of the system test and repair phase.

The evaluation work plan considered the following steps:

- Establish evaluation requirements,

- Specify the evaluation,

- Design the evaluation,

- Run the assessment,

- Conclude the evaluation,

A test plan was made according to the requirements such as Unit, Instrumentation, and UI provided by the Android Studio tool. Once completed, corrections and repairs of possible errors are made.

For the evaluation of the quality of use of the Mobile App, what was proposed by the ISO/IEC 25022 standard was used [12].

The ISO / IEC 25022: 2016 standard contains the following:

- a basic set of measurements for each quality feature in use;

- an explanation of how quality in use is measured.

It provides a suggested set of in-use quality measures to be used with the in-use quality model in ISO / IEC 25010.

Table II shows the Definitions of quality-in-use characteristics as defined by the ISO 25022 standard.
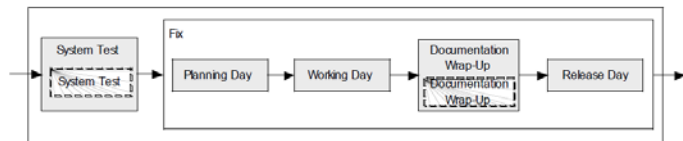
Fig. 9. System Test and Fix Phase [20].

TABLE II.   METRICS DEFINED IN THE ISO/IEC 25022

Characteristic	Definitions
Effectiveness	Accuracy of achieving the expected objective after performing an action.
Efficiency	Resources expended about the accuracy and completeness with which users achieve goals.
Satisfaction	*Usefulness*: the degree of user satisfaction with the achievement of the objectives, including the results of use and the consequences of use. *Trust*: the degree to which a user has confidence that a software product will perform as intended. *Comfort*: the degree to which the user is satisfied with physical comfort.
Freedom from risk	*Environmental Risk Mitigation*: The degree to which a software product mitigates the potential risk to property or the environment in the intended contexts of use.
Context coverage	*Context Completeness*: the degree to which a software product can be used with effectiveness, efficiency, freedom from risk, and with adequate satisfaction.

According to the ISO/IEC 25022 metric specification format, Table III presents an example of the Usage metric, belonging to the Usefulness sub-feature with its measurement details.

Once all the phases are finished, it should have a product that can be delivered to the end-user.

TABLE III.   FORMAL SPECIFICATION OF THE QUALITY IN USE METRIC

Sub-Charact.	Metric	Purpose	Method	Formula	Value
Usefulness	Discretionary use of functions	What percentage of users choose to use the system functions?	Observation	X=A/B where B>0.	0≤X≤1 The closer to 1 is the best.

IV. RESULTS

The quality metrics in the use of the product based on ISO / IEC 25022: 2016 allow to analyze the characteristics of the interactions in different interest groups with the product. This model is the most recent where the metrics used to measure the quality of product use are effectiveness, efficiency, satisfaction, freedom from risk, and coverage of the software context.

Table IV lists the quality metrics in the use of the software product, as well as its subcategories, the assigned weight (depending on the importance of this study), the value of the measures used, and the result achieved (ISO / IEC 25022: 2016).

TABLE IV.   RESULTS OF THE QUALITY ASSESSMENT OF ISO/IEC 25022 (QUALITY IN USE) SOFTWARE

Category	Subcategory	Weight category	Weight sub category	Measure	Result achieved
Effectiveness	Completed tasks	35%	20%	0.95	33%
	Objectives achieved		10%	0.95	
	Task errors		5%	0.90	
Efficiency	Tasks time	30%	10%	0.90	28%
	Time efficiency		15%	0.95	
	Profitability		5%	0.95	
Satisfaction	Usefulness	25%	10%	0.90	22%
	Trust		10%	0.90	
	Comfort		5%	0.95	
Freedom of Risk	Reduction of environmental risks	5%	5%	1.00	5%
Context coverage	Context Completenes	5%	5%	0.95	5%
Total		100%			93%

As shown in Table IV, it got 33% effectiveness, 28% efficiency, 22% satisfaction, 5% freedom of risk, 5% context coverage adding up to 93% in quality of use.

V.   Conclusions

The Android-based Mobile Application for Calculation of BMD of the students was developed following the steps of the Mobile-D methodology and using Google's Firebase technology, which proved to be very efficient for the massive handling of data.

Tests carried out using the ISO / IEC 25022: 2016 standard resulted in 93% quality in use concerning the implementation of the system, so it is considered acceptable for use in taking anthropometric measurements and calculating the BMD in school children.

The execution of the mobile app showed that the BMD calculation is obtained in a fast, scalable, and reliable way allowing to determine the degree of osteoporosis that a scholar has or not.

It is widely known that an evaluation of BMD in children and adolescents has high economic costs, however, the use of an anthropometric technique can save time, money and perform a quick evaluation.

The use of this application can serve as an instrument that allows controlling bone development and mineral accumulation during the growth stage. It is a non-invasive alternative that asks to be used daily in the school system.

For future work, this study can be extended the Mobile App with Machine Learning technologies to determine the classification of bone health; Likewise, the calculation of the BMD can be carried out, which would also support the diagnosis of osteoporosis.

This Mobile App can be used as a baseline by specialists in the academic and school health fields for quick decision-making regarding the BMD.

Acknowledgments

References

[1]   M. Li, F. Wan, J. Liu, Q. Hao, and B. Chen, "Analysis of Clinical Risk Factors for Refracture in Osteoporosis," J. Med. Imaging Heal. Informatics, vol. 10, no. 10, pp. 2337–2341, Jul. 2020.

[2]   J. A. Kanis, L. J. Melton, C. Christiansen, C. C. Johnston, and N. Khaltaev, "The diagnosis of osteoporosis," J. Bone Miner. Res., 1994.

[3]   F. Alvear-Vasquez, R. Gomez-Campos, P. Pezoa-Fuentes, C. Urra-Albornoz, J. Caceres-Bahamondes, C. Luarte-Rocha, J. Sulla-Torres and M. A. Cossio-Bolaños, "Maximum expiratory flow and handgrip strength predict bone health in children and adolescents," Retos, 2020.

[4]   S. K. Choi, B. Yelton, V. K. Ezeanya, K. Kannaley, and D. B. Friedman, "Review of the Content and Quality of Mobile Applications About Alzheimer's Disease and Related Dementias," J. Appl. Gerontol., 2020.

[5]   S. Knez and L. Šajn, "Food object recognition using a mobile device: Evaluation of currently implemented systems," Trends in Food Science and Technology. 2020.

[6]   S. D. Birkhoff and S. C. Smeltzer, "Perceptions of Smartphone User-Centered Mobile Health Tracking Apps Across Various Chronic Illness Populations: An Integrative Review," J. Nurs. Scholarsh., 2017.

[7]   M. T. Baysari and J. I. Westbrook, "Mobile Applications for Patient-centered Care Coordination: A Review of Human Factors Methods Applied to their Design, Development, and Evaluation," Yearbook of medical informatics. 2015.

[8]   W. Liu, Y. Liu, L. Pei, and N. Cao, "Design and Implementation of Family Doctor App on Android Platform," in Proceedings - 9th International Conference on Information Technology in Medicine and Education, ITME 2018, 2018.

[9]   P. Ryan, R. L. Brown, M. E. Csuka, and P. Papanek, "Efficacy of Osteoporosis Prevention Smartphone App," Nurs. Res., 2020.

[10]   P. Ravn Jakobsen, A. P. Hermann, J. Søndergaard, U. K. Wiil, and J. Clemensen, "Help at hand: Women's experiences of using a mobile health application upon diagnosis of asymptomatic osteoporosis," SAGE Open Med., 2018.

[11]   R. Alnanih, N. Bahatheg, M. Alamri, and R. Algizani, "Mobile-d approach-based persona for designing user interface," Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 5, pp. 2597–2607, Sep. 2019.

[12]   C. Guevara-Vega, J. Hernández-Rojas, M. Botto-Tobar, I. García-Santillán, A. Basantes Andrade, and A. Quiña-Mera, "Automation of the Municipal Inspection Process in Ecuador Applying Mobile-D for Android," in Advances in Intelligent Systems and Computing, 2020.

[13]   V. H. Goh and Y. W. Hau, "Android-based mobile application for home-based electrocardiogram monitoring device with google technology and bluetooth wireless communication," in 2018 IEEE EMBS Conference on Biomedical Engineering and Sciences, IECBES 2018 - Proceedings, 2019.

[14]   ISO/IEC, "ISO/IEC 25022:2016 - Systems and software engineering — Systems and software quality requirements and evaluation (SQuaRE) — Measurement of quality in use," ISO/IEC 25022:2016, 2016. [Online]. Available: https://www.iso.org/standard/35746.html. [Accessed: 12-Jul-2020].

[15]   I. Biscoglio and E. Marchetti, "An experiment of software quality evaluation in the audio-visual media preservation context," in Proceedings - 2014 9th International Conference on the Quality of Information and Communications Technology, QUATIC 2014, 2014.

[16]   C. D. Farinango, J. S. Benavides, J. D. Cerón, D. M. López, and R. E. Álvarez, "Human-centered design of a personal health record system for metabolic syndrome management based on the ISO 9241-210:2010 standard," J. Multidiscip. Healthc., 2018.

[17]   A. Idri, M. Bachiri, J. L. Fernandez-Aleman, and A. Toval, "ISO/IEC 25010 Based Evaluation of Free Mobile Personal Health Records for Pregnancy Monitoring," in Proceedings - International Computer Software and Applications Conference, 2017.

[18]   A. Idri, M. Bachiri, and J. L. Fernández-Alemán, "A Framework for Evaluating the Software Product Quality of Pregnancy Monitoring Mobile Personal Health Records," J. Med. Syst., 2016.

[19]   H. Nakai, N. Tsuda, K. Honda, H. Washizaki, and Y. Fukazawa, "Initial Framework for Software Quality Evaluation Based on ISO/IEC 25022 and ISO/IEC 25023," in Proceedings - 2016 IEEE International Conference on Software Quality, Reliability and Security-Companion, QRS-C 2016, 2016.

[20]   Agile, "Electronics -AGILE - Agile Software Technologies," Software Technologies Research Programme, 2020. [Online]. Available: http://virtual.vtt.fi/virtual/agile/mobiled.html. [Accessed: 01-Jun-2020].

[21]   M. C. Sáiz-Manzanares, R. Marticorena-Sánchez, Á. Arnáiz-González, J. F. Díez-Pastor, and S. Rodríguez-Arribas, "Computer Application for the registration and automation of the correction of a functional skills detection scale in early care," in INTED2019 Proceedings, 2019.

[22]   R. Gómez-Campos, C. L. Andruske, M. de Arruda, C. Urra Albornoz, and M. Cossio-Bolaños, "Proposed equations and reference values for calculating bone health in children and adolescent based on age and sex," PLoS One, vol. 12, no. 7, p. e0181918, Jul. 2017.

[23]   R. L. Mirwald, A. D. G. Baxter-Jones, D. A. Bailey, and G. P. Beunen, "An assessment of maturity from anthropometric measurements," Med. Sci. Sport. Exerc., vol. 34, no. 4, pp. 689–694, 2002.

Fuzzy based Reliable Cooperative Spectrum Sensing for Smart Grid Environment

Laila Nassef[1], Reemah Al-Hebshi[2]

Department of Computer Science
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Abstract—The huge demand for spectrum has created immediate need to make available new licensed and/or unlicensed spectrum bands to satisfy the explosive growth of spectrum demands and to satisfy the quality of service requirements of diverse applications. Spectrum shortage and harsh environment have become a challenging bottleneck to achieve reliable communications in the smart grid. Cognitive radio is the emerging technology to achieve both spectrum and reliability awareness. Cooperative spectrum sensing takes advantage of spatial diversity to reduce the impact of receiver uncertainty. However, the harsh smart grid environments limit advantageous of cooperation due to variations of signal to noise ratio on which energy detection technique depends on. This paper proposes a reliable spectrum detection for a cluster based cooperative spectrum sensing in harsh smart grid environment, where cognitive cluster heads and a centralized cognitive radio based fusion center are deployed to solve both spectrum and reliability problems. The proposed fuzzy inference system is based on three fuzzy descriptors of energy difference, link quality, and local probability of detection. The results show the superiority of proposed fuzzy based fusion scheme to enhance accuracy of spectrum decision in harsh environment.

Keywords—Cognitive radio; wireless networks; cooperative spectrum sensing; reliable fusion; fuzzy inference system

I. INTRODUCTION

The recent technological developments in wireless technologies have enabled the deployment of the smart grid [1], which is the next generation of electrical power grid proposed to solve problems associated with traditional power grid [2]. The massive increase in diverse applications have resulted in exponential increase in demand for spectrum which have created immediate need to make available new licensed and/or unlicensed spectrum bands [3]. Wireless Sensor Networks (WSNs) have been considered as a promising technology to enhance capabilities of monitoring, control, and maintenance of the entire smart grid from power generation to transmission, distribution and consumption [4]. Different types of communication technologies are currently used to collect, aggregate, and analyze data collected from all sectors of the smart grid. However, the severe conditions of electrical power grid may adversely affect network reliability. Furthermore, WSNs are typically operate in the unlicensed Industrial, Scientific and Medical (ISM) frequency band which is already congested and overloaded with many other coexisting technologies. This increases interferences from neighboring

networks and makes communication a challenging task due to spectrum scarcity and interference problems which could result in failure to establish and maintain reliable communications. There is need to sustainably support the continuously growing demand and to overcome both spectrum shortage and unreliable communications [4]. Cognitive radio (CR) technology allows the unlicensed Secondary Users (SUs) to opportunistically access the available licensed spectrum assigned to Primary Users (PUs) [5]. These underutilized licensed spectrum bands are assigned temporarily without the need to purchase the spectrum which offer reliable communications to ensure reliable communication in highly congested periods by transmitting data either on the original unlicensed channel or the additional licensed channel [6].

Spectrum Sensing is the most important function to detect the presence or absence of PUs at a certain location, at a given time, and in a specified frequency band. One of the spectrum sensing techniques is the transmitter detection which is further divided according to detection method into several methods, one of which is the energy detection [7]. Energy detection is the preferred technique for resource constraint nodes due to its low computational requirements compared to other techniques. Nodes in different spatial locations individually make local decision in a non-cooperative spectrum sensing by comparing the sensed energy with a predefined threshold to decide accordingly. However, due to the effect of propagation impairments, the Signal to Noise Ratio (SNR) of the received PU signal can be very low and individual SUs may not be able to distinguish between severely attenuated signal and vacant channel. Accurate spectrum sensing is important to mitigate the effect of environments and to guarantee reliable communications [8].

Cooperative Spectrum Sensing (CSS) has been proposed to enhance local spectrum decisions, where a group of SUs share their sensing reports with a Fusion Center (FC) either centralized or distributed to overcome receiver uncertainty by exploiting the spatial diversity [9]. In centralized CSS, the FC collects local spectrum sensing reports either as one-bit, quantized multi bits, or raw energy to identify spectrum occupancy, and broadcasts final global spectrum decision back to all nodes. Whereas in distributed CSS, each node shares spectrum reports with all of its neighbor nodes. CCS consumes energy in resource constraints nodes which cannot tolerate the heavy computations in distributed CSS. On the other hand,

centralized CSS consumes large bandwidth and also consumes energy which reduce network's lifetime [10].

CR based communications can be utilized in all sectors of the smart grid [11]. Reliable communication is urgently needed to sustainably solve spectrum scarcity problem, satisfy the ever increasing demand, and support the diverse applications [12]. However, the detection of available spectrum depends on multiple parameters which form a multi-objective optimization problem. Fuzzy logic can be utilized to solve this optimization problem that is difficult to be modeled using the mathematical methods. Fuzzy logic models are relatively simple and less computationally complex which make it suitable for resource constraints WSNs [13].

This paper proposes a Fuzzy Inference System (FIS) to enhance accuracy of cooperative spectrum detection in harsh smart grid environment. The harsh environment is prone to error and spectrum sensing reports will not be received correctly. To the best of our knowledge, none of the previous work have proposed a FIS to enhance detection accuracy in harsh smart grid environment. The contribution of this paper is the development of fuzzy logic based soft fusion strategy to accurately detect the opportunistic spectrum in harsh environment. The proposed system is based on three fuzzy descriptors of energy difference, channel condition, and local probability of detection.

The organization of this paper is as follows. Section II provides the related work. Section III describes the local spectrum sensing decisions in various environments. The performance indicators are also mathematically formulated. Section IV presents the cooperative hard fusion strategies in harsh environments. Section V proposes the FIS. The simulation results are demonstrated in Section VI. Finally, Section VII provides conclusion and future research work.

II. RELATED WORK

The increasing demand for more spectrum has motivated the development of CR based network to solve both spectrum scarcity and reliability problems facing the introduction of a wide range of new applications [14]. However, most of the previous work on spectrum sensing have assumed simplified propagation model in which received signal depends only on distance between transmitter and receiver. Additive White Gaussian Noise (AWGN) channel model for sensing channels was assumed while reporting channels were assumed to be dedicated error free channels. This model cannot accurately represent the real environments where propagation of electromagnetic signals suffers from propagation impairments which negatively impact the performance and present great challenge to reliability of wireless communications. The reliability of communication channels varies significantly from channel to channel and over time which impact the diverse applications which have stringent Quality of Service (QoS) requirements with respect to throughput, delay, and availability of the communications channels [15]. The random variation of the received signal makes spectrum sensing decision not accurate and nodes may fail to detect the spectrum due to presence of fading, shadowing, and time-varying nature of wireless channels. The randomness features of communication channels and variations in SNR causes uncertainty in sensing

reports and makes it difficult to accurately achieve reliable spectrum detection decisions which could results in severe threat to reliability of network and sever threat to data integrity due to failure to satisfy spectrum demand by the diverse applications.

In Cluster-Based CSS, SUs perform spectrum sensing and send spectrum sensing reports to a Cluster Head (CH) and aggregate sensing reports to FC to perform the energy consuming functions [16]. The FC combines the individual reports either by hard-decision or soft-data fusion to increase the accuracy of local spectrum sensing. In hard combining, sensing reports are received as a one-bit decision and combined using the counting rule of "$l\ out\ of\ N$", where "OR" rule, "AND" rule, and Majority rule are special cases [17]. Hard fusion decreases overhead on reporting channel, however the use of one energy threshold introduces information loss which degrade performance. On the other hand, soft fusion rules depend on raw sensed energy reports. The conventional soft combining schemes such as Square Law Combining (SLC), Maximal Ratio Combining (MRC), and Selection Combining (SC) enhance detection performance at a cost of increasing communication overhead. Nodes send their raw data to FC which consumes energy and bandwidth due to periodical transmission of raw data [18].

A tradeoff between overhead and detection performance is achieved by using a softened quantized multi-bit hard fusion [19]. Each SU quantizes its local observations into multiple decision regions to reduce overhead and achieve reliable sensing performance simultaneously. A quantization based multibit soft fusion rule is proposed in [20] to provide a compromise between sensing performance and control channel overhead. SUs use energy detector to observe the received signal level and compare it with quantization thresholds to estimate the multibit data. The data from all SUs are transmitted to FC to perform inverse quantization for the spectrum occupancy decision [21]. However, the more the number of bits, the more the overhead and the accurate the final global decision [22]. When the number of SUs is high, increasing the quantization bits provides no performance gain and quantized two bits energy report can be used. A softened two bits hard scheme that divides energy level into four quantization levels is developed in [23]. While in [24] a soft combining rule is proposed that assumed SNR is available at SUs. Similarly, a two-bit softened hard scheme is proposed in [25] for a cluster based CSS. One bit hard fusion is used based on SNR and one bit majority vote is used to combine the results.

Fuzzy logic can be utilized with CR based WSNs to produce the spectrum decision [26]. It is similar to natural language and can deal with uncertainty and imprecise knowledge for decision making. It allows phrases of real situations to feed mathematical models of reasoning. The statement truth of fuzzy logic depends on the degree of membership which have values between 0 and 1. Fuzzy logic is used in [27] with three descriptors of spectrum utilization efficiency of the secondary user, of mobility, and its distance to the primary user. Fuzzy channel ranking estimation was presented in [28] using four descriptors of variability of the channel duration availability, duration of the channel idle time,

channel condition and distance. Similarly, in [29] fuzzy based spectrum allocation was proposed to enhance spectrum efficiency based on three descriptors. IEEE 802.22 wireless regional area network (WRAN) is proposed for the smart grid to allow unlicensed broadband network to opportunistically access the unused TV white space [30]. Fuzzy logic is also applied to the IEEE 802.22 to prioritize channels in the backup and candidate channel list [31]. Finally, fuzzy logic was also proposed for IEEE 802.22 WRAN in [32] to select channels based on required QoS using throughput, latency and reliability.

III. LOCAL DECISION IN VARIOUS ENVIRONMENTS

The FC is responsible of coordinating the spectrum availability and providing the list of available spectrum to be scanned and utilized. SUs sense the spectrum and transmit their local spectrum decisions about spectrum availability to their corresponding CR based Cluster Head (CH). The network is clustered into c clusters with cluster 1, cluster 2, and cluster j, each has cluster members of k_1 , k_2,.., k_c respectively, where $1 \leq j \leq k_c$. The CHs collect sensing reports, schedule and distribute the spectrum sensing tasks among the cluster members. Upon joining a cluster, j, a SU_i becomes a cluster member defined as SU_i^j. The sensing reports from all cluster heads CH^j are collected, to make the spectrum sensing decision, and this decision is propagated back to SUs to access the available spectrum.

The functionality of spectrum sensing is implemented at both physical and MAC layers [33]. Physical layer detects PU's signal using energy detection and MAC focuses on time to sense and channels to sense [34]. CHs schedule and transmit sensing reports in a Time Division Multiple Access (TDMA) where total sensing period is divided into a number of sensing slots, each has a duration of T_{t_i} that is divided into two sub-slots: a sensing sub-slot of T_{s_i} and a reporting sub-slot of T_{r_i}. All SUs sense the channels in the assigned time slot and then forward sensing report over the reporting channel in the scheduled reporting time slots [35]. The spectrum sensing may behave as binary with hypothesis testing problem, if there are no activities on the channel, it is considered as a null hypothesis H_0 or otherwise it is considered as busy channel H_1.

All SUs forward their sensing reports to FC about each channel. Without loss of generalities, the following is for one channel, where the local spectrum decision is decided based on local energy threshold, sensing parameters, and environment. The propagation is modeled as a simplified AWGN non fading, or realistic environments lognormal shadowing with no channel errors, or harsh environment with errors.

A. Non-Fading AWGN Environments

The spectrum sensing in AWGN assumes sensing channels to be error free with AWGN [33]. For each cluster member SU_i^j , the received signal in the τth sensing time is defined as $y_i(\tau)$, where $i = 1, \ldots, k_c$, is formulated as:

$$\begin{cases} y_i(\tau) = n_i(\tau) & : H_0 \\ y_i(\tau) = \hbar_{pi}.X_p(\tau) + n_i(\tau): & H_1 \end{cases} \tag{1}$$

where $\tau = 1, \ldots, S$, where S is the number of received samples within the spectrum sensing period T_s such as $S = T_s f_s$. The PU signal sensed over the sensing channel $X_{p_i}(\tau)$ has a carrier frequency of f_s with zero mean and variance of σ_{p_i}. The AWGN noise signal $n_i(\tau)$ has zero means and variance of σ_{n_i}. Both PU's signal $X_{p_i}(\tau)$ and noise signal $n_i(\tau)$ are assumed to be independent and identically distributed random processes. $\hbar_{pi}$ is the channel fading coefficient. When number of samples "S" is large enough, the Probability Distribution Function (PDF) of Y_i follows a central chi square distributed under hypothesis H_0 and a non-central chi-square distributed with $\mathcal{N}$ degree of freedom under hypothesis H_1. Then the observed sensed energy statistics E_i can be approximated by the Gaussian distributions and formulated as:

$$E_i = \frac{1}{S} \sum_{\tau=1}^{S} |y_i(\tau)|^2 \tag{2}$$

The local hard decision u_i of the individual SU is computed by comparing its received energy E_i with a predetermined threshold λ_{th} to decide on hypothesis $\mathcal{H}_1$ and $\mathcal{H}_0$ that is formulated as:

$$u_i = \begin{cases} \mathcal{H}_1 : E_i \geq \lambda_{th} \\ \mathcal{H}_0 : otherwise \end{cases} \tag{3}$$

The performance indicators are evaluated to know whether the sensed value is generated under hypothesis $\mathcal{H}_0$ or hypothesis $\mathcal{H}_1$ which represent idle and occupancy state the PU. The probability of detection (P_d) is the probability that decision is $\mathcal{H}_1$ when $\mathcal{H}_1$ is true which is formulated as:

$$P_d = P[\mathcal{H}_1|\mathcal{H}_1] = P[E_i > \lambda_{th}|\mathcal{H}_1] \tag{4}$$

The probability of false alarm (P_{fa}) is the probability that decision is $\mathcal{H}_1$ but is true which is formulated as.

$$P_{fa} = P[\mathcal{H}_1|\mathcal{H}_0] = P[E_i > \lambda_{th}|\mathcal{H}_0] \tag{5}$$

P_{fa} is an indicator of the spectrum utilization, where a high value of P_{fa} means less spectrum utilization and lower value of P_{fa} means higher spectrum utilization. Both higher (P_d) and lower (P_{fa}) provide more protection for the PUs and improved spectrum utilization. As E_i follows a Gaussian distribution with zero mean and variance of σ_n^2 , then performance is approximated as follows:

$$E_i|\mathcal{H}_1 = P_d \sim \mathcal{N}\left((1+\gamma)\,\sigma_n^2, \frac{2(1+\gamma)^2\sigma_n^4}{S}\right) \tag{6}$$

$$E_i|\mathcal{H}_0 = P_{fa} \sim \mathcal{N}\left(\sigma_n^2, \frac{2\sigma_n^4}{S}\right) \tag{7}$$

where γ is SNR. The performance indicators in AWGN environment are computed as follows:

$$P_d = E_i|\mathcal{H}_1 = Q\left(\left[\frac{\frac{\lambda_{th}}{\sigma_n^2}}{(1+\gamma)} - 1\right]\sqrt{\frac{T_s f_s}{2}}\right) \tag{8}$$

$$P_{fa} = E_i|\mathcal{H}_0 = Q\left(\left[\frac{\lambda_{th}}{\sigma_n^2} - 1\right]\sqrt{\frac{T_s f_s}{2}}\right) \tag{9}$$

where Q(.) is the complementary distribution function of the standard Gaussian function.

B. Error Free Fading Environment

The smart grid have noise, interferences, and multipath propagation effects [36], and the received SNR at different SUs is not the same and can be computed as follows.

$$SNR_i^j = \gamma_i^j = \frac{\hbar_{pi}(\tau) X_{pi}(\tau)}{\sum_{l=1,l\neq i}^{k_c} \hbar_{li}(\tau) X_{li}(\tau) + \sigma_l^2} \qquad (10)$$

where there is l node index and $l < k_c$, $\sum_{l=1,l\neq i}^{k_c} \hbar_{li}(\tau) X_{li}(\tau)$ is the accumulated interfering signals coming from l nodes. For the error free fading Environments (channel error probability $P_e = 0$), the sensing channels will have variable channel fading coefficient $\hbar_{pi}$ and consequently have variable SNR. The average probability of detection $\bar{P}_{d,i}$ will be computed as

$$\bar{P}_{d,i} = \int_0^\infty Q\left(\left[\frac{\frac{\lambda_{th}}{\sigma_{in}^2}}{(1+\gamma_i)} - 1\right]\sqrt{\frac{T_s f_s}{2}}\right) f(\gamma_i) d\gamma_i \qquad (11)$$

where $f(\gamma_i)$ is the Probability Distribution Function (PDF) of γ_i. Under H_0 case, the P_{fa} is independent of γ_i and the average $\bar{P}_{fa,i}$ is the same for all SU_i as computed in equation (9). Then the PDF of γ_i can be approximated.

$$f(\gamma_i) = \sqrt{\frac{\alpha_i}{2\pi\gamma_i}} \frac{1}{\gamma_i} \exp - \frac{\alpha_i(\gamma_i - \beta_i)^2}{2\beta_i^2 \gamma_i}; \ \gamma_i \geq 0 \qquad (12)$$

where β_i is the expectation of γ_i i.e., $\beta_i = \varepsilon[\gamma_i]$, and α_i is the shape parameter. In lognormal distribution, the relation between the parameters $\{\alpha_i, \beta_i\}$ and $\{\mu_i, \sigma_{i_{dB}}\}$ is specified as follow:

$$\beta_i = \exp\left(\frac{\mu}{\psi}, \frac{\sigma_{i_{dB}}^2}{2\psi^2}\right), \quad \alpha_i = \frac{\beta_i}{\exp\left(\frac{\sigma_{i_{dB}}^2}{\psi^2}\right) - 1} \qquad (13)$$

where $= \frac{10}{\ln(10)}$, μ is mean, and $\sigma_{i_{dB}}$ is the standard deviation of the SNR_i in logarithmic scale ($10 \log 10 \gamma_i$). Then $Q(x)$ function is computed as $\frac{1}{2} erfc\left(\frac{x}{\sqrt{2}}\right)$, where $erfc(.)$ represents complementary error function and defined as follows:

$$erfc(x) = 1 - \frac{2}{\sqrt{\pi}} \sum_{k=0}^\infty \frac{(-1)^k t^{2k+1}}{2^{k+\frac{1}{2}} k!(2k+1)} \qquad (14)$$

In this case, $Q(x)$ function is computed as

$$Q(x) = \frac{1}{2} \frac{1}{\sqrt{\pi}} \sum_{k=0}^\infty \frac{(-1)^k t^{2k+1}}{2^{k+\frac{1}{2}} k!(2k+1)} \qquad (15)$$

Where $t = \left[\frac{\frac{\lambda_{th}}{\sigma_i^2}}{(1+\gamma_i)} - 1\right]\sqrt{\frac{T_s f_s}{2}}$

The average probability of detection $\bar{P}_{d,i}$ at different SU_i is a function of γ_i and is formulated as:

$$\bar{P}_{d,i} = \sqrt{\frac{\alpha_i}{8\pi}} \int_0^\infty \gamma_i^{-3/2} exp\left[-\frac{\alpha_i(\gamma_i - \beta_i)^2}{2\beta_i^2 \gamma_i}\right] d\gamma_i -$$

$$\sqrt{\frac{\alpha_i}{2\pi^2}}\left(\frac{T_s f_s}{2}\right)^{2k+1} \sum_{k=1}^\infty \frac{(-1)^k}{2^k + \frac{1}{2}(2k+1)k!} \times$$

$$\int_0^\infty \left(\frac{\frac{\lambda_{th}}{\sigma^2}}{1+\gamma_i} - 1\right)^{2k+1} \gamma_i^{-3/2} exp\left[-\frac{\alpha_i(\gamma_i - \beta_i)^2}{2\beta_i^2 \gamma_i}\right] d\gamma_i \qquad (16)$$

For a target performance and if each cluster is assumed to have the same channel gain, the performance indicators can be approximated as follow:

$$\bar{P}_{d,i} = Pr\{u_i = 1|\mathcal{H}_1\}, \bar{P}_{d,i} = \bar{P}_d; \ \forall i \qquad (17)$$

$$\bar{P}_{fa,i} = Pr\{u_i = 1|\mathcal{H}_0\}, \bar{P}_{fa,i} = \bar{P}_{fa}; \ \forall i \qquad (18)$$

$$\bar{P}_{md,i} = Pr\{u_i = 0|\mathcal{H}_1\}, \bar{P}_{md,i} = \bar{P}_{md}; \ \forall i \qquad (19)$$

where $\bar{P}_{md} = 1 - \bar{P}_d$ is the probability of miss detection.

C. Harsh Environment

Due severe noise and shadow fading in sensing channels in each cluster, the interference increases packet loss rate, which increases demand due packet retransmissions and consequently deteriorate QoS due increase in retransmission. This causes errors and the performance indicators should be modified to include the channel error probability P_e as follows:

$$\bar{P}_{fe} = \bar{P}_{fa}(1 - P_e) + (1 - \bar{P}_{fa})P_e \qquad (20)$$

$$\bar{P}_{mde} = \bar{P}_{mde}(1 - P_e) + (1 - \bar{P}_{mde})P_e \qquad (21)$$

IV. Cooperative Hard Fusion in Harsh Environment

Local spectrum sensing performs poorly in harsh environment. CSS exploits the spatial diversity from multiple spectrum sensing nodes, where each SU_i submits a one bit local decision to the CH^j through the reporting channels to aggregate these reports to the FC to combine these reports using hard fusion strategies or soft fusion strategy if reports have raw data. CSS improves detection performance where nodes in different spatial locations share their sensing reports with the FC to overcome uncertainties of individual decision. Either hard or soft fusion strategies can be used [37]. Local binary decisions are transmitted to FC to combine these sensing reports to evaluate performance indicators based on OR, AND, or majority voting strategies [38]. Similar to the local decision, the final CSS decision is measured based on both P_d and P_{fa}.

Under 'AND' rule FC decides the presence of PU when all the SU's detects the PU signal.

$$P_D^{AND} = \prod_{i=1}^{k_c} P_d, \ P_{FA}^{AND} = \prod_{i=1}^{k_c} P_{fa} \qquad (22)$$

Under 'OR' rule, at least one of the SU's involved in the sensing decides the presence of the PU.

$$P_D^{OR} = 1 - \prod_{i=1}^{k_c}(1 - P_{d_i}),$$

$$P_{FA}^{OR} = 1 - \prod_{i=1}^{k_c}\left(1 - P_{fa_i}\right) \qquad (23)$$

Under majority fusion rule (half voting rule), the cooperative probabilities can be simplified and formulated

using the normal Binomial distribution (instead of Poisson-Binomial) to become:

$$Q_d{}^{MAJ} = \sum_{l=i}^{k_c} \binom{k_c}{l} P_d{}^l (1 - P_d)^{k_c - l} \tag{24}$$

$$Q_{fa}{}^{MAJ} = \sum_{l=i}^{k_c} \binom{k_c}{l} P_f{}^l (1 - P_f)^{k_c - l} \tag{25}$$

In harsh environment, performance indicators are modified to include P_e as follow:

$$Q_{de}{}^{MAJORITY} = \sum_{l=i}^{k_c} \binom{k_c}{l} P_{de}{}^l (1 - P_{de})^{k_c - l} \tag{26}$$

$$Q_{fae}{}^{MAJORITY} = \sum_{l=i}^{k_c} \binom{k_c}{l} P_{fae}{}^l (1 - P_{fae})^{k_c - l} \tag{27}$$

$$Q_{mde}{}^{MAJORITY} = 1 - \sum_{l=i}^{k_c} \binom{k_c}{l} P_{mde}{}^{k_c - l} (1 - P_{mde})^l \tag{28}$$

V. PROPOSED FUZZY INFERENCE SYSTEM

In harsh environments, depending on energy statistic as the only parameter to detect PU is not enough to correctly detect the spectrum, as the received energy samples are corrupted by propagation impairments which cause performance degradation at low values of SNR. Fuzzy logic can deal with uncertainty in data by converting imprecise data into precise data. The FIS utilizes four steps of fuzzification, fuzzy inference engine, fuzzy rule base, and defuzzification as depicted in Fig. 1.

The fuzzification converts crisp values into linguistic variables to determine the membership function in which the fuzzy rule IF-THEN is applied to form the new outputs set. Three antecedents or descriptors are combined compute the spectrum detection decision. The first descriptor is energy difference, the second is the channel condition or SNR, and third is the local probability of detection which is directly computed. Each descriptor is assigned three linguistic variables to compute the reliable probability of detection. The defuzzifier computes the crisp output from the rule output sets.

The FIS computes reliable spectrum decision based on these three descriptors. Each is divided into three levels of linguistic values to reflect different degree of membership of input linguistic variables. These values are set after analyzing their range. The Fuzzification converts crisp values into linguistic variables. The rule base consists of a number logic rules in the form IF-THEN statements, where 'IF' part of the rule is called 'antecedent' and the 'THEN' part of the rule is called 'consequent'. These rule clauses depend on the number of linguistic variables and membership functions to find optimal solutions. The three descriptors are computed as follows.

A. Energy Matrix

The energy matrix is constructed from the received reports and the absolute energy difference $DIFF_i{}^j$ for each $SU_i{}^j$ from the average energy for all other SUs in each cluster j is computed by neglecting the $SU_i{}^j$ result in each sensing interval to evaluate impact of not including this particular $SU_i{}^j$ in the result. The energy difference matrix $DIFF_i{}^j$ for each node i in cluster j is formulated as

$$DIFF_i{}^j = \left| E_i^j - \frac{1}{k_c} \sum_{j=1}^{k_c} E_i^j \ (j = 1,2,\dots,c) \right| \tag{29}$$
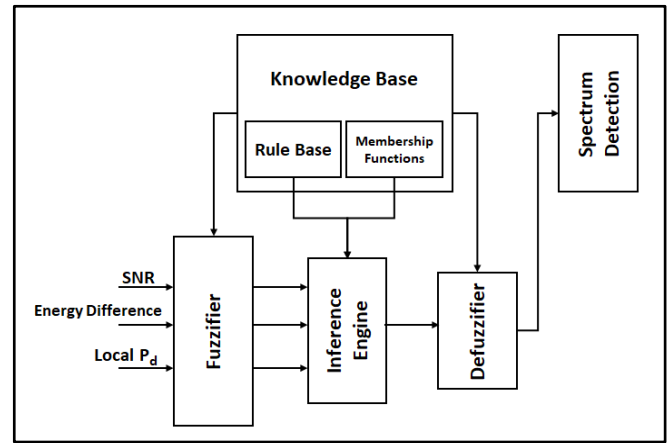
Fig. 1. Proposed Fuzzy Inference System.

The total number of $SU_{i\,i}{}^j$ in cluster j is k_j , where k_1, k_2 , ..., k_c and $1 \leq j \leq c$ means that k_1 SUs in Cluster 1, k_2 SUs in Cluster 2,... k_c SUs in cluster "c", respectively. The average value of energy reports of all other SUs is evaluated by keeping away the sensing report of the SU_i out of the average cluster's energy. This value is standardized to get its consistent dimensions as follows:

$$DIFF_i{}^j = \frac{Diff_{max} - DIFF_i{}^j}{Diff_{max} - Diff_{min}} \tag{30}$$

Energy difference to cluster average ($Diff_j{}^i$) is the Absolute differences of the sensing energy of $SU_i{}^j$ with the average sensing energy provided by all other SUs in each cluster. It has three linguistic values of Low, Medium, and High.

B. Channel Condition Matrix

The SNR, computed as $(\gamma_i{}^j)$, is standardized to get its consistent dimensions as follows:

$$\gamma_i{}^j = \frac{\gamma_{max} - \gamma_i{}^j}{\gamma_{max} - \gamma_{min}} \tag{31}$$

$\gamma_i{}^j$ has three linguistic values of Poor, Adequate, and Excellent.

C. Probability of Detection Matrix

The local probability of detection for each $SU_i{}^j$ is defined as P_{d_i} which has three linguistic values of Low, Medium, and High.

D. Reliable Probability of Detection

The output linguistic values that represent reliable decision for each $SU_i{}^j$ and defined as $P_{d_i}{}^j$ has seven linguistic values of Very High, High, Med High, Mid, Mid Low, Low, Very Low. Since there are three inputs parameters, each of them have three levels, a total of 27 fuzzy rules are used. The output is defuzzified to convert fuzzy output into crisp values again which is the actual output of the system or the reliable $P_{d,i}{}^j$ spectrum sensing decision.

E. Membership Functions

Many types of membership functions can be used to cover the complete input and output range, from which triangle and trapezoidal membership functions are the most useful types due to their simplicity to determine their degrees of membership. The middle levels of the three inputs re represented by triangle membership functions and the other levels are represented by trapezoidal membership functions. The seven-output linguistic variable are represented by triangular function as shown in Fig. 2. The two extreme IF-ELSE statement rules fall between these two cases:

Case (1): If $Diff^i_j$ is high, γ^j_i is poor, and $P^j_{d,i}$ is low then Reliable $P^j_{d,i}$ is verylow.

Case (2): If $Diff^i_j$ is low, γ^j_i is excellent, and $P^j_{d,i}$ is high then Reliable $P^j_{d,i}$ is veryhigh.
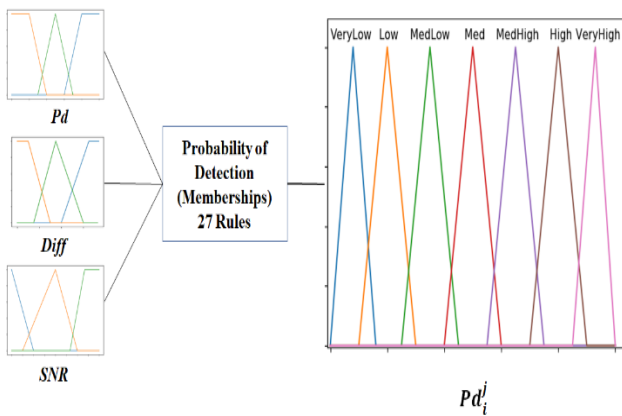
Fig. 2.   The Input and Output Membership Functions.

VI.   SIMULATION AND RESULT DISCUSSIONS

The cluster architecture is naturally suitable for all sectors of the smart grid, one of which is the Home Area Networks (HANs). HANs consist of a large number of smart appliances that communicate with utility to support various smart grid applications. This environment is simulated with all of the required simulation parameters selected based on this specific environment. The sensing reports are collected and transmitted to the corresponding CH which aggregate these reports to the FC to compute reliable spectrum decision based on the proposed FIS. The range of the inputs and outputs of the three descriptors are selected based on experience acquired from running several simulations. The numerical values are then correlated with linguistic variables describing their degree of belonging to each of the three variables. A rule table, in the form of if-then statements, is used to represent the fuzzy rules to generate the values after defuzzification is done on the output.

A. Simulation Environment

The network nodes are assumed to be static SUs and are deployed within an area of 100mx100m. SUs are assumed to have a single interface that can switch among C traffic channels accessed opportunistically, along with the predefined common control channel. These SUs coexist with M PUs that

can appear on C channels. The SUs can opportunistically access the available licensed channels only when PU is not active. The PU activity is modelled as independent and identically distributed random process with busy and vacant probabilities. Nodes are assumed to sense highly correlated sets of channels. Each cluster is assumed to have only 5 available channels to generate congestion. The propagation parameters are selected to give a high impact of environment. The traffic parameters are also selected based on data traffic load found in most HAN applications, which have a periodic traffic that is generated with constant bit rate (CBR) packets and have a rate of one data packet every 5 seconds to overload the network. Each packet has a size of 512 bytes, which corresponds to most smart grid applications. In addition, the data rate is set to the maximum data rate 250 kbps. The common simulation parameters are presented in Table I.

TABLE I.   COMMON SIMULATION PARAMETERS

Parameter	Values	Parameter	Values
Network radius	100 m	Target Probability of false alarm P_{fa}	0.1
FC location	50 m,50 m	Target Probability of detection P_d	0.9
Unlicensed Band	IEEE802.15.4	Channel Error probability P_e	0.3
Number of SUs	100	Super frame duration	10 msec
SU coverage radius	35 m	Sensing report duration T_s	1 msec
Number of channels	5	Sampling duration τ	0.1 msec
Path loss exponent	4.2	Number of samples ($\mathcal{S}$)	500
Shadowing Variance	4.0	PU coverage radius	50 m
Receiver sensitivity	-85 dBm	Number of PU	5
licensed Band	IEEE802.11	Frequency band	2.4 GHz

B. Simulation Results and Analysis

The impact of various environments, for different values of SNR, on the probability of detection for various fusion rules of OR, AND, Majority rule, and the proposed fuzzy fusion under both AWGN and harsh environment are presented in Fig. 3. Generally, the cluster based CSS enhances performance over local decision in harsh environment. The OR logic rule with and without error shows good performance compared to AND rule and Majority rule. The proposed fuzzy fusion has the highest detection probability in AWGN environment than other fusion rules due to enhancement in decision accuracy. However, the harsh environment has caused severe threat to reliability of sensing reports and threaten network operation due to failure to find any available spectrum. This has degraded performance in all schemes and reduced detection accuracy at lower values of SNR. The developed FIS was effective to enhance spectrum detection using three fuzzy descriptors of sensing node energy, channel condition, and local sensing decision. At the sensing node, lower values of SNR indicate unreliable sensing environment which introduce errors and degrade the reliability of sensing reports.

The performance of all fusion rules is degraded significantly under lower values of SNR which made spectrum

sensing difficult due to incorrect reception of sensing reports and inability to differentiate signal from noise. The degradation in detection probability for all schemes is due to inaccurate sensing reports that cause uncertainty and render the energy detector useless. For a target P_{fa_i}, as SNR increases, higher values of P_{d_i} is achieved as nodes are able to properly detect spectrum and provide more reliable sensing reports. The proposed fuzzy fusion has better performance than AND and Majority fusion rules due to enhancement in decision accuracy even at smaller values of SNR.

The harsh environment which is prone to error and sensing reports may not be received correctly have greatly affected all schemes and decreased the probability of detection. The fuzzy fusion was able to maintain the performance to a certain limit after which it could not compensate for errors introduced in environments. However, the OR logic fusion shows relatively good performance than the proposed fuzzy fusion and was able to achieve good performance even in harsh environment.

The probability of detection against probability of false alarm is presented in Fig. 4, which the proposed fuzzy fusion is compared with both majority fusion and local detection under both AWGN and harsh environments. The proposed fuzzy fusion in AWGN is shown for comparison which has the highest detection probability than other fusion rules due to enhancement in decision accuracy. The proposed FIS has enhanced the detection probability compared to other schemes and it is able to provide a sustainable solution to solve both spectrum scarcity and reliability problems in harsh environments. The reliable fuzzy based spectrum detection has accurately mitigated the impact of unreliable environment and has achieved more efficient and robust detection than other schemes. However, harsh environment is prone to error and sensing reports may not be received and this causes lower performance of the proposed fuzzy fusion under harsh environment as compared to the AWGN case.

Fig. 4. The Probability of Detection Against Probability of False Alarm.

Fig. 5 plots the probability of miss detection against probability of false alarm P_{fa} for different SNR. At lower SNR, performance deteriorates mainly because the energy detection method is based on received signal's energy and consequently higher noise levels at receiver greatly impact performance indicators. The probability of miss detection increases with the decrease in SNR and the harsh environment which introduces errors that leads to incorrect reception of sensing reports. At lower values of SNR, the probability of miss detection reaches 0 only when probability of false alarm is close to 1. However, for better values of SNR, lower values of probability of miss detection is achieved at lower values of probability of false alarm.

Fig. 3. The Probability of Detection Against Signal to Noise Ratio.
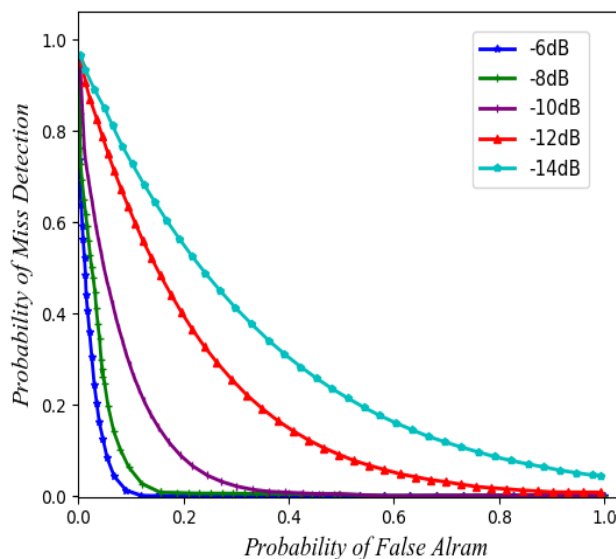
Fig. 5. Probability of Miss Detection Against Probability of False Alarm.

The lower values of probability of miss detection is achieved only at higher SNR. However, at lower SNR, nodes cannot recognize the signal and could identify the signal as a noise and consequently it will consider channel as vacant. The

harsh environment is subjected to errors that impact probability of false alarm P_{fa} which occurs if an idle channel is detected as busy and may lead to a potentially wasted opportunity for the SU to transmit. Also it impact the probability of miss detection P_{md} which occurs when a busy channel is detected as idle and could potentially lead to a collision or interference with the PU, leading to wasted transmissions for both PU and SU.

VII. Conclusion and Future Work

This paper proposes the use of a fuzzy inference system to enhance spectrum detection in a cluster based cooperative spectrum sensing in harsh environment. The proposed fuzzy based soft fusion is simulated and evaluated using the probability of detection and the probability of false alarm. The results indicated that the proposed FIS have significantly outperforms other fusion rules in harsh environment. It was effective to enhance probability of detection and provide more robust fusion decision by incorporating environment in decision making.

In this paper, the QoS was assumed to be satisfied by only accessing the available spectrum, however, various applications demand different data rates and use different data size with various transmission times which need to adapt to the PU activities. The PU's activities was fixed and the proposed FIS urgently needs to consider PU activities, sort the available channels based on their quality, and allocate available spectrum based on QoS requirements. Further maximization is expected by adapting sensing parameter to react to environment accordingly. However, these sensing parameters should be optimized along with other parameters to satisfy the QoS requirements of diverse applications. An immediate future work will be the development of this complex problem.

References

[1] Melike Yigit, V. Cagri Gungor, E. Fadel, L. Nassef, N. Akkari and I. F. Akyildiz, "Channel-Aware Routing and Priority-Aware Multi-Channel Scheduling for WSN-Based Smart Grid Applictions", Journal of Network and Computer Applications, vol. 71, pp. 50-58, 2015. https://doi.org/10.1016/j.jnca.2016.05.015.

[2] Yin XC, Liu ZG, Nkenyereye L and Ndibanje B, "Toward an Applied Cyber Security Solution in IoT-Based Smart Grids: An Intrusion Detection System Approach", Sensors (Basel), vol. 19, no. 20, 2019. DOI: 10.3390/s19224952.

[3] Laila Nassef, Remah Elhebshi and Linta Jose, "Evaluating Performance of Wireless Sensor Network in Realistic Smart Grid Environment", International Journal of Wireless & Mobile Networks, vol. 10, no. 3, pp. 27-36, 2018. DOI: 10.5121/ijwmn.2018.10303.

[4] Etimad Fadel, V.C. Gungor, Laila Nassef, Nadine Akkari, M.G. Abbas Malik, Suleiman Almasri and Ian F. Akyildiz, "A Survey on Wireless Sensor Networks for Smart Grid", Computer Communications, vol. 71, pp. 22-33, 2015. https://doi.org/10.1016/j.comcom.2015.09.006.

[5] Ian F. Akyildiz, Won-Yeol Lee, Mehmet C. Vuran and Shantidev Mohanty, "Next Generation/Dynamic Spectrum Access/Cognitive Radio Wireless Networks: A Survey", Computer Networks, vol. 50, no. 13, pp. 2127-2159, 2006. https://doi.org/10.1016/j.comnet.2006.05.001.

[6] Xin-Lin Huang, Xiao-Wei Tang and Fei Hu, "Dynamic Spectrum Access for Multimedia Transmission Over Multi-User, Multi-Channel Cognitive Radio Networks", IEEE Transactions on Multimedia, pp. 201 - 214, 2020. DOI: 10.1109/TMM.2019.2925960.

[7] James Adu Ansere, Guangjie Han, Hao Wang and Celimuge Wu, "A Reliable Energy Efficient Dynamic Spectrum Sensing for Cognitive Radio IoT Networks", IEEE Internet of Things Journal, vol. 6, no. 4, pp. 6748 - 6759, 2019. DOI: 10.1109/JIOT.2019.2911109.

[8] Yu Qiao, Alex Liu and Lejun Zhang, "EESS: An Energy-Efficient Spectrum Sensing Method by Optimizing Spectrum Sensing Node in Cognitive Radio Sensor Networks", Hidawi Journal, 2018. https://doi.org/10.1155/2018/9469106.

[9] Xue Zhang, Xiaozhu Liu, Hooman Samani and Brian Jalaian, "Cooperative Spectrum Sensing in Cognitive Wireless Sensor Networks", International Journal of Distributed Sensor Networks, 2015. https://doi.org/10.1155/2015/170695.

[10] Laila Nassef , Reemah El-Habshi and Linta Jose, "Clustering-Based Routing For Wireless Sensor Networks In Smart Grid Environment", International Journal of Advanced Smart Sensor Network Systems (IJASSN), vol. 8, no. 1, 2018. DOI:10.5121/ijassn.2018.8301.

[11] E. Fadel, M. Faheem, V. Gungor, L. Nassef, N. Akkari, M. A. Malik, S. Almasri and I. F. Akyildiz, "Spectrum-Aware Bio-Inspired Routing in Cognitive Radio Sensor Networks for Smart Grid Applications", Computer Communications, vol. 101, pp. 106-120, 2017. https://doi.org/10.1016/j.comcom.2016.12.020.

[12] Dr. Laila Nassef and Reemah Alhabshi, "Energy Efficient Fuzzy Based Clustering for Cognitive Radio Wireless Sensor Networks", International Journal of Electrical & Computer Sciences IJECS-IJENS, vol. 18, no. 5, 2018, ISSN: 2077-1231 (Online), 2227-2739 (Print).

[13] Dost Muhammad Saqib Bhatti, Nasir Saeed and Haewoon Nam, "Fuzzy C-Means Clustering and Energy Efficient Cluster Head Selection for Cooperative Sensor Network", Sensor, vol. 16, no. 9, 2016. DOI: 10.3390/s16091459.

[14] Dina Tarek, Abderrahim Benslimane, M. Darwish and Amira M. Kotb, "Survey on Spectrum Sharing/Allocation for Cognitive Radio Networks Internet of Things", Egyptian Informatics Journal, 2020. https://doi.org/10.1016/j.eij.2020.02.003.

[15] Muhammad Waqas Khan and Muhammad Zeeshan, "QoS-Based Dynamic Channel Selection Algorithm for Cognitive Radio Based Smart Grid Communication Network", Ad Hoc Networks, vol. 87, pp. 61-75, 2019. https://doi.org/10.1016/j.adhoc.2018.11.007.

[16] Laila Nassef and Reemah Alhebshi, "Secure Spectrum Sensing In Cognitive Radio Sensor Networks: A Survey", International Journal of Computational Engineering Research (IJCER), vol. 6, no. 3, p. 2250 – 3005, 2016.

[17] Pankaj Verma and Brahmjit Singh , "On the decision fusion for cooperative spectrum sensing in cognitive radio networks", Wireless Networks, pp. 2253 - 2262, 2017. https://doi.org/10.1007/s11276-016-1285-0.

[18] Thuc Kieu-Xuan and Insoo Koo, "A Cooperative Spectrum Sensing Scheme Using Fuzzy Logic for Cognitive Radio Networks", KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS, vol. 4, no. 3, pp. 289 - 304, 2010. DOI: 10.3837/tiis.2010.06.006.

[19] Jaewoo So and Wonjin Sung, "Group-Based Multibit Cooperative Spectrum Sensing for Cognitive Radio Networks", IEEE Transactions on Vehicular Technology, vol. 65, no. 12, pp. 10193 - 10198, 2016. DOI: 10.1109/TVT.2016.2536659.

[20] Junhai Luo and Xiaoting He, "A Soft–Hard Combination Decision Fusion Scheme for a Clustered Distributed Detection System with Multiple Sensors", Sensor, vol. 18, 2018. DOI: 10.3390/s18124370.

[21] Yin Mi, Guangyue Lu, Yuxin Li and Zhiqiang Bao, "A Novel Semi-Soft Decision Scheme for Cooperative Spectrum Sensing in Cognitive Radio Networks", Sensor, vol. 19, no. 11, 2019. DOI: 10.3390/s19112522.

[22] Yuanhua Fu, Fan Yang and Zhiming He, "A Quantization-Based Multibit Data Fusion Scheme for Cooperative Spectrum Sensing in Cognitive Radio Networks", Sensor Journal, vol. 18, no. 2, 2018. DOI: 10.3390/s18020473.

[23] Dongho Seo, Hyeongyun Kim and Haewoon Nam, "SDR Implementation of Energy Detection With Non-Uniform Quantization Scheme", in International Conference on Information and Communication Technology Convergence (ICTC), Jeju, South Korea, 2017. DOI: 10.1109/ICTC.2017.8190814.

[24] Huayan Guo, Nima Reisi, Wei Jiang and Wu Luo, "Soft Combination for Cooperative Spectrum Sensing in Fading Channels", IEEE Access, vol. 5, pp. 975 - 986, 2016. DOI: 10.1109/ACCESS.2016.2628860.

[25] Ibrahim Mustapha, Borhanuddin Mohd Ali, Mohd Fadlee A Rasid, Aduwati Sali and Hafizal Mohamad, "An Energy-Efficient Spectrum-

Aware Reinforcement Learning-Based Clustering Algorithm for Cognitive Radio Sensor Networks", Sensor, vol. 15, no. 8, 2015. DOI: 10.3390/s150819783.

[26] Dost Muhammad Saqib Bhatt, Nasir Saeed and Haewoon Nam, "Fuzzy C-Means Clustering and Energy Efficient Cluster Head Selection for Cooperative Sensor Network", Sensor, vol. 16, no. 9, 2016, https://doi.org/10.3390/s16091459.

[27] Hong-Sam T. Le, Hung D. Ly and Qilian Liang , "Opportunistic Spectrum Access Using Fuzzy Logic for Cognitive Radio Networks Hong-Sam", Springer, vol. 18, p. 171–178, 2011, https://doi.org/10.1007/s10776-011-0148-y.

[28] Sylwia Romaszko and Petri Mähönen, "Fuzzy Channel Ranking Estimation in Cognitive Wireless Networks", in IEEE Wireless Communications and Networking Conference (WCNC), Istanbul, Turkey, 2014. DOI: 10.1109/WCNC.2014.6952964.

[29] Giri Prasad Raman and Venkatesan Perumal, "Neuro-Fuzzy Based Two-Stage Spectrum Allocation Scheme to Ensure Spectrum Efficiency in CRN–CSS Assisted by Spectrum Agent", IET Circuits, Devices & Systems, vol. 13, no. 5, pp. 637 - 646, 2019. DOI: 10.1049/iet-cds.2018.5128.

[30] Vasudev Dehalwar, Akhtar Kalam, Mohan Lal Kolhe and Aladin Zayegh, "Compliance of IEEE 802.22 WRAN for Field Area Network in Smart Grid", in IEEE International Conference on Power System Technology (POWERCON), Wollongong, NSW, Australia, 2016. DOI: 10.1109/POWERCON.2016.7754046.

[31] Gyanendra Prasad Joshi, Srijana Acharya and Sung Won Kim, "Fuzzy-Logic-Based Channel Selection in IEEE 802.22 WRAN", Information Systems, 2015. https://doi.org/10.1016/j.is.2014.05.009, vol. 48, pp. 327-332.

[32] Muhammad Waqas Khan and Muhammad Zeeshan, "Fuzzy Inference Based Adaptive Channel Allocation for IEEE 802.22 Compliant Smart Grid Network". Telecommunication Systems, vol. 72, p. 339–353, 2019. https://doi.org/10.1007/s11235-019-00570-y.

[33] Thompson Stephan, Fadi Al-Turjman, K. Suresh Joseph, Balamurugan Balusamy and Sweta Srivastava, "Artificial Intelligence Inspired Energy and Spectrum Aware Cluster Based Routing Protocol for Cognitive Radio Sensor Networks", Journal of Parallel and Distributed Computing, 2010, https://doi.org/10.1016/j.jpdc.2020.04.007, vol. 142, pp. 90-105.

[34] Mir Mehedi Ahsan Pritom, Sujan Sarker, Md. Abdur Razzaque, Mohammad Mehedi Hassan, M. Anwar Hossain and Abdulhameed Alelaiwi, "A Multiconstrained QoS Aware MAC Protocol for Cluster-Based Cognitive Radio Sensor Networks", International Journal of Distributed Sensor Networks, 2015, https://doi.org/10.1155/2015/262871.

[35] Noor Salout, Faroq Awin, Esam Abdel-Raheem and Kemal Tepe, "Combined Fusion Schemes for Cluster-Based Spectrum Sensing in Cognitive Radio Networks", in IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Louisville, KY, USA, USA, 2018, DOI: 10.1109/ISSPIT.2018.8642636.

[36] V. C. Gungor and M. K. Korkmaz, "Wireless Link-Quality Estimation in Smart Grid Environments", International Journal of Distributed Sensor Networks, 2012, https://doi.org/10.1155/2012/214068.

[37] J. Wang, I.-R. Chen, J. J. Tsai and D.-C. Wang, "Trust-Based Mechanism Design for Cooperative Spectrum Sensing in Cognitive Radio Networks", Computer Communication, vol. 116, pp. 90-100, 2018, https://doi.org/10.1016/j.comcom.2017.11.010.

[38] Ju Ren, Yaoxue Zhang, Qiang Ye, Kan Yang, Kuan Zhang and Xuemin Sherman Shen, "Exploiting Secure and Energy-Efficient Collaborative Spectrum Sensing for Cognitive Radio Sensor Networks", IEEE Transactions on Wireless Communications, vol. 15, no. 10, pp. 6813 - 6827, 2016. DOI: 10.1109/TWC.2016.2591006.

Modeling and Assessing the Power Consumption Behavior of Sensor Nodes using Petri Nets

Alaa E. S. Ahmed[1, 2]

College of Computer and Information Sciences
Al Imam Mohamed Ibn Saud Islamic University (IMSIU), Riyadh, Kingdom of Saudi Arabia[1]
Department of Electrical Engineering, Faculty of Engineering at Shoubra, Benha University, Egypt[2]

Abstract—Power Consumption is an influential and important concern in Wireless Sensor Networks (WSNs). Assessing and determining the impact of the power behavior of sensor nodes is an inclusive concern that should be executed at the network pre-deployment phase rather than post-deployment phase. Providing an accurate power modeling improves the development process of WSN applications and protocols. This paper introduces the use of Colored Petri Nets (CPNs) to model the power behavior and relations among different sensor node components when operated in an event driven environment. The objective is to figure out the overall power behavior of the nodes by considering power consumed accompanied with different states transition at a specific packets arrival and service rate values. Colored Petri Nets is a modeling language employed for validating and evaluating concurrent and distributed systems. The introduced model is beneficial since it provides the network designer with a way to determine surrogate designs or to check the compatibility of an existing power model behavior. The proposed model has been validated through the comparison of the power behaviors of two sensor nodes Mica2 and Telos. The results demonstrate the efficiency of using the proposed model to draw and analyze the WSNs energy consumption behavior.

Keywords—*Wireless sensor networks; power modeling; event-trigger; colored petri net; WSN protocols; distributed systems*

I. INTRODUCTION

Wireless Sensor Networks (WSNs) are made up of spatially distributed motes with sensing, processing and communicating abilities. Maximizing the network life time is fundamental issue in developing WSN applications. Network life time is strongly related to the problem of optimizing the amount of consumed power throughout different stages executed by the network. It is the responsibility of the protocols or application designers to employ best strategies that lead to minimum power consumption. Power consumptions evaluation is considered as prerequisite for the deployment process to be able to avoid critical issues such as power consumption bottleneck, short network lifetime or higher cost maintenance due to replacing nodes with drained batteries.

Saving energy consumption in wireless sensor network can be achieved through the design of efficient strategies for scheduling the active and sleep moods of the node. Through these strategies, a time threshold value needs to be determined to allow optimal switching between the CPU sleep and active moods [1]. The benefit of this is to keep the cost of sensors replacement or maintenance as low as possible. However, having a power efficient designs need a good understanding of the power consumption behavior of the sensor nodes since they form the basic unit for constructing the WSNs structure. Understanding power consumption behavior can be addressed whether through real implementation, simulation or modeling.

The most accurate strategy of evaluating the power consumption is to use real measurement for the operated hardware. This could be done by applying battery level check pointing at regular period of time bases [2]. The cons of this procedure is the high cost due to the necessity of reproducing the environment, the huge number of nodes involved in the network and finally errors resulted from the measurement read by the human.

Modeling is another alternative way to evaluate the power consumption. This modeling process could be executed either analytically or by simulation. Even though results achieved from the modeling process is less accurate than those achieved by real measurement, it avails a flexible tool that help designers to evaluate and check various scenarios without affecting the actual environment. Modeling is considered an efficient way that enables the WSN designer to check and analyze these models in case of different states transitions [3, 4].

According to the previous argument, in this paper a method for modeling the power consumption behavior of sensor nodes is presented. A set of models has been defined to express power consumption accompanied with different state transition occurs within the sensor node component which are CPU, Transmission unit and sensing unit. The proposed models are implemented using Colored Petri Net [5–7] which represent a well-structured extension of Petri net that can be used in implementing and analyzing concurrency and asynchronous behaviors of distributed systems such as WSNs.

The remainder of the paper is organized as follows: In Section 2, the related work is introduced. Modeling bandwidth allocation scheduling strategies with CPNs is defined in Section 3. Section 4 shows a discussion for the simulation results. Finally, the paper ends the con collusion and future work in Section 5.

II. RELATED WORK

Technology development and user requirements both are increasing in such a way that necessitate the designers of limited energy sensor devices to consider many challenges and requirements such as energy consumption, reliability, bandwidth allocation and data aggregation [8][9] [10]. Finding,

modeling and analyzing strategies for predicting the WSN lifetime and optimizing the consumed energy is essential.

In literature, many approaches have been presented for modeling the power and energy consumption behaviors in embedded systems. In [11], Hyprid Automata is used in modeling TinyOs. TinOs is a component-based operating system designed for low power wireless devices. The provided automata model is used for analyzing the Node power dissipation based on its geographical location in the field area. Another model based on automata is presented in [12]. The aim of the generated model is to be used as a test beds for sensor network to provide priori information about the expected power consumed by the network.

In [13], the active and sleep operation modes are modeled using Markov models. The authors provided a Steady state equation that determines the number of packets needed when applying different modes of operation which in turn give indication about the system energy performance. Authors in [14] presented a method for predicting the energy level of nodes in WSNs using the Hidden Markov Model. They build the model based on the assumption that the energy level per nodes is divided into several levels and the state of these levels at any time is considered the random variable that can be used by Markov chain. The generated model performs training process while the WSN is executing their tasks so that an adjusted energy consumption prediction can be determined.

Another Markov based model is presented in [15]. They provide a model for the sensor node. They estimated the average life time in both triggers driven and duty cycle driven modes for different packets arrival time. The estimated value takes into consideration the effect of both communication and state transition overheads. Modeling using Markov chains is beneficial only in case the appropriate equation that describes the system is generated. However, the task of getting the appropriate equation can be time consuming and difficult in some cases.

Petri net is considered as an effective tool that can be used to model complex concurrent systems. Petri net is a directed graph where nodes consist of two components places and transitions. The directed arcs always connect places with transitions. The example shown in Fig. 1 consist of two places P0 and P1 and one transition T1. Places may contain tokens which used in firing the next transition connected to that place. In the figure, Place P0 contains one token that enable the transition T0. The enabled transition will fire either instantly or according to different timing parameter and the token will be moved to the output place P1. This behavior can be repeated in different ways to model the behavior of systems that utilize the flow of tokens to simulate a control mechanism adopted by different system processes.
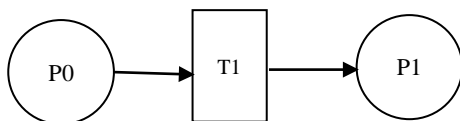
Fig. 1. Simple Petri Net Example.

Colored Petri Net (CPN) [16] is an executable model that combines the abilities of Petri Net and the abilities of high-level programming languages. It also provides the ability of providing time base models that can be used to simulate stochastic behaviors that can be utilized in evaluating concurrent systems performance. In literature, it has been used in modeling different aspects related to WSN. For example, in [17], authors used colored Petri net to validate and evaluate the WSN EQ-MAC protocol and they were able to validate its correctness. Average delay and packets delivery ratios are the two metrics used in evaluating their proposed model. Also, in [18] authors modeled an attack-resistant and efficient localization scheme using Petri net. They showed that there are no insecure states are reachable by adopting the state analysis methods. In [19] Petri nets were used to construct a formal model for the wireless sensor network. The model is utilized in testing and simulating the behaviors of different applications used by the network. They looked at the effect of QoS Parameters on the system from the application point of view. In [20] colored Petri nets are used to model the Wireless sensor network in order to capture the effect of different data aggregation techniques on the overall network performance. In this work the node behavior is represented by the M/M/1 queuing model. Results and analysis are implemented using the CPN tool.

In this paper, we handle the power consumption from the hardware point of view by providing a colored Petri net model that simulates the behavior and the interaction between these components.

III. MODELING OF SENSOR NODE USING COLORED PETRI NET

In this section, we propose a colored Petri net-based model to characterize the power consumption behavior of sensor nodes. The model utilizes the component-based architecture of node hardware components. As in Fig. 2, there are three main components included in the operation of the sensor node, Processing Unit (PU), Sensing Unit (SU) and Transceiver Unit (TU). These components operated in an event driven mode because every transaction executed by the sensor node was originally generated as a reaction to external events. Events are characterized by its time of occurrences which may follow different distributions. Events are classified into two categories source-driven or sink-driven. The first type "source-driven" is generated by either the surrounding environment or the neighbor nodes and then captured by the sensing unit of the sensor node. The second type "sink-driven" is generated by a query generated by the sink node. The proposed work deals only with the source-driven events type.
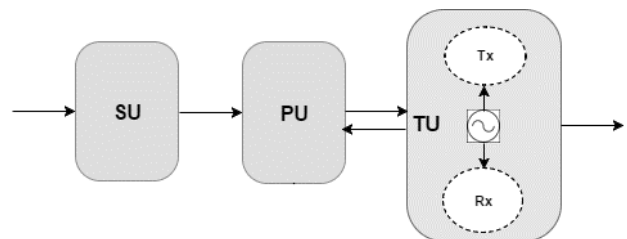
Fig. 2. Component in Operations of the Sensor Node.

A. State-Event-Transition Description

Here, a deeper look in the model dynamics is presented. These dynamics is specified in terms of sequencing and timing and using the state diagrams shown in Fig. 3. All components stay in a certain state until it is stimulated to change by an event. Each component passes over different power states. For example, the SU unit has two states on and off. The sensing process cycle starts when an event generated from the environment and read by the sensing unit (SU). Then, the event is forwarded to be processed by the processing unit (PU). The processing unit has three states, idle, sleep and run. Later, the event is either transmitted to or received from another node. The transceiver unit proceeds in five states, off, idle sleep, Tx and Rx which are connected according to the paths indicated in Fig. 2.

B. Modeling Power Consumption of a Sensor Node

In the Petri net model, places are used either to represent the power state of a specific component of the sensor node or to emulate a particular event-driven situation. In the other hand tokens are also employed to indicate either events or communication channel allocations. When tokens migrate among different places, it causes different system states to occur. The node model is stimulated by receiving sensed data from the environment emulated by the workload generator model shown in Fig. 4. When transition T1 fires, data is stored in Place "New Arrival" and a token goes back to the "Data Sampling" place to allow transition T1 to fire again to receive another generated packet. The process of Firing transition T1 is repeated randomly and according to an exponential distribution attached to the arc E1 (Table I). After the packets generated, they are formatted using a function attached to arc E2 (Table I). The packet format contains a record with three fields: Data reading "Rvalue", Initial consumed power "RPower" and "Type" which indicates whether the processed packet is to be received or to transmited mode. Table I shows arcs inscriptions used by the packets generator module.
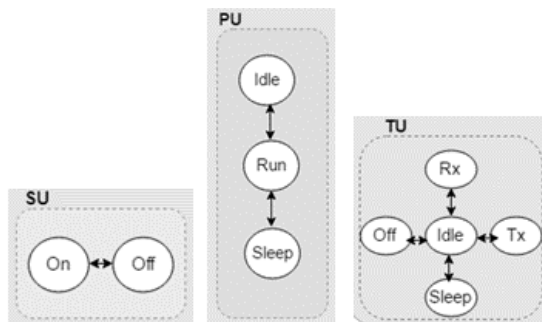
Fig. 3. Event-Driven State Transition Diagram of Sensor node Components.
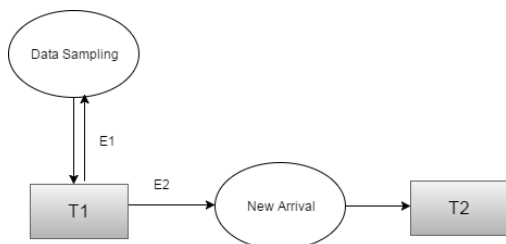
Fig. 4. Packets Generator Module.

TABLE I. SHOW THE ARC INSCRIPTIONS IN PACKETS GENERATOR MODULE

ARC name	inscription
E1	()@+expTime(ArrivalTime)
E2	fun newX() = {Rvalue = randRange, Rpower = 0, Type =0}

Fig. 5 shows the Petri net model of SU unit where initially the unit is in "off" state, and there is a token reside in place "Off-State". When data packets arrived at transition T2, it is fired and a token is put in "On State" place to indicate that the unit is "on" state. As a result, T3 transition fires and forward the received data packets to the PU module through the arc E7. A token is returned back to the "Off State" place to convert the unit to "off" state again and then wait for next packets arrival.
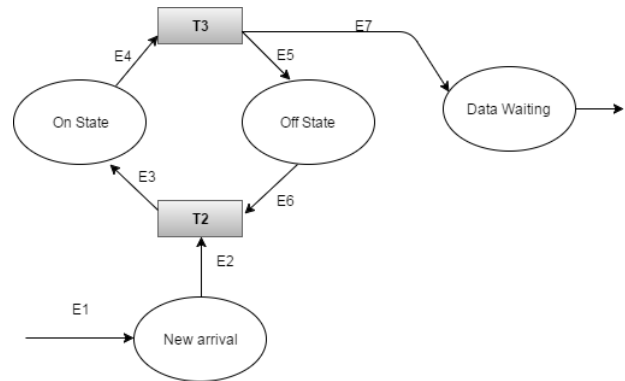
Fig. 5. SU Model.

Fig. 6 illustrates the PU model which has nine places and six transitions. The model is an implementation to the different states mentioned in Fig. 2. Initially, the unit is in "sleep" state which is implemented by having a token reside in place "Sleep". The PU unit switches to the "RUN" state to process the data and a token is generated at place "ON". PU stays in the "RUN" state as far as there are tokens exist at place "Buffer PU". The "Counter" place is used to indicate the number of tokens on the buffer. Transition T9 is fired when the "Counter" place indicates zero packets in the buffer and PU switch to "sleep" state to preserve power.

The PU model executed using the following steps:

1) When data packets arrive at transition T4, it fires and tokens are generated and forwarded to both "Buffer PU" and "Only One" places. Since the PU was previously in the sleep state, transition T7 fires and a token will be generated and move to the "ON" place to indicate that the PU unit is now switched from sleep to Run state. Transition T7 has a guard function used to add up the delay and the power consumed values due to the move from "Sleep" to "Run" state. These values are determined from the sensor node hardware specification data sheet as in Table III.

2) Now, there is a token in both "Buffer PU" and "ON" places and, the "idle" place has another token by default. This combination enables the transition T5 and a token will be moved from the "idle" place to the "run" place to announce the start of the Run state. As a consequence, Transition T8 fires and triggers the execution of a function that simulates the PU

service rate. The service rate is represented in terms of a randomly generated delay with an exponential distribution. After this delay is ended up, two tokens will be generated. The first token is moved back to the "idle" place to allow the next arrived packet to be processed. The second moves to the TU unit and per its type (Send or Receive), the TU will perform the appropriate action.

3) In case another packet arrived and the PU is still ON, the previous steps will be repeated.

Fig. 7 illustrates the TU model which has nineteen places and thirteen transitions. The TU unit receives input from two sources. The first is the data that comes from the PU unit after has been sensed by the SU. In this case, the TU unit acts as a transmitter to send the data to the next node. The second is the data comes directly from other neighboring nodes to be processed by the PU and then retransmitted to other neighboring nodes.
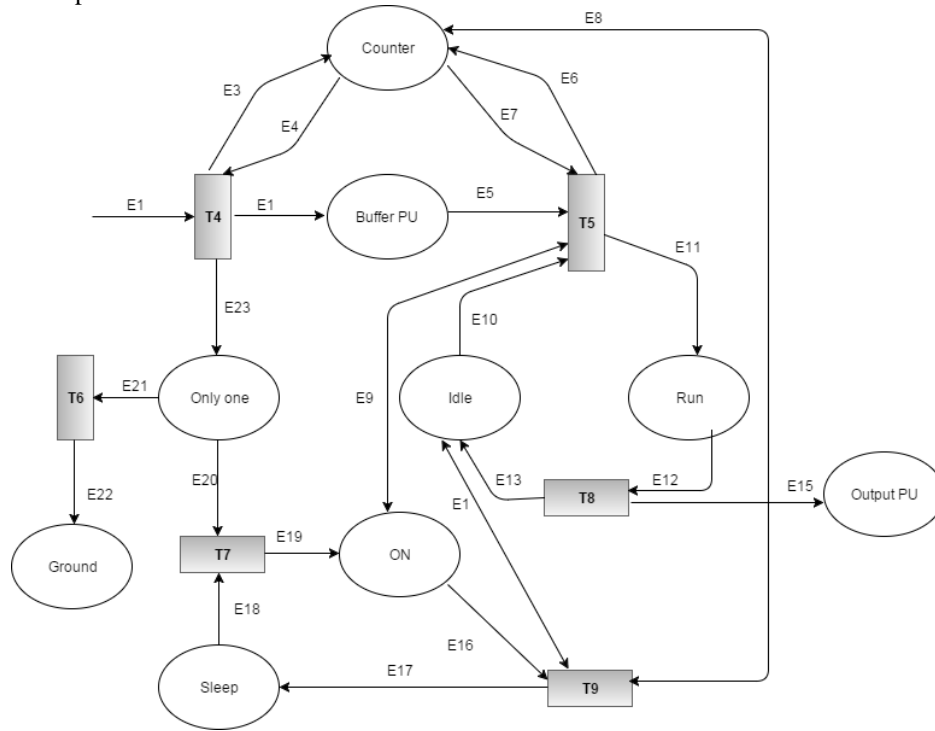
Fig. 6. PU Model.
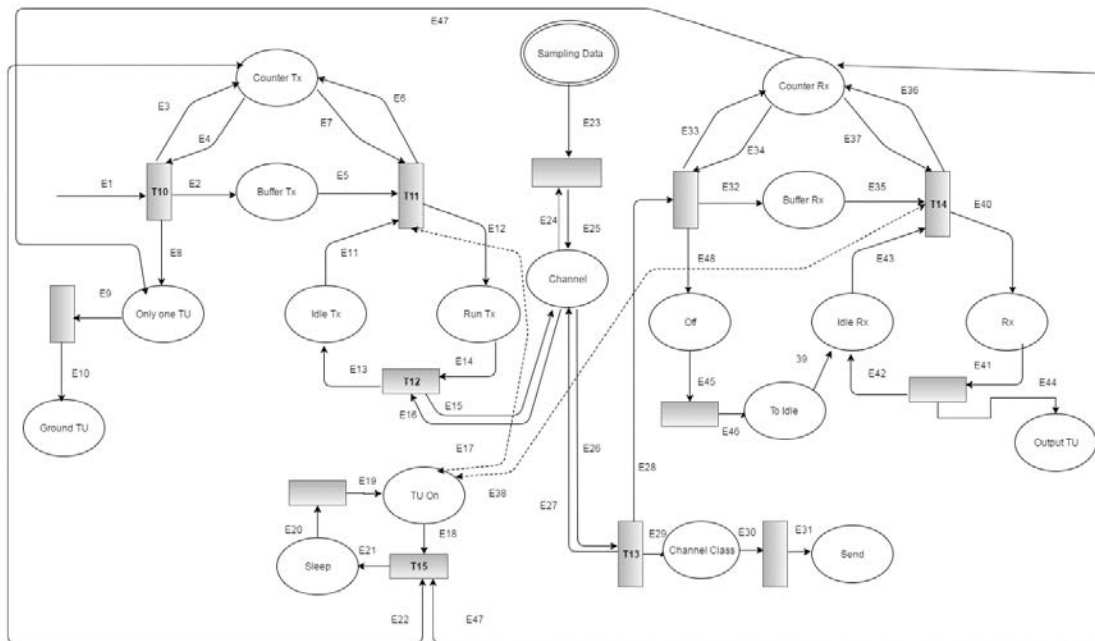
Fig. 7. SU Model.

The TU model executed using the following steps:

1) When data packets comes from the PU unit, Transition T10 fires and tokens generated and moved to the place "Buffer Tx". The place "Counter Tx" is used to count the number of packets received by the TU to control the firing rate of transition T11. As long as there is a token in the "buffer Tx" place, T11 is enabled. There are default tokens in the place "idle Tx" to indicate that the TU is initially in the idle state. When T11 fires, a token move from the "idle Tx" place to the "Run Tx" place. In other words, the TU switches from "idle" to "Tx" or transmission state. Transition T12 fires and a randomly exponentially generated delay is added up to the total delay of the packets. Also, power consumed by the Tx is added to the total power consumed due to the packet transmission.

2) Now, Tokens generated and moved to the places "idle Tx" and "Channel. "Channel" place is used to implement the channel interface used in transmitting the packets. Tokens moved to the "idle_tx" are used to enable the receiving of other new arrived packets. While tokens moved to the place "Channel" place are characterized by type "T" that stands for "Transmit" so that they will be forwarded outside the node. The place "Channel" receives another input token as a result of receiving packets from other neighboring nodes. Packets related to these tokens are marked as "R" that stands for "Receive" to be forwarded to the PU unit. Packets Types classifications is executed at transition T13. The "T" type tokens are moved to the places "Channel Class" and "send". The "R" type tokens moved to the "Buffer Rx" place to enable the Rx state.

3) When "R" types tokens exist in the "buffer Rx" place, the "Counter Rx" indicates their counts and at the same time fires transition T14. Since by default the receiving part of the TU is idle state, i.e. there are tokens in the "idle Rx" place, place "Rx" fires and tokens move to the "Output Tu" place.

4) TU unit switch from the idle state to the sleep state when there are no tokens in both "Buffer Rx" and "Buffer Tx". This transition implemented in the model using

5) Transition T15 that will be enabled when both the counter Tx, counter Rx is zero and the token in the "TU On" place move to the "Sleep" place.

6) The transition from the "off" state to the "idle" state is implemented in the model using the places "off" and "To idle". When tokens received in the "buffer Rx", the token is generated and is passed to the "idle" place which consequently moved to the "to idle" place and finally reach the "idle Rx" place to indicate that the current state is the idle state. This step is done to prepare the node to switch to the idle to either Rx or Tx.

IV. POWER CONSUMPTION EVALUATION OF THE PROPOSED MODEL

In this section, we evaluate the performance of the proposed model using the CPN tool. The evaluation process is conducted by considering two different nodes architecture telos and Mica 2. We call the first one SN1 and the second SN2.

SN1 is composed of microcontroller-MSP430F4794 to act as a PU, a transceiver-CC2420 to act as a TU, and a temperature sensor called Dallas Semi.DS1820 to act as a SU. SN2 is composed of ATmega 128L to act as a PU, component CC1000 to act as a TU and the same temperature sensor used in SN1. Table II summarize he components parameters and the consumed power due to states transition [21].

The simulation parameters used in the evaluation process in case of different workload conditions are listed in Table III.

TABLE II. PERFORMANCE PARAMETERS AND POWER CONSUMPTION AND STATE TRANSITION TIME PARAMETERS FOR SN1 AND SN2

Node Architecture			Value	
Component		*State*		
Sensor Node			*SN2*	*SN1*
Component Power consumption	PU	Idle	6 mW	3.9 μW
		Run	15 mW	1.2 mW
		Sleep	24 μW	0.3 μW
	SU	ON	3 mW	3 mW
		Off	0 W	0 W
	TU	Idle	28.8 mW	14.058 mW
		Tx	49.5 mW	62.04 mW
		Rx	28.8 mW	57.42 mW
		Sleep	3 μW	66 μW
		Off	0.6 μW	0.066 μW
Time of Transitions between states	PU	Idle to Run	5 μs	6 μs
		Run to Idle	5 μs	6 μs
		Idle to Sleep	40 μs	50 μs
		Run to Sleep	40 μs	50 μs
		Sleep to Run	65 ms	80 ms
	SU	On to Off	36 μs	36 μs
		Off to On	400 ms	400 ms
	TU	Tx to Idle	2 μs	2 μs
		Idle to Tx	270 μs	192 μs
		Rx to Idle	2 μs	2 μs
		Idle to Rx	250 μs	192 μs
		Sleep to Idle	200 μs	600 μs
		Idle to sleep	250 μs	192 μs
		Off to Idle	2.2 ms	1 ms

TABLE III. SIMULATION PARAMETERS

Parameter Configuration	Value	
Sensor Node	SN2	SN1
Simulation Time	50 sec	50 sec
Packet arriving distribution	1 to 5 sec	1 to 5 sec
Service time in PU	Exponential distribution ☐=1	Exponential distribution ☐=1
Sensor sampling period	5 sec	5 sec

The simulation process is conducted by computing the node overall power consumption by adding up different states power during the simulation time. As time proceeds, the average number of tokens deposited in various places is computed to indicate the percentages of time that packets occupy different states. The power consumed per state is computed by multiplying the computed percentage by the state power values listed in Table II. The summation of power values of different states gives the node overall value. Packets arrival rate, components services rates and time delay between different states affect and control the number of tokens exist in places and consequently the value of the computed percentages.

Results in Fig. 8, 9 and 10 evaluate the effect of multiple workloads on the power consumed by the node components on different states. As the packets inter-arrival time increases the arrival-rate decreases and the power consumption becomes lower. These results can be explained as when sensor nodes exist near active areas of the WSN such as the sink node and the cluster heads, the inter-arrival rate of packets arrival and the energy consumed increases. These experiments could be used to evaluate the effect of changing the sampling rates on the power consumed by the sensor node components. In Table IV, the power evaluation in SN1 with different inter-arrival time is shown.

TABLE IV.   SIMULATION RESULTS OF POWER CONSUMPTION IN SN1

Component	States	Arrival Time (1 sec)	Arrival Time (3 sec)	Arrival Time (5 sec)
PU	Idle	0.8	0.41	0.22
	Run	180	86.4	48
	Sleep	0.04	0.021	0.021
TU	Idle	3753	1743	815
	Tx	9306	4466	2481
	Rx	6890	3100	1148
	Sleep	17.5	8.18	3.83
	Off	0.007	0.003	0.001
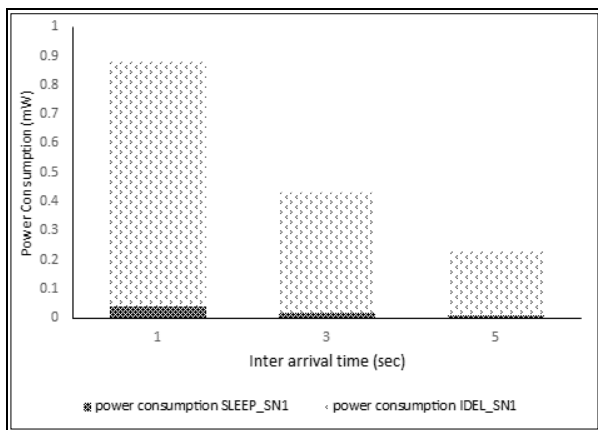
Fig. 8.   PU Power Simulation for SN1 (SLEEP and IDLE States).
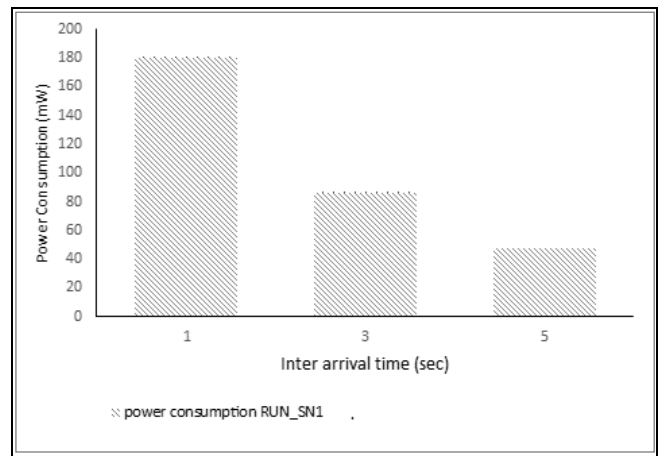
Fig. 9.   PU Power Simulation for SN1 (Run State).
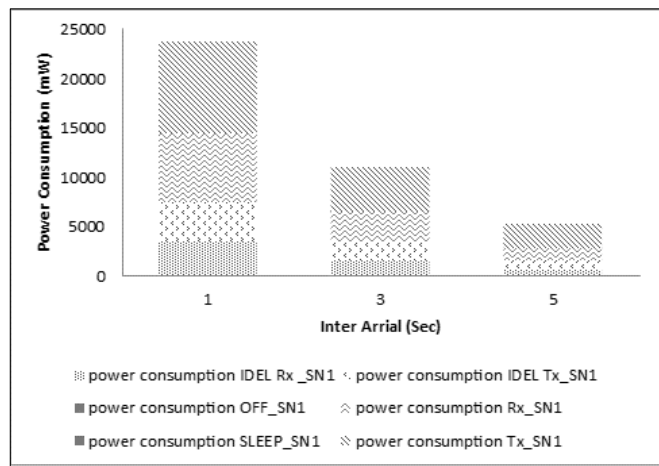
Fig. 10. TR Power Simulation for SN1.

To evaluate the effect of node architecture on power consumption level, the experiment repeated for another node architecture termed SN2. Fig. 11 and 12 show the power consumed in all components whereas the inter arrival rate increases the power consumed by the node increases.
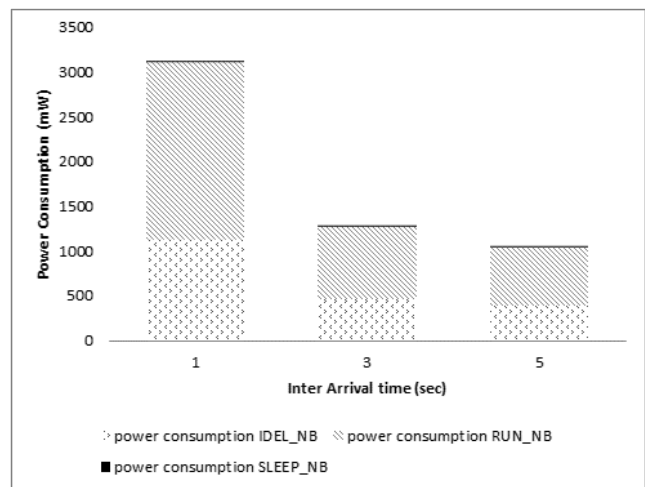
Fig. 11. PU Power Simulation for SN2.
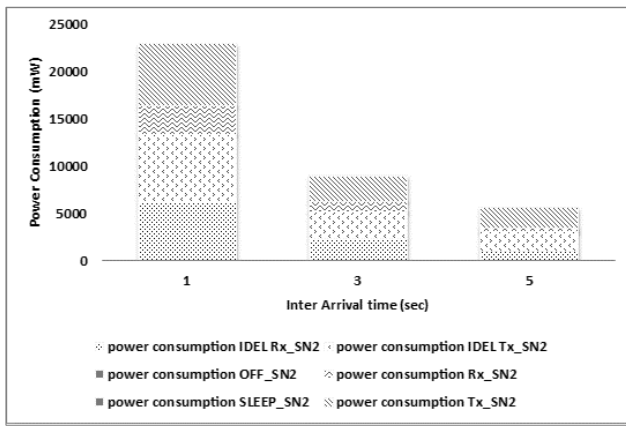
Fig. 12. TU Power Simulation for SN2.
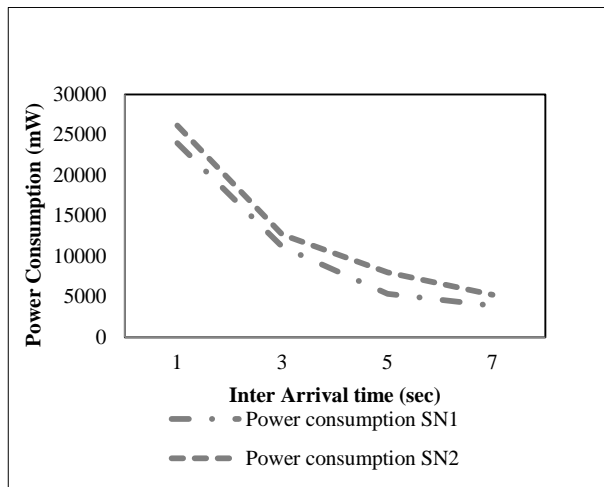
Fig. 13. Power Consumption in SN1 and SN2.

Table V shows the values of these results under different simulation conditions. Comparing the power consumption of SN1 and SN2, Node SN1 has less power consumption than the SN2 as shown in Fig. 13. This is due to the main functional components of SN1 have more ability to operate in low power situation than SN2. The model can be used to characterize the power behavior of various sensor nodes architectures and guide the network designers toward using the architecture that fit his needs.

Assuming another WSN node (termed SN3), which has different PU component ARM SA-1100. Table VI summarize the components parameters and the consumed power due to states transition which obtained from the datasheets of SN3 [21].

In Table VII, Fig. 14 and Fig. 15, the power evaluation of SN3 with different inter-arrival time is shown. From these results, we notice that most the power is consumed in the TU unit However, in Node SN1, most of the power consumed in the PU unit. This result indicates the ability to use the proposed model to help WSN designers in understanding the power consumption behavior for different sensor node architectures. Comparing the power consumption in SN1 and SN3, Node SN1 has less power consumption than the SN3 as shown Fig. 16.

TABLE V. SIMULATION RESULTS OF ENERGY CONSUMPTION IN NS2

Component	States	Arrival Time (1 sec)	Arrival Time (3 sec)	Arrival Time (5 sec)
PU	Idle	1128	468	390
	Run	1980	810	660
	Sleep	2.73	1.2	1.05
SU	On	90	54	102
	Off	0	0	0
TU	Idle	6710	2592	1555
	Tx	6543	2673	2178
	Rx	7113	2822	2016
	Sleep	0.69	0.27	0.16
	Off	0.06	0.022	0.007

TABLE VI. PERFORMANCE PARAMETERS AND POWER CONSUMPTION AND STATE TRANSITION TIME PARAMETERS FOR SN3

Node Architecture			Value
Component		*State*	
Sensor Node			*SN3*
Power consumption	PU	Idle	50 mW
		Run	230 mW
		Sleep	0.16 mW
	SU	ON	3 mW
		Off	0 W
	TU	Idle	14.058 mW
		Tx	62.04 mW
		Rx	57.42 mW
		Sleep	66 µW
		Off	0.066 µW
the transition time between states	PU	Idle to Run	10 µs
		Run to Idle	10 µs
		Idle to Sleep	90 µs
		Run to Sleep	90 µs
		Sleep to Run	160 ms
	SU	On to Off	36 µs
		Off to On	400 ms
	TU	Tx to Idle	2 µs
		Idle to Tx	192 µs
		Rx to Idle	2 µs
		Idle to Rx	192 µs
		Sleep to Idle	600 µs
		Idle to sleep	192 µs
		Off to Idle	1 ms

TABLE VII.   SIMULATION RESULTS OF POWER CONSUMPTION IN SN3

Component	States	Arrival Time (1 sec)	Arrival Time (3 sec)	Arrival Time (5 sec)
PU	Idle	9150	5600	2750
	Run	29440	17940	8740
	Sleep	17.92	11.2	5.76
SU	On	108	126	60
	Off	0	0	0
TU	Idle	3092.76	1602.612	787.248
	Tx	7941.12	4839.12	2357.52
	Rx	5397.48	2181.96	1148.4
	Sleep	14.52	7.524	3.696
	Off	0.006204	0.002508	0.00132
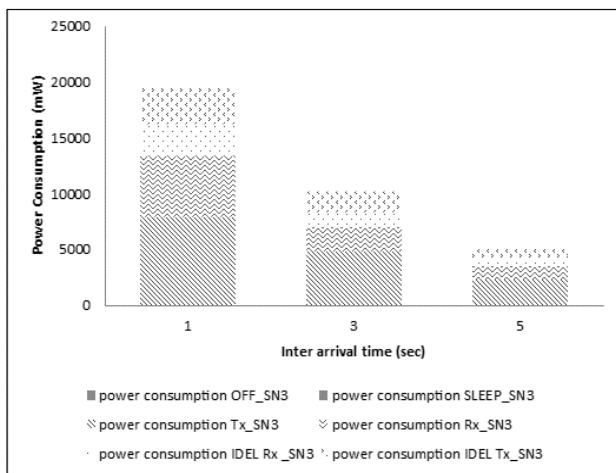
Fig. 16. Power Consumption in SN1 and SN3.

V. CONCLUSION

In this paper, we provided a petri net model of power consumption of WSN node that is described through the event-driven. The model has been evaluated by emulating the power behaviors of two real sensor nodes Mica2 and Telos. The results demonstrated the efficiency of using the proposed model to draw and analyze the WSNs power consumption behavior. Thus, the provided results and analysis showed that the proposed model is valid and easy to debug and evaluate the power consumption for different WSN node architectures. As a future work, the proposed model can be considered as an initial template where other performance parameters could be employed. For example, the effect of data aggregation, compressive sensing and bandwidth allocation on power consumption can be modeled and evaluated.

Fig. 14. PU Power Simulation for SN3.

Fig. 15. TU Power Simulation for SN3.

REFERENCES

[1]   J. Liu and P. H. Chou, "Idle energy minimization by mode sequence optimization," ACMTransactions on Design Automation of Electronic Systems, vol. 12, no. 4, article 38, 2007.

[2]   F. A. Katsriku, J. Abdulai, K. S. Adu-Manu, and F. K. Banase, "Prolonging the Lifetime of Wireless Sensor Networks: A Review of Current Techniques,", Wireless Communications and Mobile Computing, vol. 2018, 2018.

[3]   H. P. Xie, H. Y. Zhou, D. C. Zuo, and P. Zhou, "Energy optimization and modeling in wireless sensor networks: a survey," Computer Science, vol. 10, pp. 15–20, 25, 2012.

[4]   N.R. Roy, P. Chandra, "Energy dissipation model for wireless sensor networks: a survey," International Journal of Information Technology, 2019.

[5]   K. Jensen, L. Kristensen, "Coloured Petri Nets: Modelling and Validation of Concurrent Systems," Springer: New York, NY, USA, 2009.

[6]   K. Jensen, "Coloured Petri Nets: Basic Concepts, Analysis Methods and Practical Use," Springer Verlag: London, UK, 1995.

[7]   K. Jensen, L. M. Kristensen, "Colored Petri nets: a graphical language for formal modeling and validation of concurrent systems," Communications of the ACM, vol. 58, no. 6, pp. 61-70, 2015.

[8]   K. Alshammari and A E. S. Ahmed, "An efficient approach for detecting nodes failures in wireless sensor network based on clustering," International Symposium on Networks, Computers and Communications (ISNCC), Marrakech, pp. 1-6, 2017.

[9]   A. Dˆamaso, D. Freitas, N. Rosa, B. Silva, and P. Maciel,"Evaluating the power consumption of wireless sensor network applications using model," Sensors vol.13, no.3, pp. 3473–3500, 2013.

[10] E. Troubleyn, J. Hoebeke, I. Moerman, and P. Demeester, "Broadcast aggregation to improve quality of service in wireless sensor networks," International Journal of Distributed Sensor Networks, vol. 10, no. 3, 2014.

[11] S. Coleri, M. Ergen, and T. J. Koo, "Lifetime analysis of a sensor network with hybrid automata modelling," in Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications (WSNA '02), pp. 98–104, ACM, New York, NY, USA, September 2002.

[12] S. Kellner, M. Pink, D. Meier, and E. O. Blaß, "Towards a realistic energy model for wireless sensor networks," in Proceedings of the 5th Annual Conference on Wireless on Demand Network Systems and Services (WONS '08), pp. 97–100, January 2008.

[13] Y. Zhang and W. Li, "An energy-based stochastic model for wireless sensor networks," in Proceedings of the Wireless Sensor Network, vol. 10, pp. 322-328, 2011.

[14] P. Hu, Z. Zhou, Q. Liu, and F. Li, "The HMM-based modeling for the energy level prediction in wireless sensor networks," in Proceedings of the 2007 2nd IEEE Conference on Industrial Electronics and Applications (ICIEA '07), pp. 2253–2258, May 2007.

[15] D. Jung, T. Teixeira, A. Barton-Sweeney, and A. Savvides, "Model-based design exploration of wireless sensor node lifetimes," in

Proceedings of the 41rth European Conference on Wireless Sensor Networks (EWSN '07), 2007.

[16] K. Jensen, L. M. Kristensen, and L. Wells. "Coloured Petri Nets and CPN Tools for modelling and validation of concurrent systems," International Journal on So.ware Tools for Technology Transfer vol. 9, no.3-4, pp. 213–254, 2007.

[17] J. Ben-Othman, S. Diagne, L. Mokdad, B. Yahya, "Performance evaluation of a hybrid MAC protocol for wireless sensor networks," international conference on Modeling, analysis, and simulation of wireless and mobile systems, Pages 327-334 , ACM, 2010.

[18] D. He, L. Cui et, al. Design and Verification of Enhanced Secure Localization Scheme in Wireless Sensor Networks. IEEE Transactions on Parallel and Distributed Systems, vol. 20, pages 1050 – 1058, IEEE, 2009.

[19] X. Fu, Z. Ma, Z. Yu, and G. Fu, "On wireless sensor networks formal modeling based on petri nets," in Proceedings of the 7th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '11), pp. 1–4, 2011.

[20] A. E. S. Ahmed and M. E. A. Ibrahim, "Colored Petri Net Models for Clustered and Tree-Bases Data Aggregation in Wireless Snsor Networks", International Journal on Information Technologies & Security, vol 10, no. 3, 2018.

[21] J. Li,1,2 H. Y. Zhou, et, al, "Energy Consumption Evaluation for Wireless Sensor Network Nodes Based on Queuing Petri Net" , International Journal of Distributed Sensor Networks, Vol. 10, no. 4, pp. 1-11, 2014.

The TPOA Telecentre: A Community Sustainable Telecentre Architecture

Chong Eng Tan[1], Poline Bala[2], Sei Ping Lau[3], Siew Mooi Wong[4]

Institute of Social Informatics and Technological Innovations[1, 2, 4]
Faculty of Computer Science and Information Technology[3]
Universiti Malaysia Sarawak, Kota Samarahan, Malaysia

Abstract—This paper presented the telecentre implementation for the Orang Asli villages in remote rural areas under the Telecentre Program for Orang Asli (TPOA). TPOA telecentre architecture aims to assist the achievement of a rural community sustainable telecentre through innovation and strategic adoption of ICT technology. Lessons learned from our past telecentre experience have outlined various challenges in the technical aspects of the telecentre implementation and operation. The TPOA telecentre ICT architecture has been designed to address the outlined issues hence producing a smoother telecentre operation that enables the rural communities to self-sustain their own telecentres. The technical support for the remote rural telecentre can be very expensive and impractical due to the extreme physical access condition. Hence, the rural communities themselves have to carry out the support and maintenance to sustain the operation of the telecentre. The TPOA telecentre architecture has enabled a relatively friendly to operate ICT platform in order to assist and make it possible for the Orang Asli to sustain, support, and maintain the telecentre operation.

Keywords—Telecentre; sustainability; TPOA; telecentre architecture; ICT4D; rural development

I. INTRODUCTION

Telecentre has been a recognized model for bridging the digital divide that encourages the use of ICT tools as the medium to access resources beyond physical reach. The telecentre is particularly important to remote rural communities on social development. The telecentre model implemented in Malaysia started as early as 2002 in the Kelabit Highland of Bario, Sarawak called the eBario Telecentre, followed by the replication of the eBario Telecentre model for other remote rural communities in Long Lamai, Ba'Kelalan, Buayan, and Larapan. Since the implementation of eBario, the telecentre ICT architecture has evolved and improved based on lessons learned along the way. The telecentre model covers a much wider scope beyond its associated physical ICT systems and services. Telecentre's roles and programs, sustainability, and the social issues of the community it serves are among the important factors in overall telecentre development. Telecentre comes in many forms and different names in different countries to serve some special needs of the rural communities. In this paper, we focus on the technological development of the telecentres and showcase our recent telecentre implementations for the Orang Asli communities situated in the remote rural of Peninsular Malaysia. The Telecentre Program for Orang Asli (TPOA) was the third-generation telecentre implementation

after the eBario telecentre and the second-generation eLamai, eBa'Kelalan, eBuayan and eLarapan by researchers from the Universiti Malaysia Sarawak (UNIMAS). Experiences gained from the previous telecentre implementations have laid a strong foundation to continue to innovate the ICT technology used for TPOA telecentres.

The TPOA telecentre implementations have moved the ICT team to innovate new ICT architecture that would overcome some of the limitations and shortcomings encountered in the past telecentre implementations. It is crucial to review these technical challenges which will act as fundamental expectations for the next telecentre implementation. All these challenges have direct and indirect impacts toward telecentre sustainability especially for telecentres in remote rural areas that do not have a lot of local resources and the luxury to receive lots of physical technical support from the urban.

The TPOA telecentre architecture is designed to increase its technical sustainability allowing the local community to self-sustain the telecentre. Design considerations that have been taken into consideration include energy efficiency of ICT equipment, friendly user experience, gentle learning curve, ICT system robustness, seamless device interoperability, reduce dependency on Internet bandwidth, and protection against negative impacts of the digital revolution that brought in together through the telecentre implementation. Innovated from the list of considerations, the TPOA telecentre architecture and its ICT components are presented.

The setup and operation methodology of the telecentre is crucial to influence the long-term sustainability of the telecentre in remote rural areas. An effective setup approach and operation mechanism will strengthen the concept of community-driven and community sustained telecentre. Important concepts such as the sense of belonging, the community elected management committee and caretakers, community-driven operation schedule, and community communal work towards self-sustainability were adopted to further promote community sustainable telecentre.

Lastly, the proposed TPOA telecentre architecture was to strengthen telecentre sustainability, especially by the Orang Asli communities. Although the ultimate telecentre sustainability is affected by a wider scope of other factors, we are playing our part to add more values to its sustainability through ICT innovations.

II. Background

Telecentre has been widely recognized as the tool for bridging the digital divide. ICT technologies have enabled a wide variety of services to act as a catalyst to accelerate rural social development. Various telecentre models have been proposed and implemented across the world for different domestic conditions and the availability of telecommunication infrastructure. The benefits and positive impacts of implementing telecentre for rural development are undeniable but there are still plenty of challenges to overcome in order to make the telecentre more effective in achieving its goals. After all, there are still 44% of the world population still living in rural areas according to the World Bank data in 2019 [1].

Many studies have been conducted on the effective Internet access option through the advancement of telecommunication technologies, the various telecentre services for a wider scope of applications in development, and as well as various telecentre sustainability issues. Besides, the effectiveness of telecentre in serving its roles and how telecentre can play a new role alongside the wave of development, are amongst the on-going studies [2]–[7].

Owing to the difficult physical access and lack of local resources in the rural areas, the use of ICT and telecommunication technology are the key to bridge the inequality. The definition of rural areas differs from country to country mainly based on the population size per square kilometer. Wiggins and Proctor [8] further categorized rural areas into three different sub-categories, namely Peri-Urban, Middle rural, and Remote rural, where the further the distance from the urban centres, the less economic advantage hence being considered as more rural. In this case, the remote rural is the furthest to reach with the least economic activities. This paper focuses on TPOA telecentre implementations at the remote Orang Asli villages, our discussion will be specifically on the telecentre implementations in the remote rural areas. Some of the common characteristics of remote Orang Asli villages include extremely difficult off-road access of 30km to 60km, completely without electrical power supply, no means of any communication, and the majority of the villagers are low in literacy or never see a computer in their life. The off-road access is only possible via 4WD vehicles and during the rainy season, landslide and broken bridges may cut off the access road for weeks or even months. Hence, erecting telecentre in the mentioned Orang Asli villages is not only physically challenging but the low literacy level will certainly impose another level of challenge to the introduction of technology and its adoption in the community. The difficult physical access to the remote telecentre sites has imposed a very tough on-site support coming from the urban.

A. Telecentre Evolution

We are reviewing some of the prior telecentre initiatives particularly in Malaysia owing to most of them share similar characteristics as the TPOA telecentre implementation. The advancement in the telecentre ICT model throughout the years will certainly set a reference guideline for future government adoption on bridging the digital divide. The eBario telecentre implemented in 2002 [9] has set a successful telecentre model for Malaysia. The eBario initiative has sparked social-economy development in the rural village, Bario and the impact of the telecentre can still be noticed even until now although the eBario telecentre has eventually disappeared and has changed its form to serve a different role under the wave of development. The impact of the telecentre is so promising until later in 2007, a few more telecentres were initiated for other remote rural communities through the replication of the eBario successful model [9], [10]. The five telecentre sites under the telecentre replication program include eLamai, eBa'Kelalan, eBuayan, eLarapan and eBario itself as eBario 2.0 [11]. During the same period, several types of telecentres have been initiated throughout Malaysia, namely Universal Service Provision (USP), Rural Internet Center (RIC), Medan Infodesa (MID), USP Communications Center (UCC), Rural Broadband Library, etc. [12]. It is worth noting that most of the widely deployed telecentres were not located in remote rural areas but more in middle rural and peri-urban areas which are relatively relaxed in local resources such as consistent electricity supply, good road access, and more Internet connection options.

The eBario telecentre model in its early day consists of 10 desktop computers and a VSAT satellite gateway that provides Internet access as well as public telephony service. The eBario telecentre architecture is shown in Fig. 1 [13]. The telecentre was powered by a diesel power generator then changed to solar-diesel hybrid and at a later stage into fully solar-powered. The change in the power supply system was due to the inconsistency supply of diesel fuel into the remote rural and also the hiked of fuel price. The telecentre experienced technical issues where it still suffered from the inconsistent solar power supply which prevented it to operate continuously. The occasion shortage of power supply was caused by the design capacity of the system failed to cope with the weather condition in the Kelabit highland. The battery banks were not charged in time to provide power to the telecentre [14].

Under the eBario model replication to other remote communities, the ICT equipment power consumptions have been taken into consideration hence lower power requirement mini desktop and laptop computers have been adopted. Other equipment such as the projector and printer were also adopting a lower power model. This would have eased the issue of insufficient power as experienced in the eBario telecentre. With a lower power requirement, the bad weather condition would not affect the solar power system too much hence avoiding power blackout at the telecentre. In the same period, the eBario computing systems also undergone an upgrade to adopt low power mini desktops and laptops. The awareness on the impact of power requirement towards the operation sustainability has caused a reduction in computer system power requirement, from 300W per computer down to 50W per desktop and 30W per laptop. The decision on changing the computer model has created plenty of power reserved for the telecentre. The telecentres did not experience much power supply inconsistency problem until five years later when the battery banks reached the end of its usable lifespan. This had shown improvement in the energy efficiency of an ICT architecture has greatly influenced the continuous operation of a telecentre.
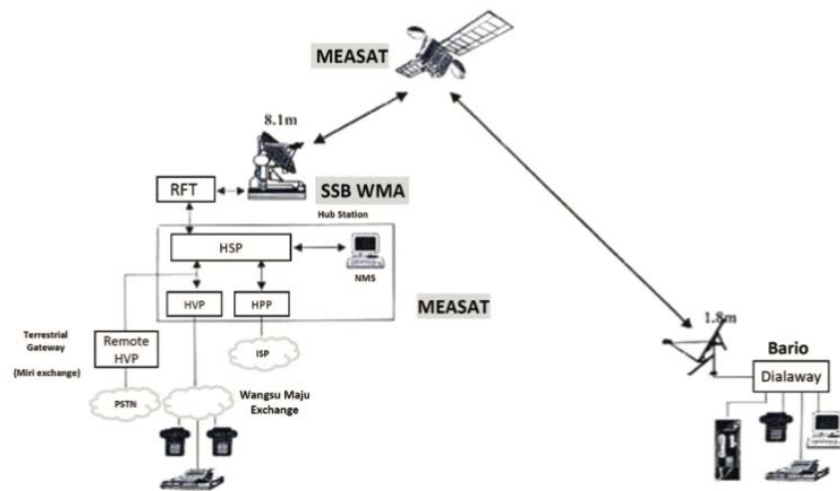
Fig. 1. The eBario Telecentre Architecture [13].

B. Telecentre Sustainability

The sustainability of a telecentre can be seen from various aspects such as financial, social, operational, policy, and organizational [15] and they are interrelated in the order to achieve the ultimate overall telecentre sustainability. While a large amount of studies focuses on investigating the various factors that affect the telecentre financial sustainability [5], [16]–[18]. In this paper, we will be focusing on the technical operational sustainability of the telecentre instead. From the technology point of view, technology challenges in operation can have a great impact that could result in telecentre failing to fulfill its roles in rural development. Michael highlighted the sustainability failure of SARI project was partly related to the availability of a trained telecentre operator [19]. The telecentre with well-trained operators will sustain a longer period compared to otherwise. Telecentre operational challenges have not gained much attention in the telecentre research because most telecentres have telecentre manager to run the operation when they deprive of trained operators. For the case of telecentre implemented in remote rural areas, the luxury of having a competent team of managers and operators is not always there. Remote rural telecentres' operation is usually run by volunteers coming from the benefiting community itself. External ICT experts will provide technical training to the voluntary future operators during the early stage of telecentre implementation. The natural state of the voluntary operators is that they are just helping to operate the telecentre based on their current capacity which may not be fully competent to perform the tasks.

Over the past two decades, technical challenges in telecentre operation include dealing with computer hardware and software hiccups, guide how to use an application, maintaining ICT equipment that encountered service intermittent, housekeeping the computer operating system in order to keep them in optimum performance state, etc. Failure to handle these technical challenges, the telecentre may not operate at its optimum state to serve its purposes. New users from a remote community with relatively low literacy levels would need a lot of help from the operators in order to learn new technology and demand help when they got stuck with some unfamiliar functionalities. Hence, the task of telecentre operation is not just to open the telecentre so that users can come to use it but also to provide technical assistance. Without competent technical assistance, users may find the telecentre less useful if they do not have the capability to master all knowledge for them to carry out what they wanted. This scenario is particularly true for remote rural telecentres that do not receive a lot of on-site supports from external parties. A community sustained telecentre will experience this a lot because there are very limited competent operators available within the community.

The rural telecentres implemented under the eBario model replication were purely community-driven where there were local champions to lead and manage the telecentre operation. Telecentre operators or caretakers were trained and worked together with the local champions. There have been cases where the trained operators left the telecentre service to attain jobs in the urban or leaving the village, hence under such a situation, activities at the telecentre will be interrupted and the opening hours of the telecentre will also be affected. Besides, any departure of the trained operators will indirectly weaken the telecentre management ability to perform technical maintenance. If such a situation prolongs, the telecentre may slowly become less accessible and lose its role in catalyzing community development. Before personal smart devices become popular, Internet access can only be performed using conventional computer systems. The only means to learn and use digital tools is through the telecentre's computer units. If the telecentre operation cannot keep to its opening hours, the interest and the will to adopt ICT will be discouraged. Such a phenomenon had been observed in a few telecentres and to restore it, a new group of operators need to undergo another round of intensive training conducted by external technologists. Without competent telecentre operators to manage and provide guidance, the computer systems were way too complex for novice villagers to operate. The computer system at the time was a complex desktop system with a relatively complex operating system, the Windows. Although at the time of mention, the telecentres were used for relatively basic applications such as emailing, web surfing, document processing, etc., the telecentre operator's guidance is crucial to

assist the accomplishment of tasks on these computer systems. Without their help, villagers have been giving their feedback that they still could not operate the computer on their own, owing to ICT is still something very new in their daily routine. The adoption of ICT is subject to the amount of exposure the villagers get from the ICT ecosystem, in order to find out what the technology can do for them. Hence, in a community with very low computer ownership, the consistency of telecentre operation will certainly affect the adoption of ICT among the community members.

Today's ICT world has evolved into personal smart devices that everyone can afford to own one. The digital revolution has changed its form from desktop / portable computing to personal smart appliances such as smartphones and tablet computing. Application software has been shifted from desktop applications to simpler mobile applications. The boom in mobile application development has created a wider variety of features on the smart devices hence there will be more to be learned. The telecentre ICT architecture shall be upgraded to catch up with the new wave of the digital revolution and at the same time to address some of the limitations encountered. The future telecentre architecture will not only have to address some of the technical limitations but also enable a simpler operation and system maintenance to allow easier telecentre sustainability especially those that are sustained by the local community.

III. THE CHALLENGES

The telecentre designed for Orang Asli villages will be addressing some of the shortcomings of the previous telecentre implementations in the remote rural areas of East Malaysia. Limitations were being observed from the past telecentre experience hence some of the critical technical issues need to be overcome so that they can be ease in the future telecentre implementations.

A. Unpredictable Power Outage

In the remote rural environment, there has been no grid electrical power supply hence standalone solar power system has been a very popular choice due to plenty of sunlight available in this region. Even though some may be running on a micro-hydropower system from the river but solar power is a still more flexible and easily available option. The nature of the solar power system is that it has limited capacity subject to the initial design. Owing to this limitation, the power usage of the telecentre cannot go too far beyond the rated supply. If the usage goes beyond the maximum rated power, total system blackout will occur. In a conventional solar power system, the output of the system is usually converted to the standard AC format, in this case, 220VAC. All electrical and ICT appliances were draining power directly from this supply and any power blackout will cause complete telecentre shut down. Even with the adoption of UPS backup power, it will not effective due to its very short backup time and not economical in long run. With the convenient 220VAC standard output supply, many other unplanned appliances can be plugged into the solar power system hence putting the power system at a higher risk of overloading. Designing a higher capacity of solar power systems with higher reserve capacity may be seen as a solution but it incurs much higher initial capital investment. Again, no

matter how much additional power capacity is being provided, they still have their limit and this limit can always be breached as long as the usage is not controlled or regulated. Furthermore, the actual capacity of a solar system is greatly influenced by the weather condition, for the same amount of usage under the rainy season, the power system could still be overloaded. It is not wise to put the users to the blame on unplanned usage that could cause unpredictable power outage at the telecentre, a clever power system design would have either prevented or minimized the occurrence of such a situation. The unpredictable power outage problem is not a design problem but more of an operational mismatch that could happen at any time if the power usage is not being carefully monitored. From our experiences with telecentres for eBario, eLamai, eBa'Kelalan, and eLarapan which are powered by standalone solar power systems, the unpredictable power outage issue is common and has been distracting telecentre services from time to time. The telecentre power supply architecture shall be redesigned to minimize the outage issues.

B. Fragile End user Computer System

The computer system is the interface of computing experience for end-users. Users would need to familiar with all the computer system components such as the CPU unit, the monitor, the keyboard, the mouse, etc. The majority of the telecentre users are rural folks who have very little experience dealing with the relatively advanced computer system. Their experience with computer systems is a novice and only know using the computer system as it is, as what they have been taught on how to operate them. The steps of operation given can be as clear and precise but things do not always turn up the same way all the time. Faulty component, loose connection, and missing operating steps are common problems in daily computing experience that require basic troubleshooting skills to solve. But for the user group that makes up the majority of the rural community, such simple problems can be the cut-off points that stop them from using the computer systems. In our experience, many computer systems in the telecentre being labeled as "broken" were actually "usable" by performing some simple or common sensed troubleshooting. The solution may look straightforward but due to the large knowledge and experience gap between the urban and rural, the rural telecentre users were bog down by very simple problems that in the eyes of many urban folks, are not a problem at all. For example, a loose power connector, monitor cable, keyboard, or mouse could cause a problem and stop them from using the computer system. The standard computer system with multiple components connecting to each other for it to work is too complex to manage and to troubleshoot especially for someone who just wants to use for simple Internet browsing. To achieve interrupt free computer usage, the users will have a lot more to catch up, learning how to deal with computer components resulting in a steeper learning curve in order to be effective in using a computer system. A more robust end-user computer system is crucial for rural telecentre especially since the majority of users have relatively low computer literacy.

C. Inconsistent Application Experience

Software applications are the actual ICT tools to enable ICT services in a computer system. Similar to the problem of computer system components and hardware, inconsistent

software application behavior will certainly produce situations users never seen before and unable to handle. There is no way for users to learn and see every behavior of all software applications. Our past telecentres were adopting a corporate-grade Microsoft operating platform with Office tools and other Internet browsing applications. No doubt the operating platform is feature-rich and sufficiently powerful in handling decent computing needs, but the highly dynamic and flexible platform is also a foe to the rural users. To perform a simple function, there may be tens of options to choose from and users can get very confused, and sometimes they do get pop-up windows that asking them questions that they do not know how to answer and how to proceed from there. Even during system boot-up, unexpected pop-up windows may have paused the start-up process and asking users to choose "OK" or "Cancel" in order to proceed. Without having enough knowledge to understand what the pop-up window is asking for, some users do not even dare to make the choice, worry that they might cause damage to the computer system. Many such scenarios in our experience, users chose to turn off the computer system and labeled it as unusable or broken. They will not touch the computer system until the so-called problem of pop-up windows being rectified by technical personnel. The situation turned worse if there were no competent technical support on-site to rectify the issue, the computer system could stay idle for months without anyone using it. Sadly, the pop-up window could be just a casual reporting message from the operating system and because the message sounded technical, it could be treated by users as the system is having a problem or broken. Hence, the application experience shall be made as consistent as possible to avoid confusion and increase productivity.

D. Complex and Steep Learning Curve

The combination of a conventional desktop computer system with a modern desktop operating system such as the Windows operating platform, can be very complicated and having a steep learning curve. For the users to be able to use the computer system as a tool for their purposes, they have to first master the operating of the computer system as well as all the associated operating system features. Without understanding the process flow and features of the operating environment, users may not be able to make good use of this digital tool to assist them in their communication, learning, information gathering, and productivities. Owing to the steep learning curve, the growth of the digital adoption rate through the use of a conventional computer system could be relatively slow. Hence, instead of making the users learn more about the digital tools before they can use the tools effectively, it would be much relaxed if the computing system can be made or designed with much gentle learning curve for faster and less stressful adoption.

E. Lack of Competent On-site Technical Support

The biggest problem of technology deployment in rural areas is the lack of competent technical support personnel on-site. Technology deployment is a relatively new thing to many rural communities, the knowledge, and experience of local support is always lacking. Technical training can be provided to develop the local human capital in technical support but it takes time and experience for technical personnel to become competent. The challenge of training the local technical support is not only time consuming but the high turn-over rate of trained technical personnel has put the effort of earlier training back to square one. From our experience, the high turn-over rate was due to trained personnel leaving the village for other jobs in town. The telecentre technical support shall be made as less technical as possible so that a telecentre caretaker can easily acquire the knowledge to carry out the task of technical support.

F. Lack of Autonomous Service Monitoring

For many rural areas, the only communication means to the outside world is through the telecentre Internet gateway. In rural areas of Malaysia, most rural telecentres utilize the VSAT satellite connection for Internet access and telephony. This is due to satellite coverage that cannot be blocked by any on land obstacles such as mountain and thick forest. Regardless of how remote the telecentre is, as long as there is good sky visibility, the VSAT connection will be able to bridge the communication gap. The problem is, any failure of VSAT equipment or service interruption caused by local equipment, will cut off all communication between the telecentre and the outside world. This including there will be no way to file a complaint on equipment fault and request for technical support from the VSAT operation team. When the issue of such is happening, someone from the local community shall be traveling to the nearest town in order to file a fault report. For rural villages that are extremely remote, it may take months until someone is making a trip to the nearby town and make that fault complaint. During the fault period of critical equipment, the telecentre ICT services will be completely cut off and made redundant. Under normal circumstances, it is difficult for the VSAT service operator to detect the faulty equipment through a link down indicator status because a VSAT link down could be due to a temporary power outage or the equipment being shut down on purpose. The VSAT operator will not send the technical support team until they received an official complaint from the users. Our experience with telecentres, the VSAT service outage could persist up to a few months without being noticed until the problem is reported to the VSAT support team. Such a long period of communication outage has defeated the primary purpose of a telecentre to act as a communication hub to bridge the digital divide in the rural areas. Owing to that, the communication link monitoring mechanism is essential in order to keep the telecentre to its maximum service uptime.

IV. TPOA TELECENTRE DESIGN PHILOSOPHY

A. The Mission of TPOA

TPOA is the Telecentre Program for Orang Asli project started in 2013 with prior engagement with Orang Asli as early as 2011 on the need analysis. Telecentres have been implemented at four remote Orang Asli communities in Peninsular Malaysia namely Pos Sinderut and Pos Lenjang at Pahang state, Pos Gob and Pos Balar at Kelantan state. The project was a collaboration with the Department for Development of Orang Asli (JAKOA). The main objectives were to facilitate communication, provide access to information, create resource centres for new knowledge and skills, provide education through ICT, provide training for local capacity building, provide health-related information to

villagers, and drive rural development through the telecentres as Rural Transformation Centres through the execution of the six telecentre programs; Agribusiness, Eco-tourism, Education, Health, Indigenous Knowledge and ICT Training [20].

Apart from serving the in-house telecentre programs under the TPOA project, the telecentre should also serve the local community in aspects where deem needed. In the era of the rapid growth of smart mobile communication, more people own smart mobile devices than conventional computing devices and the growth of personal smart devices among the Orang Asli community has also increased over years. The TPOA telecentre would need to expand its coverage scope to accommodate the demand of WiFi connection for Internet access as well as facilitating more socio-economic activities. The smart devices enable personal communication with the outside world, allowing more exchange of information on matters of personal and community interest. At some of the TPOA telecentre sites, the service of charging the community smart devices and the power bank has to be provided as many do not have access to the electricity supply. Hence, the TPOA telecentre is no longer just providing resources to serve in-house programs and equipment but also acts as a community communication hub and power supply centre. On the other hand, the increased ownership of personal smart devices among the local community could be directly or indirectly due to the availability of WiFi Internet connection provided by the telecentre, hence accelerated the local digital adoption rate. Since the full deployment of TPOA telecentres in 2017, apart from visiting the telecentre for its programs, the telecentre has become a daily visit place for the local community to pick up and send their personal messages over WhatsApp and WeChat. Some even make use of the messaging platform to raise and discuss cases on community land rights issues, to contact buyers for their farm products, and also to arrange with visitors coming to the village.

B. The Telecentre Design Philosophy

Sustainability is a key element in TPOA telecentre for remote Orang Asli communities. Looking from the past telecentre experience, telecentre implementation in its early stage can be relatively relaxed as everything is still brand new and less of a problem. When the telecentre goes on for a few years, more and more maintenances are required until a point where the telecentre may require equipment refresh in order to continue its existing role or be upgraded to serve a new function. The transition in community needs is taking time subject to how soon they get used to existing technology and be ready to move on to higher-level applications. This transition could take many years to happen hence the deployed telecentre ICT systems should last as long as possible until their service is no longer required. It is undeniable that some system parts will wear out with time and shall be maintained or replaced throughout its service life, but with additional considerations on maintenance, the overall telecentre service lifespan can be sustained and prolong. Hence, the telecentre ICT system architecture plays an important role to enable greater sustainability with minimal expenditures. More importantly, the telecentre at the end of the day, ideally shall be sustainable by the community itself within their financial capability instead of waiting for new large funding to refresh

the entire telecentre. Some older telecentres are way too expensive to be self-sustained by the community owing to the components to be replaced are beyond the community financial capability. With TPOA telecentres, the ICT systems have been optimized in order to achieve more community-friendly sustainability. The TPOA telecentre architecture focuses on improving the following aspects.

C. Energy Efficiency

The energy supply for remote telecentre is extremely critical as all equipment depends on it to operate. On a single large power supply system such as a standalone solar power system, once the power supply system is interrupted for whatever reasons, the telecentre will be shut off completely. Instead of providing an even larger power supply system, which is much more costly, we look into optimizing the ICT systems' power requirement to minimize the use of power and reduce the overall power expectation. ICT equipment with lower power requirements shall be carefully adopted. Also, instead of using one single large power supply system, multiple smaller independent power supply systems can be implemented to reduce the risk of putting all eggs in a single basket, but to dedicate independent power supply for different equipment based on their criticality. Hence, segregating critical equipment from being affected by power supply failure at other parts of the telecentre. Portable and mobile computing devices that come with their internal battery system shall be adopted to reduce the dependency on the always-on power supply in order to operate.

D. User Experience

The user experience of using the telecentre systems and services is crucial in determining the acceptance of technology in the community. Not everyone is technology savvy hence poor user experience in using the technology may put them off and thinking that the ICT skill is too difficult to acquire. A friendly and easy to use ICT system experience will certainly give motivation and confidence to novice users, to give them an impression that the use of ICT technology should be a straightforward process and anyone can learn the skill just like anyone else. A strategic combination of technology will give a better user experience even users with minimal literacy should be able to use it. Hence, user experience is a key expectation when comes to packaging the hardware and software solution for the telecentre. Given an example of poor user experience in urban ICT, connecting a laptop to a client projector for presentation sometime can be frustrating because the technology is supposed to be just plug-and-play but in reality, this never always the case.

E. Gentle Learning Curve

Adopting ICT technology requires a learning process and the learning curve of any telecentre tools should be as gentle as possible, moreover if the targeted community is low in literacy. ICT technology gets more complicated now but they can be made simple by focusing on the primary purposes the technology serves. By simplifying the ICT systems to serve just only a few relevant functionalities or re-designing the system to allow a simpler operation, would have made the learning curve very gentle. This is particularly important for new learners who have very little exposure to ICT systems and

services. The mastery of ICT requires a lot of knowledge but ICT is just a tool in the telecentre hence if the operation of the tools can be made simpler without the need to learn the in-depth ICT knowledge, then the technology adoption can be less stressful. How to achieve simpler operation and gentle learning, the ICT technology for telecentre shall be adopted strategically and innovatively. The concept of simplicity and gentle learning shall not only applicable to technology operation and learning, but should also apply to technical troubleshooting, for the minimum, to simply the recovery of equipment from operational hiccups. More importantly, the telecentre caretakers can have a higher confidence level in managing the technology and provide training. The training of new caretakers can also be done by the existing caretakers themselves. This shall contribute to telecentre operational sustainability in long run.

F. System Robustness

The system's robustness refers to its robustness against vandalism. The telecentre equipment that requires a higher level of robustness will be the shared end-user computing devices. From experience, shared end-user devices especially computer systems require frequent housekeeping and maintenance as users like to introduce all sorts of customization, changing some of the system settings, and even exploring every single function of the system. Under such a scenario, many computer systems do not work as expected after a few sessions of usage. Particularly children users, they will explore every corner of the system with their curiosity or even vandalize the system. Hence, end-user appliances should be something that could withstand extreme usage conditions especially maintaining its software functionalities. This is important to reduce the amount of effort in maintaining every piece of the end-user device in order to keep them in working order every day.

G. Seamless Interoperability

There will be many ICT appliances co-existed in the telecentre where information exchange and sharing of files somehow are unavoidable across devices. Simpler information and file sharing for printing or display on the projector will certainly create a smoother telecentre operation. In many cases, in order to share a file across devices wirelessly, many setup steps and configurations are required, making it a complicated task for telecentre caretakers to manage. Since information and file sharing is a common task, creating a seamless interoperable ICT platform within the telecentre shall ease the process. This seamless operation across devices shall also make the training session or remote conferencing over the projector screen easily accomplish. Setting up ICT devices to work together can be a frustrating task even it looks straightforward in the first place. Many advanced users still occasionally experiencing problems connecting their laptops to a projector for their meeting or presentation. Hence, it is important to minimize the potential hiccup in device interoperability especially for rural telecentre environments with no experienced technical support personnel.

H. Reduce Dependency on Internet Bandwidth

The telecentre receives all information and data through its Internet connection. In principle, as long as the Internet connection is active, all data can be obtained in digital form, but the case is not always ideal. Most rural telecentre's Internet connection is through the use of VSAT satellite connection, at least it is the case in Malaysia. Owing to it is a long-distance satellite connection, the reliability of connection especially for the Ku-band VSAT is subject to the weather condition and the actual bandwidth available is subject to contention with other VSAT terminals grouped under the same channel for bandwidth sharing. Hence, the Internet experience over the VSAT satellite system is far poorer than the urban fiber-based Internet experience. The VSAT connection is expected to experience fluctuate bandwidth performance throughout the day and this is common nature for a shared data connection with an affordable monthly rate. Hence, knowing the nature of the satellite connection, a 4Mbps Internet link may be meeting the bandwidth requirement of applications such as video streaming over the Internet but the experience could be bad from time to time. Having all applications executed over the Internet connection at all times is not a brilliant idea. This is particularly crucial for the delivery of multimedia-based education content. A new initiative is needed to adapt to the VSAT bandwidth nature and to reduce the full dependency on telecentre Internet bandwidth for all applications.

I. Protection against Negative Impact of the Digital Revolution

The introduction of the Internet through the telecentre opens the portal that allows the rural community to reach the rest of the world. The portal can deliver all required information, messages, interaction with other Internet communities, etc. This is a new digital revolution that the rural community never experience before and users can be naïve as not be able to judge between genuine and threat. Furthermore, the attractive social media network can be over additive to some newcomers to the world of the Internet. ICT technology has the ability to connect must also have control over parental new users. Instead of letting them swim on their own in the open sea of the Internet on day one, a guided exploration on the Internet will reduce the negative impact on the rural community. From experience, the widespread of social media has already caused social problems among the rural community hence the implementation of telecentre shall strike a balance between benefits and negative impacts.

V. TPOA TELECENTRE ARCHITECTURE AND COMPONENTS

The TPOA telecentre consists of a telecentre building that house all ICT systems, activity area and a 4kW standalone solar power system next to it. The full TPOA telecentre program implementation took several years to accomplish because it is not just the telecentre itself but there are many crucial non-technological stages involved such as community need analysis, community engagement, telecentre management committee formation, leaderships and community protocol development which blend the human and technology to achieve the telecentre objectives, bringing greater community development impact through intensive community participatory and lastly, to achieve a full community self-sustain telecentre in its operation. In this paper, we will be focusing on the technology portion of the TPOA implementation which will be on the ICT architecture and its

components. Adopting different ICT technology under the same roof will certainly experience many technical hiccups especially in device interoperability and operation intermittent. The job of managing and operating a telecentre can be an overwhelming task for inexperienced local caretakers. Hence, a good blend of ICT architecture is crucial to ensure the telecentre can deliver its intended services. The TPOA telecentre ICT architecture aims to provide a simple operating platform for telecentre caretakers and at the same time minimize the effort of technical troubleshooting by proposing a good and strategic combination of software-hardware solution that allows all components to work together seamlessly. Fig. 2 shows the TPOA telecentre components and services.

A. The Modular Power Supply

As we gained a good understanding of the standalone conventional solar power system's nature and problem, where it may be the single point of failure under overwhelming usage. Unfortunately, we still need the conventional solar power system to provide compatible power supply to some general AC-based appliances, we cannot change the conventional solar power system into a new proprietary standard for our purpose but to adopt additional smaller modular power supply systems as a backup measure. The modular solar power supply module was first introduced in our pilot deployment of the long-range wireless system for extending eBario Internet connection to nearby villages back in 2010 [21] and the modular solar power supply concept is effective in providing isolated solar power for dedicated equipment and yet it is a relatively small solar system. Hence, under the TPOA telecentre design, we have adopted a few modular solar power systems for several more critical equipment in the telecentre. The modular and conventional solar systems run in parallel to support each other in the event of an unpredicted outage. The Modular solar power systems provide backup power for tablet computers re-charging, local portal server, and local WiFi network equipment. By having more power supply options some of the services will still be active if the main conventional solar system fails. Owing to the modular solar systems are operated independently, they are not affected by any faults in other power supply systems. The modular solar supply is relatively light-weight and can be easily relocated to a new location for future service that extends beyond the telecentre.

Fig. 2.   TPAO Telecentre ICT Components and Services.

B. The Mobile End-User Computing

The end-user computing device selection is critical as they are to be deployed in large numbers. We have chosen to adopt tablet smart devices as the end-user interface to the ICT applications and the Internet. Tablet computer has the advantage of power supply due to it has a built-in battery with relatively long operating hours before it requires a power re-charge. The USB-based charging standard on the device enabled flexibility to be recharged by any universal power source, for example, it can also be charged using the modular solar power charger that delivers USB power output. Tablet computer comes in relatively large screen size, in this case, we have chosen the iPad Mini that comes with close to 8 inches of screen size which is sufficient for most applications. The screen is not larger as to strike a balance to the amount of power required to fully charge the device. The iOS platform is by far the most stable and user-friendly user platform to our standard. It meets our expectations of good user experience and for our in-house developed education app. We also leverage on the parental control feature to impose restriction for making the tablet more robust against out of scope usage and user activities. Others may see adopting iOS devices would increase the cost of implementation as they are generally more expensive but for better device's built quality and the associated iOS platform, it is a strategic provision of technology for a rural telecentre. The iOS devices have met our ICT system expectations on simplicity, robust, seamless interoperability with other devices and parental control. Since the telecentre was implemented in 2016, we have not encountered many issues in maintaining the tablet in working order.

C. The Local Information Portal

In order to reduce the dependency of applications on the telecentre Internet bandwidth which is over the limited VSAT satellite connection, we have built a UNIX based microserver to act as a local information portal to serve heavy-weight multimedia contents for the telecentre education applications. The microserver power consumption has been optimized to operate under 20W of power and could also be independently powered by the modular solar power supply in the event of the main solar power supply not available. With the contents hosted on the local server, even in the event of an Internet outage, the education contents will still be available for the education program. Owing to the power supply at the telecentre that may not always be stable and blackout may occur unexpectedly, the operating system of the local information portal has been designed and set up in such a way to survive a sudden power cut and boot to its normal operation when the power is back again. The entire recovery process does not require human intervention hence reduce the need for maintenance after a blackout incident. The local information portal also playing the role of synchronizing its contents and acting as an Internet service availability check-point. The local server has been programmed to execute a monitoring and check-in mechanism over the cloud service to allow monitoring of telecentre Internet service from a public network
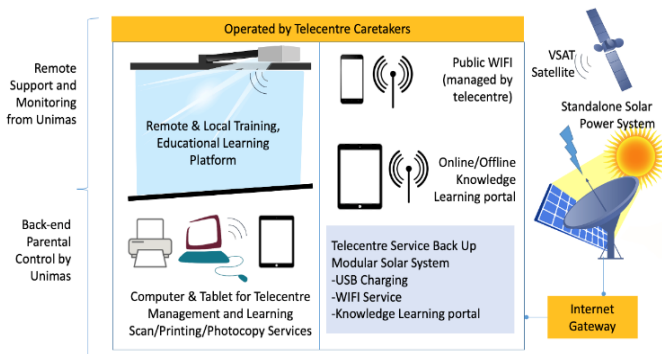
D. The Presentation Platform

The presentation platform is also known as Remote & Local Training, cum Education platform which consists of a projector, wall screen, and a TV unit. Its primary purpose is for conducting group training and online group meeting. Both projector and TV were connected to the micro-console digital media player called the AppleTV. The micro-console takes advantage of the AirPlay feature over the WiFi network to allow tablet computers or the management workstation to playback its content to the connected display equipment, in this case, the projector or the TV unit. This seamless display porting from user device to the big screen with only a few taps on the user device has created a wonderful user experience for the telecentre caretakers to conduct training activities, share display, and demonstrate the use of various apps for a bigger group of audience. The front-facing camera of the tablet computer combined with the presentation platform's big-screen display makes up a perfect tool for online conference meetings and online training. The presentation platform allows switching of the display from multiple user devices to share their screen seamlessly has enabled presentation to a big group of audience easier than ever. Regular online meetings and education sessions have been conducted over this platform and the caretakers find it very easy to set up these sessions.

E. The Wireless Network

The wireless network of the telecentre allows easier integration of the telecentre equipment. It does not only provide Internet access to all authorized telecentre equipment via the WiFi access network but also have to service connection from the personal WiFi devices of the local community. The WiFi network is a convenience to access but its nature is notorious for its quick network performance degradation under an overcrowded situation. We experience a lot of similar situations even in urban WiFi access. Concurrent connections from more than 30 devices to the same access point will certainly cause the performance problem to the WiFi network. The WiFi network needs segregation in terms of the number of connected devices and bandwidth prioritization so that the telecentre internal networked operation will not be affected by the seasonal peak usage coming from the community's personal devices. The public access WiFi is on another dedicated network segment where access control has been implemented for fair usage. Owing to the Internet VSAT gateway has limited bandwidth, and 4Mbps of bandwidth can never be enough for a large group of user access to modern multimedia intensive applications such as YouTube. The regulation on bandwidth usage is crucial to ensure the more critical messaging applications have higher priority access to bandwidth. Telecentre core applications will also be given priority and higher bandwidth allocation, for example, the online real-time conferencing will have higher bandwidth priority than other non-real-time applications.

F. The Management Workstation

Telecentre services are not limited to end-user usage but also to assist telecentre management such as caretakers to conduct activities, dealing with official paperwork, preparing new content, etc. Another more powerful workstation computer has been allocated for more advanced tasks and applications.

The workstation computer can seamlessly deploy new content such as eBooks and other media onto all the tablet computers via the AirDrop feature. Besides, it has integration with the presentation platform as well for controlling presentation sessions or video playback on the presentation big screen. The workstation housed all training materials in an interactive website hosted locally. New training materials can be synchronized into the workstation via online cloud storage service.

G. The Multi-functional Printer

The multi-functional printer is an add-on to support local community needs such as to produce hard copies of government application forms, passport photo printing, ID card photocopy, and some other paper-based reports and letters. The multi-functional printer is a low-cost inkjet printer with scanning functionality. The role of the printer is not an essential part of the ICT platform but to provide convenience to meet the needs of many conventional paper-based systems still using in rural areas. The printing is a charged service where the income is used for consumable ink replacement and printing papers. The printer is also a networked printer with WiFi capability to connect to the telecentre WiFi network. Printing of files and photos to the printer is done seamlessly via the AirPrint feature supported under iOS and macOS.

VI. TELECENTRE SETUP AND OPERATION METHODOLOGY

The telecentre setup is not just merely putting up the planned ICT systems in place and get them up running. The setup and building processes of the telecentre play another important role in determining the future sustainability of the telecentre itself. The telecentre building fund is from the Malaysian government hence by impression, the telecentre belongs to the government and the government supposes to take care of it from its operation to maintenance. Unfortunately, for remote rural telecentre projects like TPOA, the funding is usually one-off, meaning the funding is fixed to an allocated amount. The long-term sustainability plan and budget may not be part of the picture but more of a case-by-case basis. Owing to that, the benefiting local community from the telecentre project shall be brought into the picture so that the telecentre will be sustained by the local community as far as the community still needed it. For such a remote rural telecentre, external support is always very expensive and difficult to realize due to the telecentre is too remote. Hence, the telecentre requires local community commitment, participation, and support in order to be sustainable. Several setup and operation methodologies have been used to encourage community participatory so that at the end of the day, the telecentre can be managed, operated, and maintained entirely by the local community.

A. Sense of Belonging

The involvement of the local community in the building of the telecentre is crucial because their active and voluntary participation will inject a strong sense of belonging towards the telecentre. From the building structural design to the engagement of the local workforce in the construction process, local ideas and efforts will mark the community identity on the telecentre which makes it one of their own. With a strong sense of belonging, the local community will take greater

responsibility to ensure the safety of the building and also taking more initiatives to maintain it within their capacity. In the remote rural environment with minimum access to external support, telecentre can only be sustained effectively by the commitment of the local community. Fig. 3 shows the unique design of the four TPOA telecentres of Orang Asli at Pos Sinderut, Pos Lenjang, Pos Balar, and Pos Gob respectively representing their unique community identities.

Fig. 3.   The Four unique TPOA Telecentre Buildings.

B. Community Elected Telecentre Management Committee and Caretakers

The telecentre management committee shall be elected from the local community by the community itself. Our team of researchers will only provide guidance on the management skills and operation training so that the management committee together with the trained telecentre caretakers can manage and operate the telecentre services all on their own. This is an important step to cultivate community leadership and building local capacity to teach and educate the entire community. The local management committee will meet to discuss issues encountered and propose new plans to improve the telecentre and to appoint talented community members to succeed any caretakers left so that the telecentre will continue to serve its functions.

C. Community driven Operation Schedule

For a community-operated telecentre, the operation schedule shall be designed according to the living style of the respective local community. The urban 8 am to 5 pm schedule just not practical for a community that lives a rural lifestyle. Telecentre caretakers who man the telecentre are usually voluntarily and they have a living to make and a voluntary job at the telecentre should not put them into any difficult situation. Hence, the operation hours and days were to be discussed and decided through a community meeting. The local management committee even requests to stop the open WiFi service after 11 pm so that no community members hang around until midnight to avoid social issues within the community.

D. Telecentre Maintenance via Communal Work

The maintenance of the telecentre building and its surrounding requires communal work by the local community at a fixed routine, community members gathered together to clean up and maintain some of the broken pieces of the telecentre building. The solar panels of the telecentre power

supply system will require periodic cleaning and this is the time the panels get clean up in order to maintain the performance of the solar power system always at its peak. Hence, the culture of communal work will certainly help in sustaining the telecentre services to its designed lifespan. Fig. 4 shows a routine communal work of cleaning the solar panels.

Fig. 4.   Routine Communal Work of Cleaning the Solar Panels.

VII. CONCLUSIONS

In conclusion, we have presented challenges of telecentre implementation learned through 20 years of past telecentre projects where these challenges put up new expectations for future telecentres. The TPOA telecentre implementation is the showcase of the new ICT architecture for telecentre at the Orang Asli villages. Through the new expectations on ICT systems and its associated digital platform, technology has been designed and adopted strategically to assist the rural community to self-sustain the telecentre that brings social and rural development to their communities. The longer the telecentre operation can be sustained, the more rural development could be seen through the telecentre initiatives.

REFERENCES

[1] World Bank, "Rural Population (% of Total Population)," World Development Indicators 2019, 2019. https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS (accessed Jul. 20, 2020).

[2] R. Heeks, "Information and Communication Technologies, Poverty and Development," SSRN Electron. J., 1999, doi: 10.2139/ssrn.3477770.

[3] M. K. Bhattarai, "Telecentres in Nepal: lessons learned, prospects and challenges." http://www.ruralict.ftml.net/np/Bhatt.pdf (accessed Jun. 20, 2020).

[4] R. Harris, "Telecentre Sustainability: Financing ICTs for the poor," APDIP eNote 15, p. 0, 2007, [Online]. Available: www.apdip.net/apdipenote/15.pdf.

[5] M. E. K. Mphahlele and M. E. Maepa, "Critical success factors in telecentre sustainability: a case study of six telecentres in the Limpopo Province," Communication, vol. 29, no. 1–2, pp. 218–232, Jan. 2003, doi: 10.1080/02500160308538028.

[6] J. Lee and P. Sparks, "Sustaining a Nepali Telecenter: An Ethnographic Study Using Activity Theory," Int. J. Educ. Dev. Using Inf. Commun. Technol., vol. 10, no. 2, pp. 41–62, 2014, [Online]. Available: https://ezproxy.bibl.ulaval.ca/login?url=http://search.proquest.com/docview/1773225425?accountid=12008.

[7] T. Van Gevelt, "Indigenous communities, ICT and rural development: case studies in Tanzania and Sarawak, Malaysia," no. February, p. 24, 2017.

[8] S. Wiggins and S. Proctor, "How Special Are Rural Areas? The Economic Implications of Location for Rural Development," Dev. Policy Rev., vol. 19, no. 4, pp. 427–436, Dec. 2001, doi: 10.1111/1467-7679.00142.

[9] R. Harris, N. A. N. K. Ramaiyer, and J. Tarawe, "The eBario Story: ICTs for Rural Development," in 2018 International Conference on ICT for Rural Development (IC-ICTRuDev), Oct. 2018, pp. 63–68, doi: 10.1109/ICICTR.2018.8706855.

[10] Universiti Malaysia Sarawak, "ICT for rural communities in Malaysia," Asia Research News, 2007.

[11] A. W. Yeo, F. S. Hazis, T. Zaman, P. Songan, and K. A. Hamid, "Telecentre Replication Initiative in Borneo Malaysia: The CoERI Experience," Electron. J. Inf. Syst. Dev. Ctries., vol. 50, no. 1, pp. 1–14, Jan. 2012, doi: 10.1002/j.1681-4835.2012.tb00353.x.

[12] N. Razak and J. Malek, "Bridging digital divide in Malaysia: Cyber learning for the marginalized community," 2008, [Online]. Available: http://www.waseda.jp/DLI2008/program/proceedings/pdf/session9-1.pdf.

[13] Z. Hushairi, A. H. Khairuddin, P. Songan, A. W. Yeo, and J. Gnaniah, "Bridging the Digital Divide – the E-Bario and E-Bedian Telecommunication Framework," in Seventh Conference on Work With Computer Systems, 2004, pp. 277–281.

[14] M. Anyi, B. Kirke, and S. Ali, "Remote community electrification in Sarawak, Malaysia," Renew. Energy, vol. 35, no. 7, pp. 1609–1613, Jul. 2010, doi: 10.1016/j.renene.2010.01.005.

[15] S. S. B. Shadrach, "Telecentre Sustainability - Misnomers, Challenges, and Opportunities," telecentre.org, no. May, 2011.

[16] H. Ibrahim, A. Yasin, and Z. Md Dahalin, "Financial Sustainability Issues in Malaysia's Telecentres," Comput. Inf. Sci., vol. 3, no. 2, Apr. 2010, doi: 10.5539/cis.v3n2p235.

[17] F. Proenza, "Telecenter Sustainability - Myths and Opportunities," J. Dev. Comm., vol. 12, 2001.

[18] A. Bailey, "Issues Affecting the Social Sustainability of Telecentres in Developing Contexts: A Field Study of Sixteen Telecentres in Jamaica," Electron. J. Inf. Syst. Dev. Ctries., vol. 36, no. 1, pp. 1–18, Jan. 2009, doi: 10.1002/j.1681-4835.2009.tb00251.x.

[19] M. L. Best and R. Kumar, "Sustainability Failures of Rural Telecenters: Challenges from the Sustainable Access in Rural India (SARI) Project," Inf. Technol. Int. Dev., vol. 4, no. 4, pp. 31–45, 2008, doi: 10.1162/itid.2008.00025.

[20] P. Bala and C. E. Tan, Methodologies for Bridging the Digital Divide amongst Remote and Rural Communities in Peninsular Malaysia, 1st ed. Kota Samarahan: UNIMAS Publisher, 2018.

[21] K. Ab-Hamid, C. E. Tan, and S. P. Lau, "Self-sustainable energy efficient long range WiFi network for rural communities," in 2011 IEEE GLOBECOM Workshops (GC Wkshps), Dec. 2011, pp. 1050–1055, doi: 10.1109/GLOCOMW.2011.6162337.

A Computational Approach to Explore Extremist Ideologies in Daesh Discourse

Ayman F. Khafaga

College of Science and Humanities, Prince Sattam bin Abdulaziz University, Saudi Arabia

Abstract—**This paper uses a computer-based frequency analysis to present an ideological discourse analysis of extremist ideologies in Daesh discourse. More specifically, by using a computer-assisted text analysis, the paper attempts to investigate the hidden extremist ideologies beyond the discourse of the first issue of *Rumiyah*, one of the main digital publications of Daesh. The paper's main objectives are to expose hidden ideologies beyond the mere linguistic form of discourse, to offer better linguistic understanding of the manipulative use of language in religious discourse, and to highlight the relevance of using a computer-based frequency analysis to discourse studies and corpus linguistics. The paper also employs van Dijk's ideological discourse analysis, by adopting his positive self-presentation and negative other-presentation strategies. Findings reveal that Daesh discourse in *Rumiyah* is rhetorically structured to hide the manipulative ideologies of its users, which in turn functions to reformulate the social, political and religious attitudes of its readers.**

Keywords—*Computational linguistics; concordance; Daesh; frequency analysis; ideology; Rumiyah*

I. INTRODUCTION

This study presents a computer-based frequency analysis to explore the extremist ideologies in the discourse of Daesh's *Rumiyah*. This computational linguistics treatment is based on both a frequency distribution analysis conducted by the program of concordance and van Dijk's [1] ideological discourse analysis, by adopting his positive self-presentation and negative other-presentation strategies. Such a targeted linguistic treatment is emphasized by Smith [2] who reports that the emergence of some extremist religious movements, with their intentional discursive attempts to maintain their manipulative ideology, paves the way for more counter linguistic analysis, and opens new scopes of linguistic studies in the field of ideological discourse. This paper, therefore, attempts to investigate the extremist ideologies encoded in one of Daesh's publications: *Rumiyah* [3]. The study, therefore, is analytically based on both a computational approach, represented by the program of concordance, and van Dijk's [1] ideological discourse analysis.

A. Significance of the Study

The significance of this paper lies in its attempt to demonstrate the relevance of applying a computer-based frequency analysis to the linguistic analysis of texts. This is conducted by shedding light on the extent to which computer software packages, such as concordance can effectively be

used to help researchers from different research domains to arrive at concise and accurate results during the process of analysis. In light of the current study, concordance is applied to one of Daesh's publications (*Rumiyah* magazine) in order to uncover the hidden ideologies of its discourse. This might help readers of such magazine to resist the misleading information and deceptive tactics that depend on religious argumentations. The paper, therefore, might contribute to the field of computational linguistics, because it attempts to show the analytical integration of computer software programs, linguistics, and ideology.

B. Research Questions

This paper attempts to answer the following research questions:

1) To what extent does a computer-based frequency analysis help in the linguistic analysis of texts and talks?

2) What are the hidden ideologies encoded in the discourse of the selected magazine?

3) How can these ideologies be decoded by means of a computer-based frequency analysis?

C. Objectives of the Study

There are three objectives this study tries to achieve:

1) To explore the extent to which discourse is religionized to encode specific ideologies of its users.

2) To uncover the hidden extremist ideologies beyond the mere linguistic structures of Daesh's *Rumiyah*.

3) To prove the relevance of using computational analyses to discourse studies and corpus linguistics.

The rest of this study is organized as follows. Section 2 presents the literature review of the study, by shedding light on: (i) computational linguistics represented by concordance and its frequency distribution option, (ii) the notion of ideology within the scope of ideological discourse analysis, and (iii) a brief account of positive self-presentation and negative other-presentation strategies. Section 3 offers the methodology of the study, which comprises the approach of the study and data collection and description. Section 4 is devoted to data analysis. Section 5 is a conclusion and provides some recommendations for further research.

II. THEORETICAL PRELIMINARIES AND LITERATURE

A. Computer-Based Frequency Analysis

The frequency analysis attempted to be applied in this study will be conducted by the computer software program:

concordance. Concordance is a computational program which aims to access different types of texts in order to reveal the frequency each given word occurs in the text under analysis. Besides the number of occurrences a word has in text, this program also provides information about the contextual environment of any specific word highlighted for the analysis [4]. In this regard, Sinclair postulates that concordance is "a collection of the occurrences of a word -form, each in its own textual environment. In its simplest form, it is an index. Each word -form is indexed, and a reference is given to the place of each occurrence in a text" [5]. Concordance further offers certain analytical clues derived from frequencies of words. These clues help analysts arrive at results pertaining to various analytical purposes. Among these clues are collocations that shed light on the contextual world of a specific word in text. [6]. Concordance is very significant for any linguistic analysis and considerably contributes to the analytical weight of texts.

According to [7], the use of concordance in textual analysis has provided certain computational applications that deal with various analytical concepts, such as cataloguing, concordance, the analysis of form, of content and of the syntactic nature of texts. Concordance then can analytically be applied to the different types of linguistic analyses: syntactically, semantically, pragmatically, etc. Further, concordance is perceived as "a formatted version or display of all the occurrences or tokens of a particular type in a corpus [8]". According to [8], the type is usually called "a keyword but is sometimes referred to as a target item, node word or search item".

The nature of concordance is emphasized by Hockey [9], who argues that concordance is analytically manifested when the contextual background (i.e., the words in its company) of any searched word is revealed in text. This is also supported by [8], who postulates that concordance allows its users to identify the contextual environment of words and to recognize the different senses of a word type. In light of this paper, concordance is intended for providing frequency distribution analysis to the selected data. Its application functions to facilitate a much thorough and comprehensive study than would otherwise be possible. Therefore, concordance, here, aims to provide one verifiable input: *Frequency Distribution*. This computational option tends to offer the frequency of the searched word within its textual and contextual world.

Furthermore, concordance is perceived as an important tool in corpus linguistics [9]. This program provides texts' analysts three search options: the first option is frequency list, in which frequencies of specific word is generated in the corpus; the second option is collocations, which offers a combination of two words that co-occur together in text; and the third option is keywords, wherein a searched word is shown in accompany with the words preceded and followed it. For [9], the three options provide an analytical support for text analysis. That is, they help in demonstrating both the number of occurrences of a specific searched word, as well as its contextual environment.

B. Ideology and Ideological Discourse Analysis

The study of ideology in discourse has been the focus of many linguists [1, 10, 11, 12, 13, 14, 15]. From a socio-cognitive perspective, van Dijk [5] perceives ideology as a

social and cognitive form that is shared by a particular group or party. He argues that ideology is constituted by a number of ideas and beliefs that are perceived as group beliefs not personal ones. van Dijk [6] emphasizes that ideologies are close to what he calls "socially shared group knowledge", such as the specific knowledge shared by a number of individuals within the same institution or speech community. Ideologies are the driving force that shape and reshape the discursive practices among participants.

Ideologies "show a polarizing structure between US and THEM" [1], which indicates that they may be visualized as schemas of self-group. It can be assumed that each group is supposed to formulate its own conceptual schemata that frame its organizational patterns as well as its relationships with other groups. As such, ideologies constitute categories that abound in activities, norms, values and goals that are considered as dominant features that shape social positions and attitudes. These categories, therefore, are the main tenets that form the ideological schemata of any group in society. Such schemata help different group members defend their own interests and communicate their ideological beliefs [16, 17, 18].

Ideological discourse analysis is a discourse model which concentrates on investigating ideologies, their structures and representations in texts and talks. van Dijk [19] emphasizes that this model of discourse can be considered as one form of sociopolitical analysis of discourse because it related structures of discourse with structure of society. Ideological discourse analysis, according to van Dijk, is not only concerned with discovering the hidden ideologies in discourse, but also focuses on clarifying how structures of discourse are incorporated, intermingled and affected by the social structures of society.

C. Positive Self-Presentation and Negative Other-Presentation

The positive presentation of the in-group and the negative presentation of the out-group are among the most valuable ways of analyzing forms of ideological discourse. van Dijk [10] points out that these strategies function to expose the good of the in-group and the bad of the out-group. This can be conducted by ascribing good traits to US (i.e. the in-group members) via repetitions, association, and intensifying strategies, on the one hand, and by attributing bad qualities to THEM (i.e. the out-group members) through dissociation and downplaying strategies, on the other hand. Within these two processes, ideologies are formed, framed and expressed in texts and talks. The representation of an individual or a group or a political party positively or negatively attempts to affect the cognitive background of the public, and to reshape the ideological attitudes of individuals. The fundamental aim beyond such a process of positive/negative presentation is to focus on all information that beautify their image and strengthen their position, on the one hand, and de-emphasize all information that misshape their opponents and undermine their status, on the other.

According to van Dijk [20], the strategies used in the process of mollification or vilification are linguistically evidenced on the different linguistic levels of analysis: lexically, semantically, rhetorically, etc. On the lexical level, for example, the choice of specific lexis plays an effective role

in the process of communicating and reflecting ideologies in discourse. Rhetorically, the employment of euphemistic or dysphemistic terms is also significant in conveying positive presentation and negative presentation. Crucially, these strategies always set a distinction between in-group and out-group members within different discourse settings.

III. METHODOLOGY

A. Approach of the Study

This paper uses a computer-based frequency analysis and an ideological discourse analysis in data analysis. This means that the integration between the two approaches will be shown throughout the stages of analysis in this study. This is conducted by analyzing the underpinning ideologies of the selected data, and then carrying out a frequency analysis by means of concordance of specific words that are marked as indicative in each part of the ideological analysis of the magazine under investigation. Significantly, the use of concordance functions to help arrive at accurate results that support the whole linguistic analysis attempted in this paper, that is, to reveal the extremist ideologies in *Rumiyah*'s discourse.

B. Data

The corpus of this study consists of the first issue of *Rumiyah* magazine which was launched by Daesh's propaganda system and was published in September 2016. Some extracts from the selected issue are highlighted and marked as linguistically indicative in the study of ideological discourse of religious extremism. For McKernan [21], *Rumiyah* magazine was firstly published in September 2016 to replace its previous one entitled *Dabiq*. It is considered to be one of the effective propagandist tools used by Daesh to propagate its ideologies. The magazine is released in different languages, such as Arabic, English and Russian.

IV. ANALYSIS AND RESULTS

The analysis focuses on the manner through which Daesh's *Rumiyah* conveys its intended ideologies. In this regard, some discursive strategies, including positive and negative lexicalization and relational values of words, including mood and modality have been skillfully employed to reveal Daesh's extremist ideologies.

A. Positive and Negative Lexicalization

Lexis is always carriers of ideologies [22, 23]. The extremist ideology of jihad is interpersonally reflected in *Rumiyah'*s discourse via the employment of some positive and negative words. This ploy has been used to characterize the relationship between two different groups: Daesh and its opponents. Throughout the discourse of *Rumiyah*, a number of ideology-oriented vocabularies have been used to describe each group. Consequently, words such as *believers*, *muwahiddin*, *mujahidin*, and *martyrs* have been semantically antonymized by *disbelievers*, *mushrikin*, *murtaddin*, *tawaghit* and *kuffar*. The same oppositional lexicalization is also conveyed on the phrase level. This is clearly shown in phrases, such as *righteous men*, *fighters for Allah's cause, persevering brothers* and *lions of the Ummah*, which are contradictory counterparted by phrases, such as *enemies of Allah*, *people of falsehood*, *wicked scholars* and *Rafidi murtaddin*.

Crucially, this diametrically opposed lexicalization is intended to reflect an in-group positive presentation and an out-group negative presentation, which is based on the choice of words and phrases that imply positive or negative evaluation. This also supports the idea that ideologies are encoded by means of lexis, that is, words are considered ideology carriers in discourse. The use of ideologically contested words in Daesh's *Rumiyah*, therefore, is highly indicative in two ways: first, it emphasizes the meaning of jihad as one of the main religious ideologies of Daesh that is reflected by the use of words and phrases whose meanings connote the meaning of jihad, whether associatively or incompatibly; second, it demonstrates the interpersonal relationship between Daesh as a positively self-presented in-group and its opponents as a negatively other-presented out-group. Tables I and II show a computer-based frequency analysis wherein the number of occurrence of positive and negative lexicalization in the selected issue of *Rumiyah* is reflected.

TABLE I. A FREQUENCY ANALYSIS OF POSITIVE LEXIS IN *RUMIYAH*

Positive in-group lexis	
Word	Number of occurrence
mujahidin	25
brothers	20
believers	13
righteous	9
fighters	4
victorious	3
shuhada	2
muwahiddin	2
martyrs	1
martyrdom	1
righteousness	1
worshipers	1
lions	1
elite	1

TABLE II. A FREQUENCY ANALYSIS OF NEGATIVE LEXIS IN *RUMIYAH*

Negative out-group lexis	
Word	Number of occurrence
murtaddin	42
mushrikin	20
kafir	18
kuffar	12
taghut	10
disbelievers	9
kufr	9
tawaghit	7
enemies	5
enemy	5
tyrants	4
mushrik	3
fools	2
wicked	1

Tables I and II clarify that some words are frequently used to describe Daesh's affiliates positively and its opponents negatively. The high frequency of words, such as *mujahidin*, *brothers* and *believers*, on the positive side; and *murtaddin*, *mushrikin*, *kafir* and *kuffar*, on the negative side, indicates the conflicting way Daesh perceives its members and opponents. Indicatively, a simple look at the frequencies pertaining to each word in the above tables demonstrates that one can obtain specific information about the general ideological atmosphere of Daesh's discourse in *Rumiyah*. This, in turn, serves to emphasize the importance of applying a computer-based frequency analysis to discourse studies and corpus linguistics.

B. Mood

Many studies discussed the notion of mood as a discursive device which is used to communicate ideologies in discourse [24, 25, 26, 27]. This concept reflects the relational values between participants in discourse. Mood is realized in Daesh's *Rumiyah* through two lexico-grammatical patterns: the first is speech acts, which are manifested in the type of clause structure used in discourse, that is, the way of delivering the clause; declaratively, directively, or commissively. The second pattern is modality, which refers to all the non-propositional elements of a sentence, and is also demonstrated through two types of modality: truth modality and obligation modality.

C. Speech Acts

Using different speech acts in texts and talk is one way by which mood is conveyed in discourse [28, 29, 30]. Daesh uses three types of speech acts: the declarative, the directive and the commissive. The following extracts add clarification:

1) This is the way of the muwahhidin in every time and place. Whenever a generation of them passes, another generation follows, holding the banner of Tawhid overhead while plunging anew into the battle for Islam, which continues to be waged against shirk and its people. (*Rumiyah*, issue 1, p. 3)

2) By waging war against shirk and subjecting the people to the rule of the Lord of all creation, the greater injustice is eliminated. (ibid., p. 10)

3) The great gate of jihad with wealth is left wide open for the women who will make deals with their Lord, deals that will never end poorly. (ibid., p. 20)

The above extracts show different grammatical patterns of declarative sentences that revolve around the theme of jihad, which is delineated as the ultimate goal of Daesh. This is reflected by the use of some expressions, such as *the battle for Islam, to be waged against shirk, waging war against shirk, the great gate of jihad*. These clauses explicitly emphasize the significance of jihad in Daesh's ideological agenda. A computer-based frequency analysis is conducted here to demonstrate the extent to which the declarative mood is used in the selected data.

TABLE III. A FREQUENCY ANALYSIS OF DECLARATIVES IN *RUMIYAH*

Type of speech acts	No. of occurrence	Targeted ideology
Declarative mood	53	Jihad and Jama'ah

Table III displays that a computer-based frequency analysis helps in arriving at the total number of occurrences for a specific mood used in the selected data. Significantly, without the help of the computational work, it will be difficult to reach to such a credible and accurate result. Here lies the importance of applying computer software programs in accessing and analyzing any type of data.

On the interpersonal level, *Rumiya*'s discourse abounds in some discursive expressions that delineate the relationship between Daesh and its opponents. This is clearly shown through the dexterous employment of religiously-based words that are used to refer to two groups: the first group includes *muwahhidin, Tawhid, Islam, jihad, rewarded* and *jannah*, whereas the second group includes *shirk, injustice, dhimmi kafir, tyrant, hostile, sinful* and *mushrikin*. Again, this emphasizes the in-group/out-group polarization in *Rumiyah*'s discourse; those who adhere themselves to Daesh's religious ideologies (jihad and jama'ah) are members of the in-group, who are positively presented; and those who refused Daesh's ideologies are out-group enemy members, who are negatively presented. Also noticeable is the different parts of speech used in the description of the in/out-group in the magazine under investigation. Some are used to describe an action, as is the case with the words: *rewarded, shedding* and *waging*; others are employed to describe entities, such as *muwahhidin* and *mushrikin*; and a third group is used to describe a state, as in *sinful* and *hostile*. Consequently, Daesh communicates its religious ideologies interpersonally by making use of the declarative mood on the different levels of word classes: the verb, the noun and the adjective. Indicatively, the frequency analysis on such classes of words displays the extent to which both jihad and jama'ah are represented, either literally or associatively. The number of occurrences obtained from the frequency analysis adds significantly to the whole understanding of the hidden ideologies in the discourse of the selected magazine. Consider Tables IV and V.

TABLE IV. A FREQUENCY ANALYSIS OF LITERAL REPRESENTATION OF JIHAD AND JAMA'AH

Literal representation			
Jihad		Jama'ah	
Word	No. of occurrence	Word	No. of occurrence
jihad	59	jama'ah	19

TABLE V.   A FREQUENCY ANALYSIS OF ASSOCIATIVE REPRESENTATION OF JIHAD AND JAMA'AH

Associative representation			
Jihad		**Jama'ah**	
Word	**No. of occurrence**	**Word**	**No. of occurrence**
war	11	Khilafah	25
killing	32	Ummah	7
attacks	6		

Another interpersonal observation in the discourse of *Rumiya* lies in the grammatical use of the nominal gerund. In nominal gerunds are represented in *holding, plunging, waging* and *shedding*. These gerunds function as nouns of non-finite clauses within the larger structures of their sentences. As such, they serve as subjects of the larger sentences and add a sense of continuity to action. The meaning of such gerund clauses, then, may be: you (Daesh's soldiers) should continue holding the banner of Tawhid, waging war against shirk, and shedding the blood of a non-dhimmi kafir because these things are *not sinful* and are *rewarded with Jannah*. The following frequency analysis displays the high frequency of nominal gerunds in *Rumiyah*.

TABLE VI.   A FREQUENCY ANALYSIS OF NOMINAL GERUND IN *RUMIYAH*

Type of clause	**No. of occurrence**	**Targeted ideology**
Nominal gerund	27	Jihad and Jama'ah

As indicated in Table VI, nominal gerunds occur 27 times in the selected data. Again such a result can precisely be arrived at by a computer frequency analysis.

A further observation relates to the addressee's gender, that is, it is not only male participants who are supposed to commit themselves to Daesh's jihad, but women are discursively addressed as another discourse participant to share the same ideology of jihad as well. In extract (3) above, women are instigated to carry out a specific type of jihad, that is, *jihad with wealth*. This type of jihad, as understood from its name, is based on giving money to the so-called Islamic State so as to be used in military operations. Here, a new dimension of Daesh's jihad is stated; it is a type of jihad that is no longer committed physically, but rather financially. To highlight this idea, concordance has a role to play, as is shown in Table VII.

TABLE VII.   A FREQUENCY ANALYSIS OF REFERENCES TO GENDER IN *RUMIYAH*

References to Gender	**No. of occurrence**	**Targeted ideology**
Male	19	Jihad and Jama'ah
Female	11	Jihad and Jama'ah

The frequency analysis of the words *male* and *female* in Table VII casts emphasis on revealing the meaning that both men and women are targeted by Daesh discourse. This result is strengthened by the frequency analysis conducted above.

The directive mood is interpersonally communicated through the use of imperatives that are employed to accentuate its religious ideologies. The discourse of *Rumiyah* witnesses a strongly directed message of jihad through a number of imperatives. Using imperatives in discourse, according to [31] and [32], allows speakers to address their recipients clearly and directly and, thus, to practice power over them. Here, Daesh attempts to create a direct communicative channel with its recipients through which it can communicate its religious ideologies. Obviously, the use of such imperatives constitutes the meaning of jihad both explicitly and implicitly. Verbs, such as *strik, scorch, kill, stab, shoot, poison, run down,* and *fight* comprise a direct call towards violent action against Daesh's opponents, whereas verbs like *stand, die, follow, and mobilize* are indirect references to the same idea of jihad. The implicit jihad here is conveyed by the fact that acts of standing, dying, following, and mobilizing can be only realized via completing the course of Daesh's soldiers who died in their fighting against enemies. The frequency analysis in Table VIII shows that imperatives occur with the frequency of 24 occurrences, all of which are employed to communicate the meaning of jihad in *Rumiyah*.

TABLE VIII.   A FREQUENCY ANALYSIS OF IMPERATIVES IN *RUMIYAH*

Type of speech acts	**No. of occurrence**	**Targeted ideology**
Directive mood (imperatives)	24	Jihad

Another significant mood utilized in Daesh's *Rumiyah* to convey its religious ideologies is delivered commissively. Daesh's discourse presents clauses, such as *we will not rest from our jihad, their slaying will not harm the Islamic State, they shall shed many tears, we would have made effort to open the door...* and *this Ummah will be victorious* carry a strong commitment to some future actions that are expressed in the manner of vowing, threatening and promising, respectively. Daesh's commissive mood above has been characterized by three things: first, all commissives above revolve around the meanings of jihad and jama'ah. Both concepts have commissively been represented explicitly, through vowing and, implicitly, through threatening and promising; second, all commissives have been confirmed by the use of the truth modal *will*. This adds a sense of certitude and credibility to the pragmatic message of the commissive mood; and, third, the idea of jama'ah, which is indirectly communicated by the commissive *this Ummah will be victorious*, is preconditioned by the prepositional phrase *through your sacrifices*. This conditional promise aims to influence the attitudes of Daesh's soldiers and drive them to offer more sacrifices in order for the *Ummah* to be *victorious*. With the help of a computer-based frequency analysis, accurate results of the number of occurrences of commissives in the selected magazine are registered as is shown in Table IX.

TABLE IX.   A FREQUENCY ANALYSIS OF COMMISSIVES IN *RUMIYAH*

Type of speech acts	**No. of occurrence**	**Targeted ideology**
Commissive mood	14	Jihad and Jama'ah

D. Modality

A further device used to communicate ideologies is modality which is discussed by many linguists in previous literature [e.g., 33, 34, 35, 36]. Daesh employs two types of modality: obligation modality and truth modality. Both types aim to convey Daesh's jihad and jama'ah. Consider the following extracts:

1) This religion will remain established and will not be damaged by the death of any person. (*Rumiyah*, issue 1, p. 2)

2) A generation has been born in the Islamic state… that will not accept humiliation. (ibid., p. 37)

Daesh utilizes the truth modal 'will' in the clauses: *will remain, will not be damaged,* and *will not accept* in the above extracts to communicate trustworthiness, and to prove the validity of its arguments. Crucially, the use of truth modals reflects the degree of certitude which is often connected with the notion of authority a discourse participant practices over another (Yule, 1996). Here, by employing the truth modal *will*, Daesh tries to establish itself as having the discourse access of authority over its members, which makes it appear as a religion defender. This authoritative role is not only stated by the use of the truth modal *will* in extract (1) above, but also by the antonyms *established* and *damaged*. The meanings of the two words, however incompatible, remain complementary in confirming the concepts of jihad and jama'ah.

Another important type of modality, which is employed in Daesh's *Rumiyah*, is obligation modality. This type of modality is expressed by modals, such as 'must' and 'shall' as in the following extracts:

1) Men shall continue to be employed by Allah to frustrate the kuffar. (*Rumiyah*, issue 1, p. 3)

2) Muslims currently living in Dar al-Kufr must be reminded that the blood of the disbelievers is halal. (ibid., p. 36)

The obligation modals *shall* in *men shall continue* and *must* in *must be reminded* reflect the power of Daesh over its members. This nonreciprocal relationship of power is employed in discourse situations where one powerful participant dominates another. Daesh uses this type of modality to impose its own ideology over their recipients. The use of the agentless passive in *Muslims must be reminded* in extract (2) above signifies to "leave causality and agency unclear" [8]. As such, this grammatical feature can be said to have an experiential value in the sense that it leaves the responsibility of 'reminding' Muslims unspecified. Consequently, people all over the world, who believe in Daesh's ideological agenda, are responsible for reminding Muslims that *the blood of the disbelievers is halal*. In Islamic traditions, the word *halal* carries the speech function of permission and, therefore, affirms the associative meaning of Daesh's jihad. Tables X and XI present the frequencies of "will", "must" and "shall" in *Rumiyah*.

TABLE X.   A FREQUENCY ANALYSIS OF TRUTH MODALITY IN *RUMIYAH*

Truth Modality		
Word	Total occurrence	No. of indicative occurrence
Will	71	25

TABLE XI.   A FREQUENCY ANALYSIS OF OBLIGATION MODALITY IN *RUMIYAH*

Obligation Modality		
Word	Total occurrence	Number of indicative occurrence
must	6	4
shall	9	2

Tables X and XI clarify that the truth modal "will" and the obligation modals "must" and "shall" have total frequencies of 71, 6 and 9, respectively. Only 25, 4 and 2 occurrences of them are indicative in conveying the concepts of jihad and jama'ah. Significantly, the use of a computer-based frequency analysis helps in indicating the indicative and the non-indicative occurrences of the modals in the above table.

V.   CONCLUSION

This paper uses a computer-based frequency analysis and van Dijk's model of ideological discourse analysis to present a linguistic analysis of extremist ideologies in Daesh's *Rumiyah*. The analysis shows the relevance of applying computer software programs to the linguistic analysis of texts. It is also evidenced that the use of a frequency analysis helps arrive at accurate and credible analytical results, as well as yields better understanding of the textual and contextual meanings of the text under investigation.

The analysis, supported by concordance, demonstrated that Daesh's discourse in *Rumiyah* addressed two main extremist ideologies: jihad and jama'ah. These ideologies have been traced and reflected semantically, through patterns of interpersonal meanings manifested in positive and negative lexicalization, and mood (speech acts and modality). The semantic meanings of these ideological concepts have shown an increasing emphasis on religious ideas that are based on a clever process of intertextuality [37]. These ideas in turn serve to promote extremism and violence against the other, and create a group-oriented religious discourse, which abounds in meticulous ideological and discursive structures that aim to intensify the ideological polarization between in-groups and out-groups.

The study also revealed that Daesh's *Rumiyah* is apparently a propagandist protrusion to a specific ideological agenda. The textual organization of the magazine and its contextual atmosphere are integrated to produce the final discursive image of Daesh. This image is computationally and semantically delineated to establish a legitimate positive self-presentation to Daesh in a way that, on the surface, displays a reciprocal persuasive type of discourse, while, implicitly, shows a nonreciprocal extremist one. This, in turn, enables this movement to implant its extremist ideologies and to advocate its schematic violent goals.

This paper recommends further applications of computer-aided text analysis programs to other texts. This could reveal different findings other than what is approached in the current study. It might also demonstrate the extent to which computer software programs are analytically relevant to discourse studies and corpus linguistics.

REFERENCES

[1] T. A. van Dijk, "Discourse analysis as ideology analysis," In A. Wenden, and C. Schaffner (Eds), *Language and Peace,* pp. 17-33, 1995.

[2] R. M. Smith, "Religious rhetoric and the ethics of public discourse - The case of George W. Bush," *Political Theory*, vol. *36*, No. 2, pp. 272-300, 2008.

[3] Islamic State, *Rumiyah* (issue 1). Al-Hayat Media Center, available at (https://archive.org/search.php?query=rumiyah), 2016.

[4] J. Flowerdew, "Concordancing as a tool in course design," *System*, vol. 21, no. 2, pp. 231-244, 1993. https://doi.org/10.1016/0346-251X(93)90044-H

[5] J. Sinclair, *Corpus, Concordance Collocation*. Oxford: Oxford University Press, 1991.

[6] B. Kettemann, "On the use of concordancing in ELT," *TELL & CALL*, vol. 4, pp. 4-15, 1995.

[7] A. Thabet, "Applied computational linguistics: an approach to analysis and evaluation of EFL materials," *Damietta Faculty of Education Journal,* Part 1, No. 13, pp. 7-39, 1990.

[8] G. Kennedy, *An Introduction to Corpus Linguistics*. London & New York: Longman, 1998.

[9] S. Hockey, *A Guide to Computer Applications in the Humanities*. London: The Johns Hopkins University Press, 1980.

[10] T. A. van Dijk, (Ed.) "Discourse as interaction in society," In *Discourse as social Interaction: Discourse studies: A multidisciplinary introduction*, vol. 2, pp. 1-37, Sage publications, 1997.

[11] T. A. van Dijk, "Discourse, ideology and context," *Folia Linguistica*, vol. *xxx*, No. 1(2), pp. 11-40, 2001.

[12] T. A. van Dijk, (2004). Politics, ideology and discourse. In R. Wodak (Ed.), *Encyclopedia of Language and Linguistics: Second Language and Politics*. Oxford, UK: Elsevier, 2004.

[13] T. A. van Dijk, "Ideology and discourse analysis," *Journal of Political Ideologies*, vol. 11, No. 2, pp. 115-140, Routledge, 2006.

[14] N. Fairclough, *Language and power*. London and New York: Longman, 1989.

[15] R. Fowler, *Language in the news: Discourse and ideology in the press*. London: Routledge, 1991.

[16] U. Eco, *The limits of interpretation*. Bloomington: Indiana University Press, 1990.

[17] G. Fauconnier and M.Turner, "Conceptual blending, form and meaning," *Recherches en Communcation* [Communication Research], *vol. 19*, pp. 57-86, 2003.

[18] M. Dascal, *Interpretation and understanding*. Amsterdam: John Benjamins, 2003.

[19] T. A. van Dijk, *Ideology and discourse: A multidisciplinary introduction*. Barcelona, ES: Pompeu Fabra University, 2000.

[20] T. A. van Dijk, "Principles of critical discourse analysis," *Discourse & Society,* vol. 4, No. 2, pp. 249- 283, 1993.

[21] B. Mckeman, "ISIS's new mangazine Rumiyah shows the terror group is struggling to adjust to losses," *The Independent*, issue 14, 2016.

[22] M. A. K. Halliday and C. M. I. Matthiessen, *Halliday's introduction to functional grammar* (4th ed.). Revised by Christian M.I.M. Matthiessen. London & New York: Routledge, 2014.

[23] A. F. Khafaga, "Linguistic manipulation of political myth in Margaret Atwood's The Handmaid's Tale," *CDELT Occasional Papers in the Development of English Education*, Vol. 64, No. 1, pp. 71-107, 2018.

[24] J. Searle, J. (1969). *Speech acts*. Cambridge: Cambridge University Press, 1969.

[25] J. Searle, *Expression and meaning: Studies in the theory of speech acts*. Cambridge: Cambridge University Press, 1979.

[26] H. Grice, "Logic and conversation," In P. Cole and J. L. Morgan (Eds.), *Syntax and semantics,* vol.3, pp. 41-58, 1975.

[27] G. Yule, *Pragmatics*. Oxford University Press, 1996.

[28] M. Short, *Exploring the language of poems, plays and prose*. London & New York: Longman, 1996.

[29] J. Thomas, *Meaning in interaction: An introduction to pragmatics*. London & New York: Longman, 1995.

[30] A. Dylgjeri, "Analysis of speech acts in political speeches," *European Journal of Social Sciences Studies*, vol. 2, No. 2, pp. 19-25, 2017.

[31] K. von Fintel and S. Iatridou, "A modest proposal for the meaning of imperatives," In A. Arregui, M. L. Rivero and A. Salanova (Eds.), *Modality across syntactic categories*, pp. 288-319, Oxford University Press, 2017.

[32] A. F. Khafaga, "Do Directives Always Direct? Cognitive Directives and Meaning Inculcation in El-Sisi's Improvised Speeches: A Pragma-Semantic Analysis," *Journal of the Faculty of Education in Humanities and Literary Studies*, Vol. 25, No. 2, pp. 115-178, 2019.

[33] J. Flowerdew, "Globalization discourse: A view from the east," *Discourse & Society, vol. 13*, No. 2, pp. 209-225, 2002.

[34] M. A. K. Halliday, "Modes of meaning and modes of expression: Types of grammatical structure, and their determination by different semantic functions," In D. J. Allerton, E. Carney, and D. Holdcroft (Eds.), *Function and context in linguistic analysis,* (pp. 57–79). London & New York, NY: Cambridge University Press, 1979.

[35] T. Bloor and M. Bloor, *The functional analysis of English: A Hallidayan approach*. London & New York: Arnold, 1995.

[36] A. F. Khafaga, "Linguistic representation of power in Edward Bond's Lear: A lexico-pragmatic approach to critical discourse analysis," *International Journal of English Linguistics*, Vol, 9, No. 6, pp. 404-420, 2019.

[37] A. F. Khafaga, "Intertextual relationships in literary genres," *International Journal of English Linguistics*, Vol, 10, No. 3, pp. 177-188, 2020.

Optimized Cardiovascular Disease Detection and Features Extraction Algorithms from ECG Data

Sanjay Ghodake[1]
MIT Academy of Engineering
Alandi, Pune, India

Shashikant Ghumbre[2]
Govt. College of Engineering and
Research, Avasari, Pune, India

Sachin Deshmukh[3]
Dr.BAMU, Dept. of CSIT
Aurangabad, India

Abstract—A heart disease called cardiovascular diseases (CVD) is another leading cause for the death. There are several reasons that lead the CVD in human beings. The early detect of CVD helps to take necessary medical attentions to prevent the harms. The conventional techniques for CVD detection were manual and expensive which often delivers the inaccurate diagnosis. Since from the last decade the other inexpensive Computer Aided Diagnosis (CAD) based methods gained significant medical attentions. The CAD based techniques mainly based on raw Electro Cardiogram (ECG) signals of patient for the accurate and economical detection of CVD at early stage. In recent past, there are several CAD systems designed for CVD diagnosis utilizing raw ECG signals, however accuracy of CVD detection utilizing ECG bothered through several issues of research like QRS beats extraction, artefacts, efficient features extraction. This research paper present CVD novel framework, utilizing raw ECG signals and designed hybrid pre-processing algorithm for extracting artefacts and noise through raw ECG signal. Further designed simple and efficient dynamic thresholding based technique to extract the beats such as Q, R, S, and ST segment through pre-processing ECG signal. Third step perform the fusion of extracted beats and apply the feature extraction method called Normalized Higher Order Statistic (NHOS). The normalized HOS techniques asses the complexity among all the QRS based beats and delivers the more unique features for the accuracy enhancement. The final step is the classification by using five different classifiers for the CVD detection. The simulation results presented in this paper demonstrate that proposed framework achieved the significant accuracy improvement.

Keywords—*Electrocardiogram; heart disease; cardiovascular disease; hybrid filtering; features extraction; QRS and ST beats*

I. INTRODUCTION

The electrical improvement of human heart is assessed by utilizing the different body habitats, which are appropriate, masterminded over the body. From the estimation results, paper can shape a thought of heart limit called as electrocardiogram (ECG) [1]. The ECG is utilized to analyze of coronary illness these days in clinical frameworks. The heart banner winds up critical biomedical signs. A contraption called ECG banner is utilized to see heart rhythms. In heart muscles, starting and spread of electrical potential is recorded in ECG banner. Depolarization of chamber is tended to by ECG banner and ventricle in one pattern of each heartbeat [2, 3]. ECG banner is more essential than other natural signs and it has specific morphological properties. This morphology apparently is familiar along examining different

cardiovascular maladies. Banner preparing is familiar along investigating and assessment of ECG signals. The portrayal of ECG banner is six apexes and valleys. The key attributes of human ECG signal are appears in Fig. 1. As seen in Fig. 1, the key waves called P, Q, S, R, & T and so on comprises the significant fragments of ECG signal called QRS complex, ST section, and PR interim [4,5]. The ECG signal has been assorted and comprehend the improvement of exactness estimation and reproducibility. ECG signal is troublesome while examination the sign because of undermined by clamour during obtaining. One of the most testing ECG signal is the extraction of clinical parameters to frame uproarious biomedical signs.
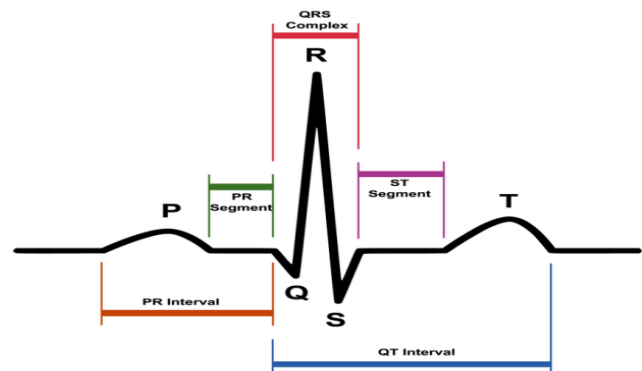
Fig. 1. Characteristics of ECG Signal [5].

By added substance high recurrence commotion, muscle relics or movement and standard meander cover signal in time and recurrence ECG signal is tainted. The issues are antiquities, which happen like an undesirable recurrence in signals. Simple channels can assist along taking care of these issues, yet they can cause nonlinear stage shifts, signal twisting. Computerized channels are progressively precise. The ECG signs can be harmed by various sorts of commotion. The way is required to have the pre-processing strategies to denoise the information ECG signal for the precise examination and finding utilizing the CAD frameworks. There are number of sifting strategies intended to denoise the ECG signals since from a decade ago detailed in [6]-[12], anyway the key worries for ECG signal processing is the multifaceted nature and productivity.

After the pre-processing of ECG signal, the following test is identified along the extraction of five pinnacles, for example, P, Q, R, S, & T from the ECG signal. The Individual pulses in ECG information are basically described by such

five pinnacles and valleys, namely, P, Q, R, S, & T, and these can be utilized to recognize atypical heartbeat in CAD frameworks. For CVD location, the QRS complex and ST portion extraction assumes the huge job, as the both QRS complex and ST fragment contains the key data identified along heart conduct. Accordingly, the dynamic advance in CVD discovery primarily embraces the strategies to extricate the genuine QRS complex area. Various calculations dependent on subordinate [13], computerized channels [14], and wavelet change [15] have as often as possible been utilized for QRS recognition. Along the improvement of equipment condition, considerably more strategies receive wavelet changes. In wavelet-based procedures, the effectiveness of wavelet change emphatically relies upon the decision of the mother wavelets. Other discovery calculations proposed in the literary works including numerical morphology [16], shrouded Markov model [17], S-change [18], standard language structure [19], quadratic channel [20], multiresolution entropy [21], meagre representation [22], and particular worth deterioration (SVD) [23]. Key concern for QRS a complicated and ST segment extraction method has been complexity and accuracy while utilizing along machine learning techniques. The QRS complex and ST segments holds the big and complex data, which may take longer time for the prediction and may leads to inaccurate detection as well. Hence the QRS complex and ST segments needs to optimized further by utilizing the features extraction and normalization techniques before applying the machine learning methods for the early CVD detection.

Under this research, first described the hybrid filtering approach based on notch filtering and high-pass band filtering method for the pre-processing of raw ECG signal. After that designed simple and efficient algorithm for Q, R, S, &T waves extraction and formation of QRS complex and ST segments. The QRS complex and ST segments further exploited for the NHOS based features extraction to boost the accuracy performance and robustness. The features extraction for training ECG dataset is further investigated utilizing five different machine learning based methods. In Section II, the review of recent filtering and CVD detection techniques presented. In Section III, the proposed methodology described. In Section IV, the result analysis, in Section V discussion. Finally in Section VI, conclusion and future work presented.

II. RELATED WORK

The CVD detection is challenging research problems due to the artefacts, baseline wander, noises, efficient extraction of QRS complex and ST segments, features construction and classification phases. This section describes review of current ECG based methods.

In [8], creator proposed the SVD approach for ECG denoising. The ECG signal is pulled in additional by white Gaussian clamour. ECG signal is deteriorating into symmetrical subspaces utilizing high choice estimation range instruments are done in this methodology. The commotion parts contained inside ECG signal that subspaces related to symmetrical change inside particular worth disintegration. SVD disposes of the subspace of clamour and the unfortunate sign parts.

In [9], creator presents the basic whole number coefficient computerized channel along Infinite motivation reaction (IIR) structure for Finite drive reaction (FIR), which is band-stop straightforward number coefficient advanced channel. The structured technique used to sift through the gauge float and force recurrence obstruction, and ECG signal without examination.

In [10], creators structured new versatile strategy for denoising the ECG signal. The proposed multistage technique can dispense along numerous sorts of ancient rarities because of various reference inputs so it tends to be applied to the frameworks where different kinds of impedance are discovered, given that the apriori information about the obstruction is accessible.

In [11], creator proposed late strategies for the pattern meander expulsion and electrical cable obstruction decrease in ECG signals utilizing experimental wavelet change (EWT).

In [12], creator proposed another commotion decrease technique that can apply to crude ECG flags all together acquire a higher sign to-clamour proportion (SNR) for additional processing. Creators utilized the wavelet change and dynamic thresholding way to deal along limit specific sorts of clamour installed in crude ECG signals.

Above all the ECG signal filtering methods, for decision-making process, it's important to work on QRS complex and ST segment extraction, there are many methods presented for the same in [13]-[23]. The features extraction and classification part of CVD detection frameworks, some of the recent methods reviewed below for ECG based disease detection.

In [24], creator presents the technique for proficient framework for acknowledgment of the premature ventricular constriction from the heart maladies thumps and ordinary pulsates. This framework incorporates three fundamental stages: de-noising module, include extraction module and classifier module.

In [25], ECG signal analysis and characterization strategy utilizing wavelet vitality histogram technique and support vector machine (SVM).They structured the cardiovascular arrhythmia discovery in the ECG signal dependent on three phases including ECG signal pre-processing, highlight extraction and pulses order.

In [26], an effective and simple to-interpret strategy of heart illness order proposed dependent on novel component extraction techniques and correlation of classifiers. Creators described the appropriations by test quintiles which beat test implies. Creator examined the highlights extraction technique utilizing three classifiers utilizing measurement decreased highlights by PCA: stepwise discriminate analysis (SDA), SVM, and LASSO calculated relapse.

In [27], the ongoing work on CVD discovery utilizing ECG signal detailed. Creators utilized sign processing and neural systems procedures for processing ECG signals comprising of extricating highlights from ECG signal so as to distinguish the kinds of CVD's.

In [28], the dynamical ECG acknowledgment structure proposed for CVD's and human location utilizing the dynamical neural learning system. The proposed technique comprises of two stages: a preparation stage and a test stage. In the preparation stage, cardiovascular elements inside ECG signals is removed (approximated)accurately by utilizing spiral premise work (RBF) neural systems through deterministic learning component. They got heart framework elements is represented and put away in consistent RBF systems.

In [29], late strategy proposed for the QRS Complex location from the ECG signal utilizing 1-D convolutional neural system (CNN). The CNN comprises of article level and part-level CNNs for extricating distinctive grained ECG morphological highlights consequently. All the removed morphological highlights are utilized by multi-layer recognition (MLP) for QRS complex discovery. Also, creator received the ECG signal pre-processing technique which just contains contrast activity in transient area.

In [31], proposed approach is implemented using ML-libs and Scala language on Apache Spark framework. Main challenge in ECG classification was to handle the irregularities in the ECG signals. ECG heartbeats are grouped under three main categories that is normal, PVC, PAC, and other.

From the recent literature analysis, QRS complex and ST segment detection features extraction may leads to the computation complexity problem, the work proposed in this paper is different in many ways as mainly focused on computationally efficient as well as higher accuracy for CVD detection.

III. METHODOLOGY

This section presents the proposed methodology for CVD detection. Fig. 2 shows the architecture for proposed technique. As observed in figure, the raw ECG signal given as input for CAD system for decision-making process whether it's functioning normal or CVD. In first step, applied the pre-processing algorithm to remove the noises and artefacts without loss of generality along efficiency and robustness in results. The pre-processing ECG signal given input to P, Q, R, S & T waves detection utilizing proposed method, the detected waves are used to form the QRS complex and ST segments. QRS complex and ST segments fused to get the final feature vector as both waves the important information of changes in heart behaviour. As the QRS a complex and ST segment both contains large amount information, this may lead the computationally inefficient while applying the classifier for the detection purpose as well as less detection rate. In this work, applied the normalized HOS technique in which first wavelet packet decomposition method applied to get total 90 unique features. Those features further normalized, hence called as NHOS. Five different classifiers are applied to perform the detection in this work.

A. Pre-Processing

Hybrid filtering method for the ECG signal pre-processing is designed. Algorithm 1 shows the working of proposed pre-processing strategy. The Butter worth channel intended to evacuate the benchmark meandering. Further, second request step channel at 60 Hz is applied, contingent upon the way that signs originate from European or US accounts to expel power-line obstruction. The use of both filtering methods assures that the artefacts available in raw ECG signal also suppressed. After successful removal of unwanted information from the ECG signal, the process of wave's extraction and features formation are applied.
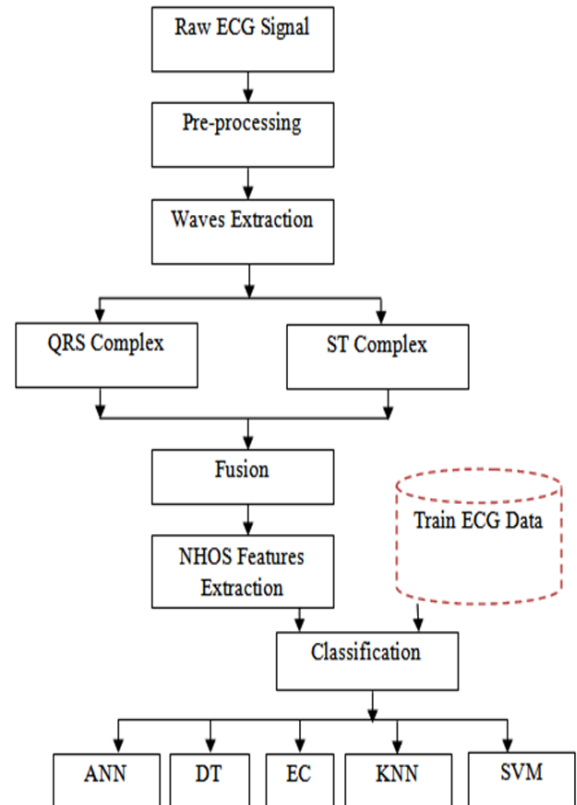
Fig. 2. Proposed CVD Detection System.

Algorithm 1: ECG Signal Denoising
Inputs
S: ECG signal
$\delta = 2$: cut off frequency
$\sigma = 1000$: data sampled frequency
$\varepsilon = 50$ Hz
Output
P: Pre-processed signal
1. Acquire input raw signal S 2. Perform noise insertion explicitly in S 3. Baseline wander removal
$B =$ Butterworth $(S, (\frac{\delta}{(\frac{\sigma}{2})}),$ 'high')
4. Design notch filter
$P =$ notch (B, ε)
5. Return P

B. Waves Extraction

The pre-processed ECG signal used for the waves detection and extraction utilizing the proposed thresholding based method. First extract all the waves such as Q, R, S, P, & T, and then form the QRS complex and ST segment for the features extraction process. Along efficiency of wave's detection and extraction, the main objective is that wave's extraction algorithm should take minimum computation efforts. The technique of thresholding is dynamic in which the threshold value is fixed and computed as per the input ECG signal. This method is start from the signal normalization to QRS complex and ST segment extraction. Algorithm 2 shows the proposed methodology designed in this paper.

As observed in Algorithm 1, the first step is to perform the normalization of pre-processed signal as:

$$A = \frac{P^2}{\max(|P^2|)} \tag{1}$$

Algorithm 2: Waves Extraction

Inputs
P: Pre-processed ECG signal
Output
QRS
ST

1. Compute length *N* of *P*
2. *A*: Signal Normalization utilizing Eq. (1)
3. *AS*: Average signalling utilizing Eq. (5)
4. *α*: Mean of Signal *AS*
5. AS: Apply the threshold on *AS* utilizing Eq. (6)
6. [L, R]: Estimate the left & right waves
7. *Rwave*: Extract *R* wave utilizing Eq. (7)
8. *Qwave*: Extract *Q* wave utilizing Eq. (8)
9. *Swave*: Extract *S* wave utilizing Eq. (9)
10. *Pwave*: Extract *T* wave utilizing Eq. (10)
11. *Twave*: Extract *P* wave utilizing Eq. (11)
12. *QRS*: [*Qwave, Rwave, Swave*]
13. *ST*: [*Swave, Twave*]
14. Return (*QRS, ST*)
15. Stop

Where, P is the pre-processing ECG signal. The normalized signal further used for the computation of average signalling utilizing the convolution operations as:

$$B = \frac{\text{Ones} (1,31)}{31} \tag{2}$$

The temporary array of matrix 1's of size 1*31 which is used to perform the convolution along normalized signal *A* as:

$$C = conv (A, B) \tag{3}$$

$$C = C (15 + [1:N]) \tag{4}$$

The convolution output used to compute the average signal as:

$$AS = \frac{C}{\max(|C|)} \tag{5}$$

The normalization and average signalling of original pre-processed signal helps to minimize the overhead of computation burden while estimating the waves, also improves the accuracy of waves extraction. The dynamic threshold value is computed by taking the mean of *AS* signal. The computed threshold value applied to select the waves those satisfies the threshold value as:

$$AS = AS > \alpha \tag{6}$$

The left and right waves computed utilizing Differences and Approximate Derivatives (DAD) operator. The DAD computes the differences between adjacent elements of AS along the first array dimension whose size does not equal. This returns the signals of matrix starting from -1 (left) to 1 (right). Thus, got the left wave and right wave for the signal. From the left and right waves perform the subtract delay operations to suppress the low-pass and high-pass filtering. After the localization of left and right waves, extract the *P, Q, R, S, & T* waves as utilizing the minimum and maximum operations on *L* (left) & *R* (right) waves detected.

$$Rwave = \max(P(L:R)) \tag{7}$$

$$Qwave = \min(P(L:Rwave)) \tag{8}$$

$$Swave = \min(P(L:R)) \tag{9}$$

$$Pwave = \max(P(L:QwaveQT)) \tag{10}$$

$$Twave = \max(P(Swave:R)) \tag{11}$$

The QRS complex and ST segments formed utilizing the extracted waves.

C. Features Extraction

As the QRS complex and ST segments are having big number of data which may be the inefficient in terms of accuracy as well as computational efforts while applying the machine learning technique for CVD detection, here introduced the higher order statistics (HOS) based features extraction technique [30]. First extract the HOS features utilizing the 4th level wavelet packet decomposition technique, followed by the operator's kurtosis and skewness.

The first and second order statistics not enough to improve the detection accuracy, used the HOS that computes the higher order moments (*m2,m3, m*4) and non-linear combinations of higher order moments which are known as cumulates (*c2, c3, c4*). The cumulates are of second *(c2)*, third *(c3),* and forth *(c4)* order statics of each wavelet packet decomposition coefficient (at each level). Applied this method on QRS+ST matrix which extracts the 90 unique and complex features. The features complex, hence here add the normalization step to HOS outcomes, called as NHOS features. Then perform the normalization of features utilizing the Eq. (1). The NHOS features are used for the classification purpose.

D. Classification

Here five different machine learning techniques for the classification process are used in which the NHOS features of input raw ECG signal consider as test input to detect whether it has CVD or not. The classifiers such as Artificial Neural Network (ANN), k-nearest neighbors (KNN), Support Vector Machine (SVM), Ensemble Classifier (EC), Decision Tree (DT) used to investigate performance of proposed method. For classifiers like SVM, EC, DT, and KNN used the 10-fold cross validation technique, whereas the ANN used the ratio 70%, 15%, 15 % ratio of training, validation, and testing, respectively.

IV. RESULT ANALYSIS

MATLAB is used for CVD detection and Physikalisch-Technische Bundesanstalt (PTB) ECG dataset [32] is used for the experiment. For both heart disease patients and healthy patients ECG signals were recorded by the Professor Michael Oeff, M.D. The 545 ECG signals were collected for different subjects consisting of both CVD and normal. There are total 468 ECG signals affected by CVD and 77 samples are normal.

Here the training is performed on all 545 ECG samples, and then performed the classifications by utilizing five different classifiers. First demonstrate the outcomes of pre-processing algorithm, and then present the results related to accuracy.

Fig. 3 shows the crude ECG signal which is misshaped because of the presence of floats, benchmark meanders, and extra clamor because of the ongoing condition catching procedures. Further the sign ECG is pre-handled utilizing the half and half methodology wherein disentangle the procedure of ECG signal denoising by utilizing the two understand channels at tentatively confirmed setups. Fig. 4 shows the result denoising which shows the reasonable distinction between unique ECG signals and denoised ECG signal. In following graphs, X axis presents milliseconds and Y axis presents millivolts.
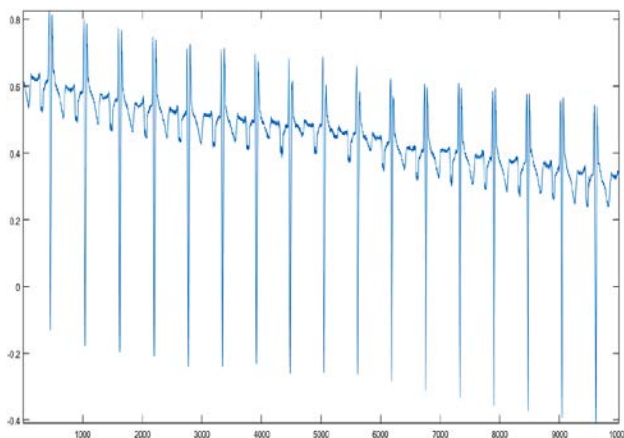
Fig. 3. Original ECG Signal.
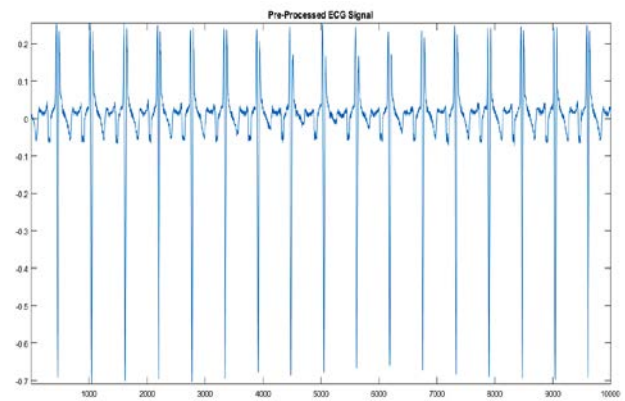
Fig. 4. Pre-Processed ECG Signal utilizing Hybrid Filtering.

V. DISCUSSION

Table I shows the performances in terms of Mean Square Error (MSE), Signal to Noise Ratio (SNR), and time complexity. The average performance is measured for 50 ECG signals from the given dataset. The time complexity is shows nothing more difference in time complexity performance. Suggested way is effective and efficient for preprocessing of ECG signal.

This paper present the performance investigation of proposed method called QRS_ST + NHOS utilizing five different classifiers. The proposed model is compared along QRS_ST (without utilizing any features extraction technique), and QRS_ST+HOS (without normalization) to claim the efficiency. Table II shows the detection accuracy performance in terms of for all classifiers.

As observed in Tables II to IV, the performance of proposed QRS_ST + NHOS shows the significant improvement in accuracy, precision, and recall rates compared to the QRS_ST and QRS_ST+HOS. In an average the ANN classifier shows the best performance as compared to other classifiers. The DT classifier shows the worst performance among other classifiers. The results of QRS_ST demonstrates that, using only QRS and ST segments, the classification results is not improved, but by using the said features extraction technique the performance is enhanced. Similarly, the using NHOS demonstrates the benefits of applying the normalization technique on extracted features. Table V shows the average computation time required for CVD detection and classification time (using KNN classifier). As the size of QRS_ST is larger, it takes long time for the classifier to perform the features matching, hence takes more time for the detection as compared to QRS_ST+HOS and QRS_ST+NHOS. There is negligible difference between both HOS based model and NHOS based model, as NHOS mainly introduced to improve the accuracy rate.

Finally, the accuracy comparison along similar works presented in Table VI. The results show the performance improvement utilizing proposed model over the methods presented recently.

TABLE I.   PRE-PROCESSING ALGORITHM ANALYSIS

Methods	MSE	SNR	Time Complexity (seconds)
Butterworth	0.3545	10.55	0.148
Notch	0.3323	10.78	0.112
Adaptive Filtering	0.2566	11.89	0.193
Hybrid	0.1973	13.33	0.156

TABLE II.   CVD DETECTION ACCURACY (%) ANALYSIS

	QRS_ST	QRS_ST+HOS	QRS_ST+NHOS
ANN	89.77	91.98	94.47
DT	80.47	84.11	86.82
EC	83.29	88.35	91.76
KNN	82.58	90.23	92.23
SVM	76.94	89.76	91.76

TABLE III.   CVD DETECTION PRECISION (%) ANALYSIS

	QRS_ST	QRS_ST+HOS	QRS_ST+NHOS
ANN	89.77	91.98	94.47
DT	88.66	84.11	86.82
EC	84.17	88.35	91.76
KNN	85.42	90.23	92.23
SVM	85.24	89.76	91.76

TABLE IV.   CVD DETECTION RECALL (%) ANALYSIS

	QRS_ST	QRS_ST+HOS	QRS_ST+NHOS
ANN	89.77	91.98	95.47
DT	65.93	75.38	74.65
EC	97.01	99.33	91.65
KNN	88.01	78.67	78.67
SVM	89.29	100	100

TABLE V.   COMPUTATION EFFORTS ANALYSIS (SECONDS)

	QRS_ST	QRS_ST+HOS	QRS_ST+NHOS
CVD Detection	3.79	3.25	3.23
Classification time	1.89	1.45	1.42

TABLE VI.   ACCURACY RATE EVALUATION ALONG SIMILAR WORKS

	Accuracy (%)
[26]	88.52
[27]	82.5
[29]	92.5
Proposed	94.47

VI. CONCLUSION AND FUTURE WORK

The aim of this paper is to present the methodology for the CVD detection utilizing the ECG signals. The proposed framework is consisting of pre-processing, wave's extraction, features extraction and classification. At first step, paper proposed the hybrid filtering technique to filter out different type's noises and artefacts. The proposed hybrid filtering technique is simulated and evaluated along existing filtering techniques in terms of MSE, SNR and Time complexity parameters. For wave's extraction, the lightweight and efficient dynamic threshold based algorithm are proposed. Here first locate the left and right waves, and then extract the Q, R, S, T, & P waves. Utilizing the extracted waves, QRS complex and ST segments extracted. To minimize the computational efforts and improve the accuracy, here applied the NHOS features extraction technique on extracted QRS complex and ST segments. In classification step, five different classifiers applied and investigate their performances in terms of accuracy, precision, recall, and computational time. The proposed method shows the significant improvement over the recent methods. For future work, use of different ECG heart datasets are suggested.

REFERENCES

[1] Tu Ya, Zhou Runjing, Zhang Fei, "ECG Signal Preprocessing Based on Change Step Iteration of the LMS Adaptive Filtering Algorithm", 2009 World Congress on Computer Science & Information Engineering.

[2] D. Balasubramaniam, D. Nedumaran, "Implementation of ECG Signal Processing & Analysis Techniques in Digital Signal Processor based System", MeMeA 2009 - International Workshop on Medical Measurements & Applications, Cetraro, Italy, May 29-30, 2009.

[3] Ioan Tudosa, Narcis Iulian Adochiei, & Razvan Ciobotariu, "New Aspects in ECG Signal Processing Utilizing Adaptive Filters", 7th International Symposium On Advanced Topics in Electrical Engineering (ATEE), 2011.

[4] Lukas Smolarik, Adriana Libosvarova, Dusan Mudroncik, Peter Schreiber, "Non-contact ECG Signal Processing", 6th IEEE International Conference Intelligent Systems, 2012.

[5] Anil Chacko & Samit Ari, "Denoising of ECG Signals utilizing Empirical Mode Decomposition Based Technique", IEEE International Conference on Advances in Engineering, Science & Management (ICAESM), pp. 6 – 9, 2012.

[6] Mehmet Ustundag, Muammer Gokbulut, Abdulkadir Sengur & Fikret Ata, "Denoising of Weak ECG Signals by utilizing Wavelet Analysis & Fuzzy Thresholding", Springer Network Modeling Analysis in Health Informatics & Bioinformatics , Volume 1, Issue 4, pp 135-140, 2012.

[7] Lukas Smital, Martin Vitek, Jiri Kozumplik, & Ivo Provaznik, "Adaptive Wavelet Wiener Filtering of ECG Signals", IEEE Transactions On Biomedical Engineering, Volume 60, Issue 2, pp. 437 - 445, 2013.

[8] Mohammed Assam Ouali & Kheireddine Chafaa, "SVD-Based Method for ECG Denoising", IEEE International Conference on Computer Applications Technology (ICCAT), pp. 1 - 4, 2013.

[9] Dong Jingwei, Jiang Wenwen, "Design of Digital Filter on ECG Signal Processing", Fifth International Conference on Instrumentation & Measurement, Computer, Communication & Control, 2015.

[10] Rizwan Qureshi, Muhammad Uzair, Khurram Khurshid, "Multistage Adaptive Filter for ECG Signal Processing", International Conference on Communication, Computing & Digital Systems (C-CODE), 2017.

[11] Omkar Singh, Ramesh Kumar Sunkaria, "ECG signal denoising via empirical wavelet transform", Australas Phys Eng Sci Med, 2017.

[12] Pandit D., Zhang L., Liu C., Aslam N., Chattopadhyay S., Lim C.P. "Noise Reduction in ECG Signals Utilizing Wavelet Transform & Dynamic Thresholding", In: Bhatti A., Lee K., Garmestani H., Lim C. (eds) Emerging Trends in Neuro Engineering & Neural Computation. Series in BioEngineering. Springer, Singapore, 2017.

[13] Arzeno NM, Deng ZD, Poon CS, "Analysis of first-derivative based QRS detection algorithms", IEEE Trans Biomed Eng. 2008;55:478–484.

[14] Keselbrener L, Keselbrener M, Akselrod S, "Nonlinear high pass filter for R-wave detection in ECG signal", Med Eng Phys. 1997;19:481–484.

[15] Sharma T, Sharma KK, "QRS complex detection in ECG signals utilizing locally adaptive weighted total variation denoising", Comput Biol Med. 2017;87:187–199.

[16] Trahanias PE, "An approach to QRS complex detection utilizing mathematical morphology", IEEE Trans Biomed Eng. 1993;40:201–205.

[17] Coast DA, Stern RM, Cano GG, Briller SA, "An approach to cardiac arrhythmia analysis utilizing hidden Markov models", IEEE Trans Biomed Eng. 1990;37:826–836.

[18] Zidelmal Z, Amirou A, Adnane M, Belouchrani A, "QRS detection based on wavelet coefficients", Comput Methods Programs Biomed. 2012;107:490–496.

[19] Hamdi S, Ben AA, Bedoui MH, "Real time QRS complex detection utilizing DFA & regular grammar", Biomed Eng Online. 2017;16:31.

[20] Phukpattaranont P., "QRS detection algorithm based on the quadratic filter", Expert Syst Appl. 2015;42:4867–4877.

[21] Farashi S., "A multiresolution time-dependent entropy method for QRS complex detection", Biomed Signal Process Control. 2016;24:63–71.

[22] Zhou Y, Hu X, Tang Z, Ahn AC., "Sparse representation-based ECG signal enhancement & QRS detection", Physiol Meas. 2016;37:2093.

[23] Jung WH, Lee SG., "An R-peak detection method that uses an SVD filter & a search back system", Comput Methods Programs Biomed. 2012;108:1121–1132.

[24] Ataollah Ebrahim Zadeh, Ali Khazaee, Vahid Ranaee, "Classification of the electrocardiogram signals utilizing supervised classifiers & efficient features", computer methods & programs in biomedicine vol. 99, pp.179–194, 2010.

[25] Prof. Alka S. Barhatte, Dr. Rajesh Ghongade, Abhishek S. Thakare, "QRS Complex Detection & Arrhythmia Classification utilizing SVM", 2015 International Conference on Communication, Control & Intelligent Systems (CCIS).

[26] Huang, Rong & Yingchun Zhou. "Disease Classification & Biomarker Discovery Utilizing ECG Data" BioMed research international vol. 2015 (2015): 680381.

[27] Shalin Savalia, Eder Acosta, Vahid Emamian, "Classification of Cardiovascular Disease Utilizing Feature Extraction & Artificial Neural Networks", Journal of Biosciences & Medicines, vol. 5, pp. 64-79, 2017.

[28] Muqing Deng a, Cong Wangb, Min Tang c, Tongjia Zheng, "Extracting cardiac dynamics within ECG signal for human identification & cardiovascular diseases classification", Neural Networks vol. 100, pp. 70–83, 2018.

[29] Xiang, Yande et al. "Automatic QRS complex detection utilizing two-level convolutional neural network" Biomedical engineering online vol. 17,1 13. 29 Jan. 2018.

[30] Yakup Kutlu, Damla Kuntalp, "Feature extraction for ECG heartbeats utilizing higher order statistics of WPD coefficients", Comput Methods Programs Biomed, vol. 105, no. 3, pp.57-67, 2012.

[31] Fajr Ibrahem Alarsan, Mamoon Younes "Analysis and classification of diseases using heartbeat features and machine learning algorithms" Springer open, journal of Big data, pp 1-15, 2019.

[32] https://www.physionet.org/physiobank/database/ptbdb/.

Facilitating the Detection of ASD in Ultrasound Video using RHOOF and SVM

Mrunal Ninad Annadate[1] iD

Department of Electronics and Telecommunication
MIT College of Engineering, Survey No. 124
Shivtirthanagar, Paud Road, Kothrud, Pune 411038
Maharashtra, India

Manoj Nagmode[2] iD

Department of Electronics and Telecommunication
Government College of Engineering and Research
Avasari (Khurd), Taluka-Ambegaon, Pune 412405
Maharashtra, India

Abstract—In the medical field various motion tracking techniques like block matching, optical flow, and histogram of oriented optical flow (HOOF) are being experimented for the abnormality detection. The information furnished by the existing techniques is inadequate for medical diagnosis. This technique has an inherent drawback, as the entire image is considered for motion vector calculation, increasing the time complexity. Also, the motion vectors of unwanted objects are getting accounted during abnormality detection, leading to misidentification / misdiagnosis. In this research, our main objective is to focus more on the region of abnormality by avoiding the unwanted motion vectors from the rest of the portion of the heart, allowing better time complexity. Proposed a region-based HOOF (RHOOF) for blood motion tracking and estimation; after experimentation, it is observed that RHOOF is four times faster than HOOF. The performance of supervised machine learning techniques was evaluated based on accuracy, precision, sensitivity, specificity, and area under the curve. In the medical field more importance is given to the sensitivity than accuracy. Support vector machine (SVM) has outperformed other technique on sensitivity and time complexity, hence chosen for abnormality classification in this work. An algorithm has been devised to use combination of RHOOF and SVM for the detection of atrial septal defect (ASD).

Keywords—Two dimensional; apical four chamber; region-based histograms of oriented optical flow; machine learning; area under the curve; support vector machine; congenital

I. INTRODUCTION

Cardiac diseases are mainly categorized into congenital and acquired. Congenital diseases are birth defects. Many of them can be treated or prevented by taking necessary corrective measures like surgeries. Acquired diseases have an impact on the overall functioning of the heart. Hence, more emphasize is given to the physical structure and overall functioning of the heart. Representation and evaluation of cardiovascular movements have become a significant step in the analysis. Heart-related disorders/diseases can be visualized and analyzed using an ultrasound, which is the easiest and most widely used real-time method. The ultrasound findings vary from patient to patient due to the differences in the structure of the heart. Hence, experts play a very vital role in capturing correct ultrasound images, accurate analysis and interpretation of results. As there is a difference in the results of a heart ultrasound, appearance-based methods are not recommended

for the machine learning (ML) algorithm classification of cardiovascular disease.

The overall objective of this study is to devise an algorithm for facilitating automatic detection of atrial septal defect (ASD) using motion vector analysis of ultrasound videos in combination with ML algorithms.

ML is a next-generation technology used worldwide in various fields and applications. Medical field is no exception to this technology. Nowadays, ML technology is increasingly used worldwide. Several studies have been conducted in the medical field using ML as one of the enablers for early and accurate detection of diseases and because of the availability of data in the form of electronic media. A huge amount of data in the form of numbers, images, videos, pictures, and reports are generated and transmitted in the medical field on a daily basis. With the appropriate use of new technology, real-time data can be provided as and when required, regardless of the location. Because of the advantages over other diagnostics techniques, ultrasound is one of the most popular techniques used in the medical field.

There is an ongoing study evaluating the ultrasound method in terms of real-time analysis, quality of the output, and disease coverage, which provided a very good opportunity for researchers to use this next-generation technology; through this experimentation, ML and other computer vision techniques were used for the prediction of heart abnormalities. Recently, researchers have evaluated the utilization of several techniques like optical flow (OF), BPNN and SVM [9], in the detection of cardiovascular abnormalities.

The field of computer vision has described several OF estimation techniques. An OF measures the angle and magnitude of brightness patterns in an image or moving objects and plots them in the form of vectors called motion vectors (MVs). OF provides lot of information on the spatial patterns based on the rate of change.

This study used the image and video processing techniques in combination with ML for the classification of ASD. As this defect usually develops in the upper chambers of the heart, our study focused more on the detection of abnormality, that is, the hole in the septal wall. Using a computer vision technology to identify the location of defects, the blood movement from the left to right atrium (LA to RA) is studied. The MVs in this area

were examined closely to establish the pattern; using the MV pattern and direction, the ML model has been trained. This trained model is then used for testing.
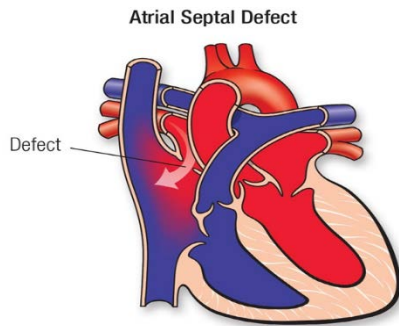
Fig. 1. Heart with Atrial Septal Defect (Source: Heart.Org).

As seen in the "Fig. 1", hole in the septal wall allows the movement of blood from the LA to RA or vice versa. A dynamic analysis of the ultrasound findings assists in detection of ASD. The movement of blood can be tracked using motion estimation analysis methods.

The sections were arranged as below:

Section II: Related Work.

Section III: Methods.

Section IV: Experimental Results.

Section V: Conclusion.

Section VI: Future Work.

II. RELATED WORK

Mrunal N. Annadate and Manoj S. Nagmode, in [1] proposed a motion vectors-based algorithm using block matching technique (Elongated Horizontal Large Diamond Search Pattern – EHLDSP) for tracking of Blood Movement in the heart for enabling the detection of Atrial Septal Defect. Mrunal N. Annadate and Manoj S. Nagmode, in [2], proposed an automatic detection of heart disorder (Hypertrophic Cardiomyopathy) using watershed segmentation technique. Achmad Solichin, Agfianto Eko and Agus Harjoko, in [3] proposed an OF feature-based method to detect the direction of an object movement. Author proved the usage of HOOF for the detection of object movement direction in the video. They also emphasized on the three main parameters affecting the success of results; number of frames, their interval and size of the grid. Results obtained for accuracy, precision, recall and direction detection error rate are 98.1%, 35.6%, 41.2% and 25.28%, respectively. The proposed approach is 22 times faster than other approaches which uses object detection and segmentation as pre-step. Mrunal N. Annadate and Manoj S. Nagmode, in their previous study [4] they have worked on the de-noising filters of seven different categories and proved that Three Ranking and Lee outperformed other de-noising filters. Rashmi, Nagaraja NS and Ashwin Kumar, in [5] presented HOFO- and support vector machine (SVM)-based techniques for detection of abnormal behavior. Here the research is accomplished in two parts, computing the histogram of optical flow orientation followed by a non-linear one-class Support Vector Machine. For optical flow Lucas Kanade method was used. Rensso Victor Hugo Mora Colque, Matheus Toledo Lustosa de Andrade, Carlos Caetano, and William Robson Schwartz, in [6] showcased the detection of anomalous events in videos. Here author proved that proposed descriptor is capable of extracting the information from cuboids. Unlike their previous work, in this research both magnitude and direction of the optical flow was used for the histogram generation. Also, showcased the improvements by comparing the results obtained with conventional HOOF. Database used by the author has various limitations like size, saliency of the anomalies and evaluation criteria.Achmad Solichin, Agfianto Eko Putra and Agus Harjoko, in [7] proposed a Histogram of Oriented OF (HOOF)-based method for determination of movement direction. The experiment results show False Positive Per Grid (FPPG) as 28.32%, and False Negative Per Grid (FNPG) as 4.08%. The role of grid size in faster identification of movements is proved in this research. Chetana D Patil and Bharathi V K, in [8] proposed Histogram of OF Orientation and Magnitude (HOFM) and K-nearest neighbor (KNN) classifiers for abnormal event detection and classification. Not only orientation measurement using temporal information but also velocity extraction using magnitude of flow vectors is done by the researcher to achieve the event detection. G.N.Balajia, N.Chidambaramc, and T.S.Subashinib, in [9] proposed method performs well for the automatic classification of echocardiogram in all the four standard views; parasternal short-axis, parasternal long-axis, apical 2 chamber and apical four chamber, using back propagation neural network (BPNN) and SVM with an 87.5% accuracy. Limitation of the existing study is non-inclusion of views such as subcostal view, Doppler view, etc. J.R.R. Uijlings, N. Rostamzadeh, I.C. Duta, and N. Sebe, in [10] presented various techniques to address the problems related to computational efficiency. The focus of this research is to evaluate the trade-off between computation time and accuracy for video classification. Authors have achieved this through fast Matlab implementation of HOG and HOF and made this publicly available.

Daniel Tenbrinck, Xiaoyi Jiang, Sönke Schmid, Jörg Stypmann, and Klaus Schäfers, in [11] described, how presence of speckle noise violates the Intensity Constancy Constraint' (ICC) assumption of optical flow techniques. To overcome this issue, they have proposed use of local statistics and introduced an optical flow method using histograms as discrete representations for motion analysis. Tian Wang Hichem Snoussi, in [12] proposed a novel spatiotemporal feature descriptor, called histogram of optical flow orientation, magnitude and entropy (HOFME), which is an improvised version of HOFM. On the data set, proposed algorithm showed 100% and 83% of accuracy in recognizing anomalous events at two different locations. Alessandro Becciu, Luc Florack, Hans van Assen, Vivian Roode, Sebastian Kozerke, and Bart M, in [13] proposed a new three-dimensional (3D) multiscale OF-based method for analyzing a true 3D cardiac motion at voxel precision. Rizwan Chaudhry, Gregory Hager, Ren´e Vidal and Avinash Ravichandran, in [14] proposed usage of generalization of the Binet-Cauchy kernels to nonlinear dynamical systems (NLDS) and represented each frame of a video using a histogram of oriented optical flow (HOOF) to

recognize human actions by classifying HOOF time-series. Balza Achmad, Aini Hussain, and Mohd. Marzuki Mustafa, in [15] extracted three successive frames from an ultrasound video and presented an OF-based enhancement technique. The focus of the research in this paper is to remove speckle. Author made use of optical flow technique in three consecutive frames and reconstructing the resultant image with reference to the preceding and succeeding image using fusion and Lukas-Kanade method for obtaining optical flow. F Wang, D Beymer, and T Syeda-Mahmood, in [16] presented a method of echocardiographic sequence registration. Aditi Roy, Arun K. Majumdar, Shamik Sural, and Jayanta Mukherjee, in [17] represented an echocardiogram video in the form of hierarchical state-based model. Author carried out experiments on 20 echo videos and compared the results with manual annotation done by two experts. View classification accuracy is 97.19% and misclassification error is in the acceptable range (less than 13%), and corresponds to the frames of state boundaries. Schlomo V Aschkenasy et al. in [18] presented an algorithm for the automatic classification of cardiac views like A4C, A2C, and PLAX. Navneet Dalal and Bill Triggs, in [19] experimentally proved that HOG descriptors outperform existing gradient and edge descriptors. Further optimization of algorithm and increase in the speed of detection are the areas of improvement. John L. Barron, Steven S. Beauchemin, and David J. Fleet, in [20] carried out the empirical comparison between different techniques to overcome the shortage of quantitative evaluation of optical flow in other researchers' studies. In this study, evaluation is done on the basis of density of velocity measurement, accuracy, and reliability.

Berthold KP Horn and Brian G Schunck in [21] presented a method to find optical flow patterns, based on the assumption of smooth variation of brightness pattern of the velocity across the image. This method is popularly used by the researchers till date. Bruce D. Lucas, and Takeo Kanade, in [22] presented a spatial intensity gradient-based innovative technique used for image registration. The differential method presented by the authors for optical flow estimation assumes constant flow in a local neighborhood of the pixel under consideration, and solves the basic optical flow equations for all the pixels in that neighborhood. Navneet Dalal, Bill Triggs and Cordelia Schmid, in [23] proved that a differential OF histogram provides the best performing motion-based descriptors. The combined detector reduces the false alarm rate by a factor of 10 relative to the best appearance-based detector. J. H. Park, C. Simopoulos, J. Otsuki, S. K. Zhou1, and D. Comaniciu, in [24] built a system for automatic classification of the heart views through use of machine learning which extracts the knowledge from an annotated database. This system automatically classifies four standard cardiac views: apical four chamber and apical two chamber, parasternal long axis and parasternal short axis. Author achieved a classification accuracy of 96%. Current system can handle pre-defined views only. SS Beauchemin and JL Barron, in [25] investigated the computations performed using the OF methods and further scrutinized the hypothesis and assumptions used by these methods. .Matthew Eric Otey et al., in [26] used Markov Random Fields (MRF) model for part-based representation and automatic recognition and grouping of the heart chambers in the different views. Current system can handle simple distributions for the properties of parts, complex distributions are not possible at this stage. Rensso Victor Hugo Mora Colque, William Robson Schwartz, and Carlos Antˆonio Caetano J´unior, in [27] proposed a feature descriptor called HOFM for detection of an anomalous event.

III. METHODS

The dataset chosen for experimentation comprised 20 MPEG4 (MP4) videos because AVI does not support HEVC/H265 codecs. The length of each video is considered as 1 second because one heartbeat or cardiac cycle takes about 0.8 sec to complete. The images are cropped from these videos to the same size to avoid impact on the overall analysis. All our videos have a standard frame rate of 30fps.

A. Proposed Methodology – Region-based HOOF (RHOOF)

"Fig. 2", demonstrate the block schematic of the end-to-end process flow for abnormality detection; each step in this block diagram is covered in more detail in the subsequent sections of this study.
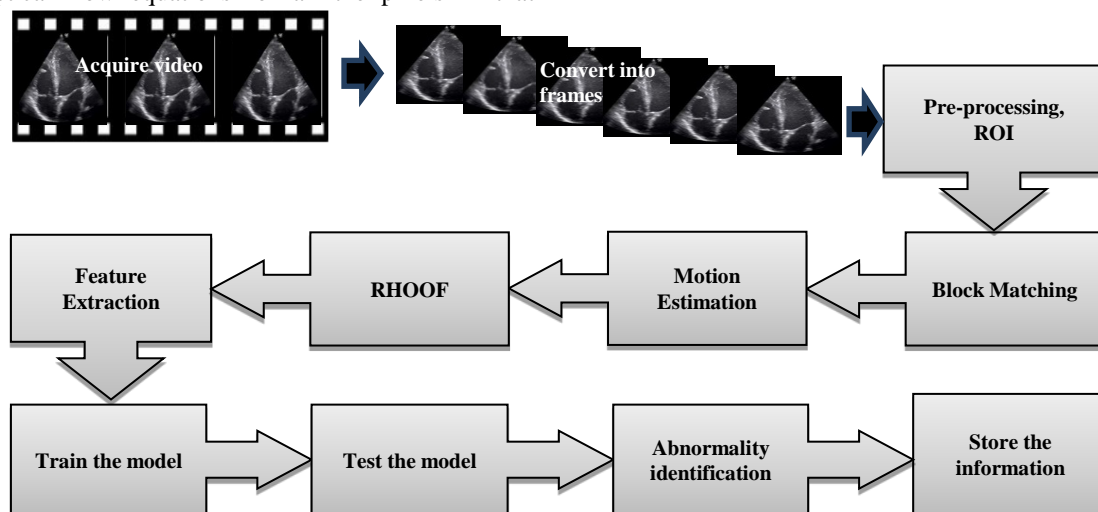
Fig. 2.   Block Schematic of a Proposed Methodology.

B. Pre-processing[4]

Multiple noise models and de-noising filters were studied and evaluated based on various performance parameters, as part of a pre-processing combination of best performing Lee and the three ranking filters were used for de-noising, as reported in a previous study.

C. ROI Identification

During run time image cropping is done to focus on the septal wall area only, and the algorithm automatically marks the ROI in each frame. By doing this we are avoiding the unwanted motion vectors from other regions or rest of the portion of the heart, as a result ASD identification and diagnosis becomes easy and accurate.

D. Motion Estimation[1]

In the last decade, several studies regarding motion estimation have been conducted. Motion analysis is performed using two conventional methods: Block Matching and Optical Flow. These techniques were used to quantify the velocity pattern of moving objects. The distribution of velocity movement based on the brightness pattern is evaluated in the image sequence of the video. The three-dimensional arrangement of the object and rate of change-specific information can be derived from these image sequences.

1) Optical Flow (OF)[1]: The motion estimation in terms of image intensity gradient change can be evaluated based on OF. The vector that provides movement between pixels of each frame is called OF. The sequence of continuous frames makes a video, and its signal is a three-dimensional function (f) of m, n, and t (where m and n are spatial or plane coordinates and t is time) as shown in equation 1. The assumption of brightness consistency across successive frames can be stated using the following equation:

$$f(m,n,t) = f(m+dm, n+dn, t+dt) \tag{1}$$

Where, m+dm and n+dn are displacements and t+dt is new time.

Here, left-hand side of the "(1)" is the 1st frame and right-hand side is the subsequent frame; as mentioned above, brightness across the frame is constant, which means that brightness at m and n of the first frame is equivalent to m+dm and n+dn in the subsequent frame. Equation 1 can be represented as follows:

$$I_x m + I_y n + I_t = 0 \tag{2}$$

Where, spatiotemporal image brightness derivatives are defined by Ix, Iy, and It. And m is horizontal and "n" is vertical optical flow and "t" is time.

2) Block Matching[1]: Here current frame of a video is divided into macro blocks of equal sizes and comparison between macro blocks of adjacent frames is carried out. Location wise movement of macroblock is traced by drawing a vector called as motion vector (MV). Likewise, movement of all macro blocks in the frame is calculated to estimate the motion.
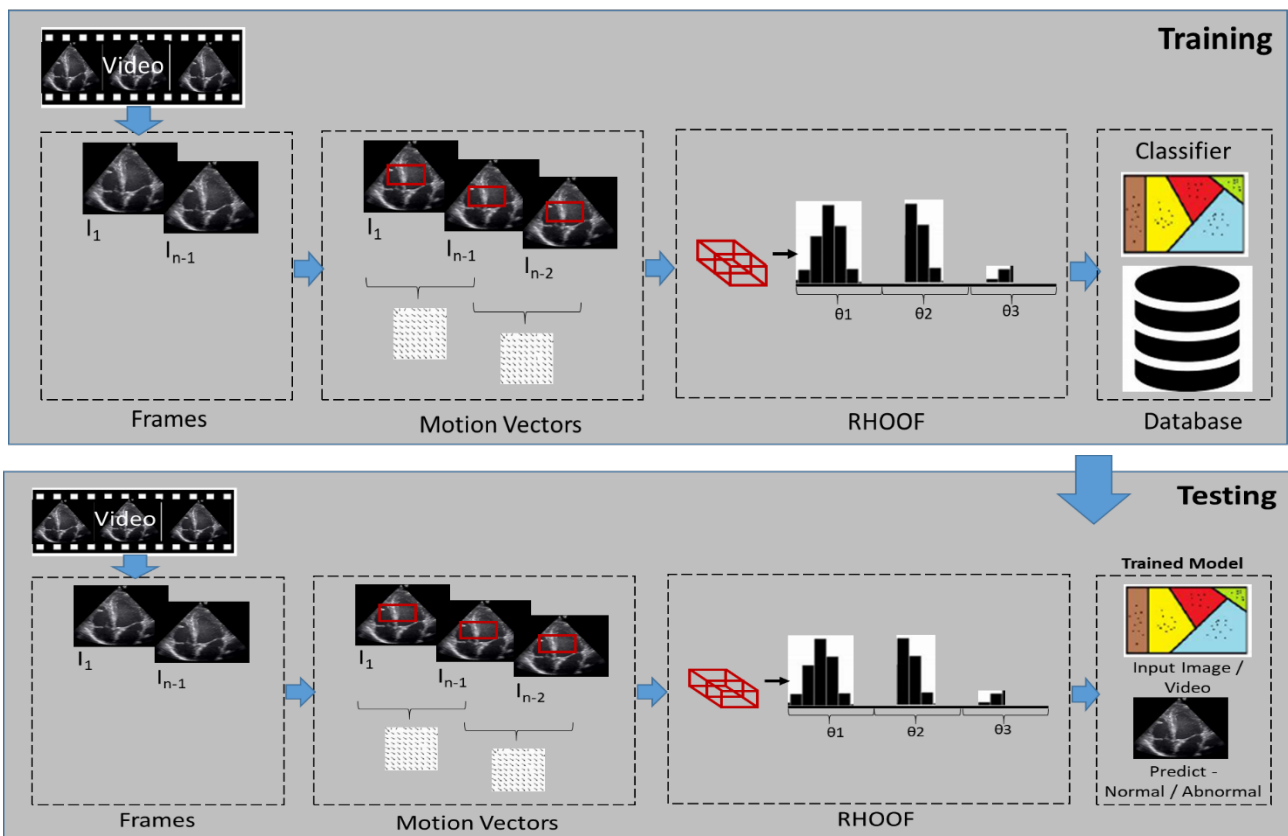
Fig. 3.   Diagrammatic Illustration of a Proposed RHOOF based Approach to Classify Heart Abnormality.

"Fig. 3" shows the diagrammatic illustration of proposed RHOOF, wherein the acquired video was converted into frames using an algorithm; the region of interest was identified; the optical flow was extracted for each frame to calculate MV; and the features like magnitude and direction were provided as input to the algorithm to calculate region-based HOOF (RHOOF). These inputs were further fed into the ML algorithm for heart abnormality classification. For each block, a RHOOF feature was computed, and the vector was created and denoted as Fx for frame x as shown in the figure. Every frame in RHOOF was denoted as a vector of n x n. Equally spaced bins were spread over 0 degree to 360 degrees. As seen in the "Fig. 4", two neighboring blocks overlap 50% on each other during RHOOF calculation. The computation time depends on the size of the block.

The values extracted from motion vector has horizontal and vertical components: U and V at pixel(x,y), respectively. Angle θ denotes motion vector direction and whose value is between 0 degree and 360 degrees. It was derived based on U and V values. The direction and magnitude of the motion vector are calculated as follows:

$$\theta(x,y) = \tan^{-1}\frac{V(x,y)}{U(x,y)} \qquad (3)$$

And magnitude as

$$magnitude = \sqrt{x^2 + y^2} \qquad (4)$$
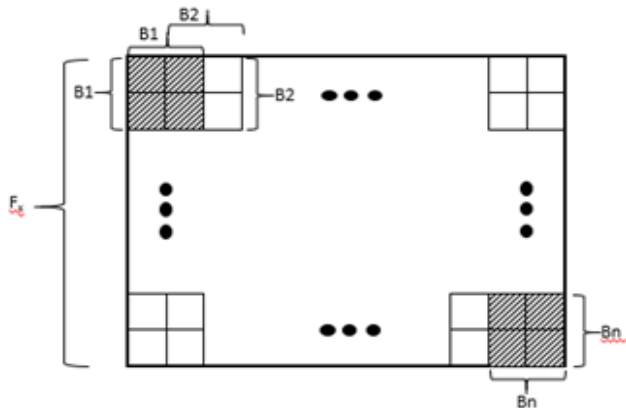
Fig. 4. Region-based Histogram of Optical Orientation of Frame k.

After obtaining the direction and magnitude of the motion vector at each point, it was normalized and the movement directions were clubbed into various Bins like Bin1, 2,..,9, as per θ value. No movement direction points were denoted as Nil. RHOOF features were analyzed to identify the movement direction, and histogram was obtained. Grid processing will not process the grid if there is no motion in it; this reduces the overall processing time. The RHOOF values were accumulated using the following equation:

$$Hoof_b(m,n) = \sum_{frame=1}^{P} Hoof_b(m,n) \qquad (5)$$

Where, "b" stands for bin, P is the number of frames and (m,n) values corresponds to grid(m,n).

The representation of histograms is not dependent on motion directions.

"Fig. 5" explains how the histogram magnitude is distributed in nine bins.

E. Classification

ML is a subset of artificial intelligence and used for data analysis so that the system can learn from data and patterns, and decide on its own with/without human intervention. In the initial stage, started with pattern recognition, now various algorithms are developed to perform many tasks. Through an artificial intelligence, computer can identify whether the trained model is exposed to a new set of data and automatically re-train the model.

ML is not a new technique, has been used for a long time, and is applied in today's applications; its use has given a new face and momentum to it. ML is broadly classified into two classes, supervised and unsupervised. In the first category, the model is trained using labeled examples; the desired output is already known based on the input. In the second category, there is no historical labels available against the input data, the system or algorithm is trained to find some structure or pattern within, and works best on transactional data.

Various ML algorithms are applied to determine the best performing classifier; the following section describes the various ML algorithms that were used in the current study.
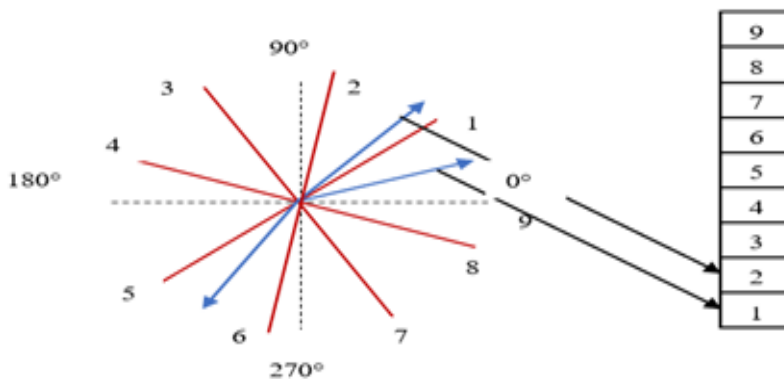
Fig. 5. Motion Vector Distribution in Nine Bins.

Decision tree analysis is a predictive modelling tool for supervised learning. It is constructed using an algorithmic approach by splitting the data set based on certain conditions, subsequently used for classification and regression tasks. The model is created based on the "if-then-else" principle or series of questions and conditions to predict the value of the target variable; if there are more validation conditions, the tree will be more complex.

Labelled training data is used to build a predictive model in discriminant analysis. It is further divided into two categories, linear and quadratic. In this technique, it is assumed that within a group there exists an equal covariance matrix and normally distributed variables. A hyper-plane is created based on the number of classifying variables to minimize the miscalculations constructed between two different groups.

Logistic regression is a predictive analysis used to define the data, because of its association with dependent binary variables and other independent variables like ordinal or nominal. In this technique, it is assumed that the dependent variable is binary, and there is no high correlation within predictors. The good example could be to calculate the probability of having a heart attack through analysis of influential parameters like body weight, calories/fat intake, and age.

SVM is another technique used for regression and classification. In this technique, every data element is plotted as a point in the feature dimension space. The hyper plane is drawn to best fit the two classes; based on that, classification is carried out. Identifying the right hyper plane is very crucial in SVM; usually, the plane that divides the two classes well is chosen as a hyper plane.

KNN is a supervised algorithm referred for classification and regression; this method is easy and simple to implement. The algorithm works on the principle that similar things are near or close to each other, or they are in close proximity by capturing the idea of similarity. Choosing the right value of K is of prime importance in this algorithm; the value that reduces the errors during iterations and has the ability to make accurate predictions is chosen as the best K value. If the value of K decreases, the predictions become less stable and vice versa.

An ensemble is ML algorithm that makes use of various base algorithms or a combination of algorithms to predict the optimal model. The use of a combination of algorithms provides better predictive models in certain situations. The purpose of combining various algorithms is to decrease bias (boosting), reduce variance (bagging), and increase predictions. This technique is further divided into two categories, sequential (e.g., AdaBoost) and parallel (e.g., Random Forest). Homogeneous base learners make use of a single base learning algorithm, whereas heterogeneous learners make use of a multiple learning algorithm.

Images of twenty videos were given as input to the ML algorithms/models. We have split the data into training and testing in the ratio of 80:20.

F. Training the Model

All of the models mentioned in previous sections were applied for analysis; the best performing model is chosen for further testing.

G. Testing the Model

Once the best performing model is selected for prediction, a test video is provided as an input to the trained model to check their performance and the output is recorded.

H. Identification of Abnormalities

Based on training, the selected model will predict whether input video is normal or abnormal.

I. Storing the Information

Once the abnormality is identified, the same gets stored against the video in the database; that is, the said video is marked as normal/abnormal in the database for future reference and training.

Is carried out on the acquired dataset in three phases as explained below. Various ML techniques performance is evaluated based on the following statistical parameters.

IV. EXPERIMENTAL RESULTS

Performance parameters:

$$\text{Accuracy} = (TP+TN) / (TP+FP+TN+FN) \qquad (6)$$

Where, TP= true positive, TN=true negative, FP=false positive, and FN=false negative.

$$\text{Precision} = TP/(TP+FP) \qquad (7)$$

$$\text{Sensitivity} = TP/(TP+FN) \qquad (8)$$

$$\text{Specificity} = TN/(TN+FP) \qquad (9)$$

$$\text{Area under the curve (AUC)} \qquad (10)$$

Phase 1: Process of calculating OF

1) The video is fed and converted into frames
2) The region for analysis is determined.
3) Pre-processing procedures are carried out like de-noising and region cropping.
4) Two consecutive frames are read.
5) $f_x d_x$, $f_y d_y$, $f_t d_t$, and MV are calculated.
6) The MV is plotted.
7) The next two consecutive frames are obtained.
8) Steps 5 to 7 are repeated.

Phase 2: Process of calculating RHOOF

1) The MVs are obtained from the OF.
2) The features like a magnitude and direction of the MVs using above "(3)" and "(4)", respectively.
3) The values obtained in step 2 are divided into nine different bins depending on the "Θ" value and spread over "0" degree to "360" degrees.

The bin values are fed into the algorithm to calculate the histogram.

The histograms of RHOOF are calculated.

The histogram is plotted.

Phase 3: Process of identifying the heart abnormality based on RHOOF

1) The RHOOF descriptor data are fetched.

2) ML algorithms like a decision tree, discriminant analysis, logistic regression, SVM, KNN, and Ensemble are applied.

3) The TP, TN, FP, and FN values are obtained.

4) The statistical parameters like accuracy, precision, sensitivity, and specificity are derived using the algorithms.

5) The performance of the above ML algorithms is evaluated based on the statistical parameters.

6) The best performing algorithm is selected, and the model is trained.

7) The sample is tested.

8) The abnormality is identified.

ASD is seen in the upper two chambers of the heart in the form of hole on the septal wall separating left atrium (LA) and right atrium (RA), as such septal wall is our focused area or ROI in this study. During run time image cropping is done to locate the septal wall, and the algorithm automatically marks the ROI in each frame. We focused more on the region where abnormalities can be detected correctly by avoiding the unwanted motion vectors from the rest of the portion of the heart, allowing better time complexity, also ASD identification or diagnosis becomes easy and accurate.

We used 2D echocardiography–apical four chamber view videos of 20 different patients, gathered from the hospital and the open source database, for experimentation.

Scatter Plot

a.	Decision tree
b.	Discriminant analysis
c.	Logistic regression
d.	SVM
e.	KNN
f.	Ensemble

Confusion Matrix

a.   Decision tree   b.   Discriminant analysis   c.   Logistic regression

d.   SVM   e.   KNN   f.   Ensemble

ROC and AUC

a.   Decision tree   b.   Discriminant analysis   c.   Logistic regression
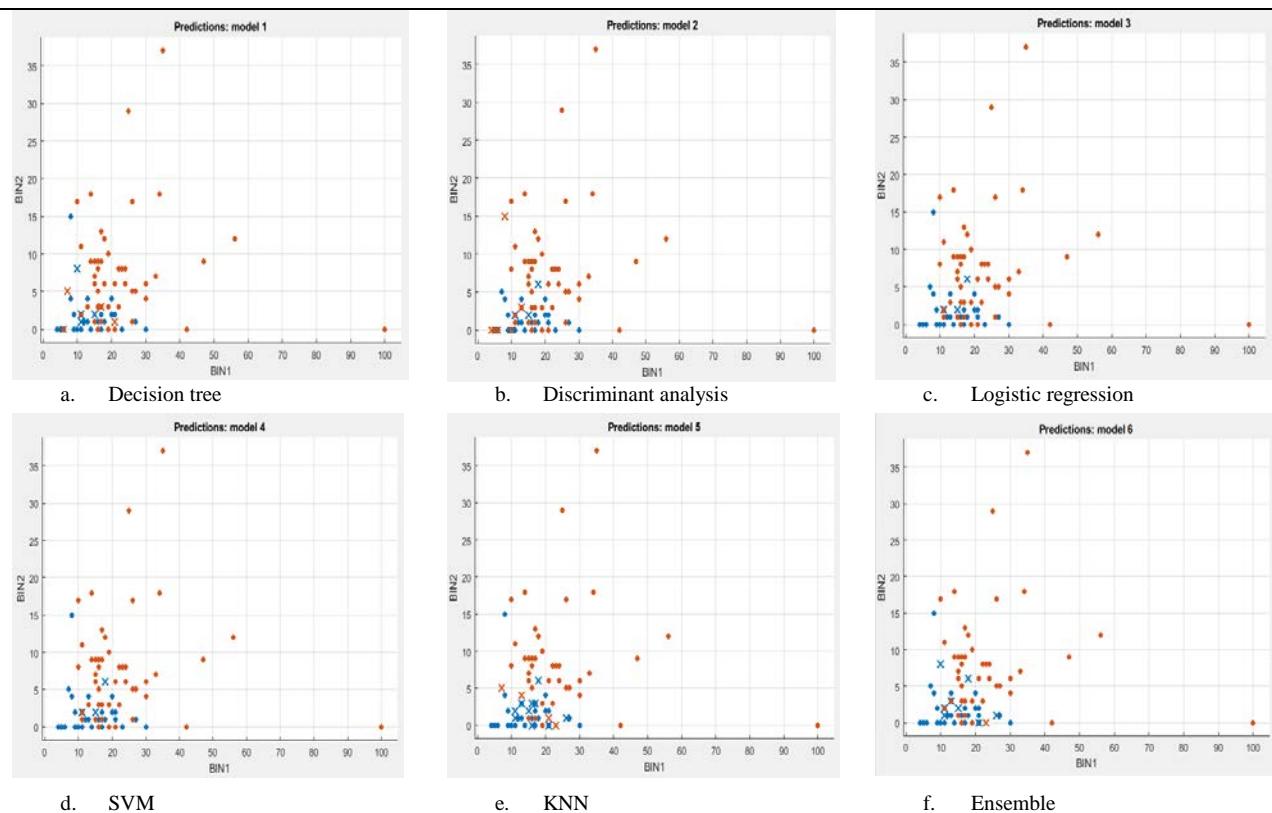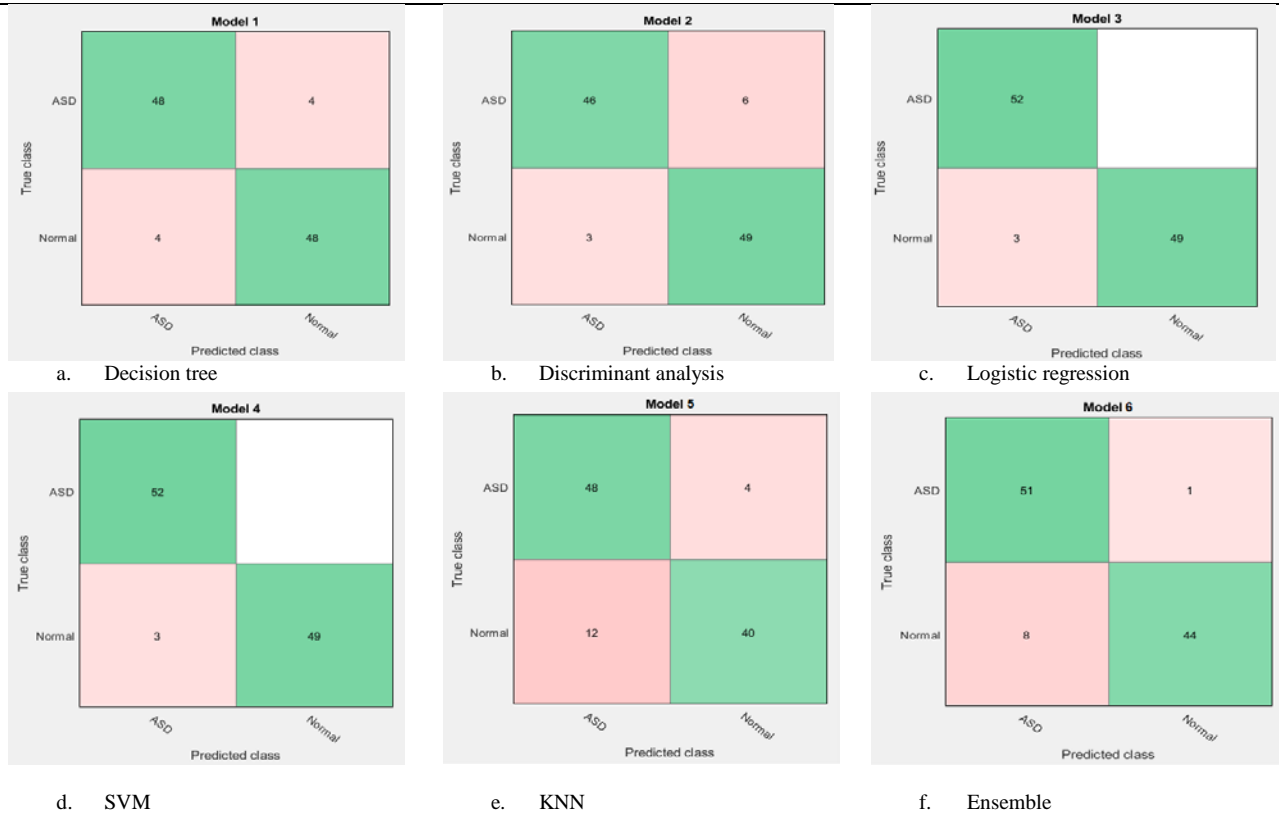
d.   SVM   e.   KNN   f.   Ensemble
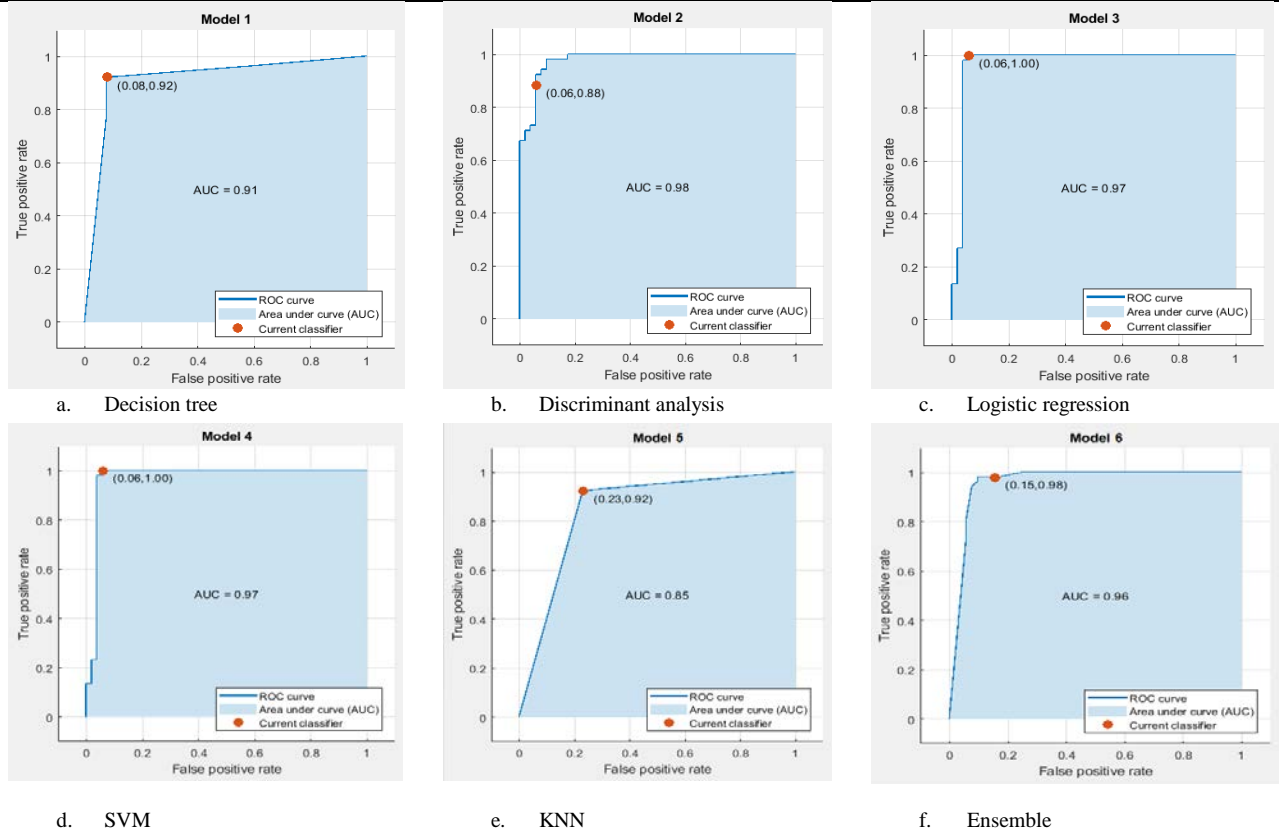
Fig. 6.   Machine Learning Algorithm(s) output. Scatter Plot of (a) Decision tree. (b) Discriminant analysis. (c) Logistic regression. (d) SVM. (e) KNN. (f) Ensemble. Confusion Matrix of (a) Decision tree. (b) Discriminant analysis. (c) Logistic regression. (d) SVM. (e) KNN. (f) Ensemble ROC and AUC of (a) Decision tree. (b) Discriminant analysis. (c) Logistic regression. (d) SVM. (e) KNN. (f) Ensemble.

TABLE I.   ANALYSIS OF RESULTS

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
Logistic regression	98.1	100	96.1	96.2	0.97
SVM	98.1	100	96.1	96.2	0.97
Discriminant analysis	95.2	96.1	94.2	94.3	0.98
Ensemble	95.2	96.1	94.2	94.3	0.96
Decision tree	91.3	94.2	88.4	89.0	0.91
KNN	89.4	94.2	84.6	85.9	0.85
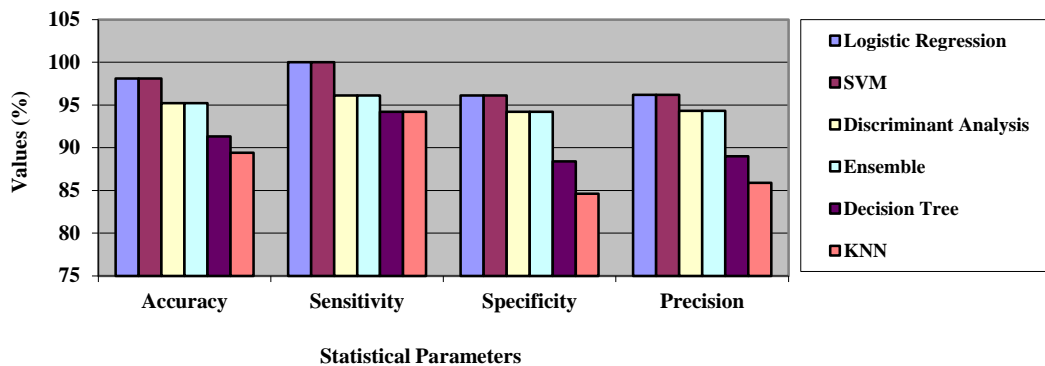
Fig. 7.   Graphical Representation of Results.

Using block matching technique EHLDSP proposed in our previous work [1], the movement pattern and direction of vectors can be determined. Based on the values of RHOOF, MV direction and magnitude is obtained. MVs with the same direction/angle were grouped together and were put into nine different bins based on the magnitude and angle of the direction to form a histogram. The movement of pixels from each frame of the video is used to train various classifiers and treat them as descriptors.

During the training stage, the extracted features from RHOOF were used to train the model and is saved, which is used for testing the sample data and determine the abnormality. The model was trained using 20 ultrasound videos (comprising 520 frames), 10 each of Normal and Abnormal.

"Fig. 6" showcases the performance of various ML classifiers, in the form of scatter plot, confusion matrix, the ROC and AUC. The scatter plot displays the value of two variables or data points and their co-relation. As a sample representation, in this manuscript we have included the scatter plots of comparison between bin1 and bin2 using six different ML models/classifiers.

As seen in the scatter plot, there are very few outliers and wrongly identified points which are marked as "x". Confusion matrix is drawn between the predicted class (x-axis) and true class (y-axis). In our case comparison was done using two parameters; ASD and Normal.

Based on the model/classifier performance, data gets divided into four main categories; TP, TN, FP, and FN. This information is further used for calculating accuracy, precision, sensitivity, and specificity parameters using the "(6), "(7)", "(8)" and "(9)", respectively.

In the medical field, algorithms ability to correctly identify the disease (true positive rate) is known as sensitivity, while the ability to correctly identify the non-disease (true negative rate) is known as specificity. As a result, more importance is given to sensitivity over specificity in the medical field; in our study, the LR and SVM has the highest sensitivity (100%). Sensitivity and specificity have an inverse relation; an increase in the value of either parameter will decrease the value of the other and vice versa.

In the confusion matrix, values in the green boxes represent TP and TN, respectively, and those are in the red/white boxes represent FP and FN, respectively. The receiver operating characteristics (ROC) is a curve of probability, while area under curve (AUC) indicates the degree of separability. AUC and ROC are used in ML, particularly during classification to visualize the performance of algorithms.

AUC can have a value between 0 and 1; when the AUC is 0.7, this indicates that 70% of the time model will be able to accurately identify positive and negative classes. An AUC value is considered as acceptable if it is between 0.7 and 0.8, excellent in case it is between 0.8 and 0.9, and outstanding if more than 0.9 [28].

For drawing the ROC, false positive rate (FPR) and true positive rate (TPR) values are plotted on the x-axis and y-axis, respectively. The blue line in the graph indicates the ROC curve and shaded portion or the area under the blue line is called as AUC. The red point on each classifier graph, indicates the performance of that classifier. The AUC values are obtained based on the performance of various classifiers and are given in Table I. "Fig. 7" is the graphical representation of accuracy, sensitivity, specificity, precision values from Table I.

In our study, SVM algorithms had an AUC of 0.97, which was well above 0.9 and thus categorized as outstanding. Although the AUC of a discriminant analysis is slightly higher (0.01) than that of SVM, this will not have any major impact on our overall classifier selection.

The accuracy obtained by another researcher [9] for histogram plus SVM was 85% and that for histogram and BPNN was 87.5%, for ultrasound video. In this research, the RHOOF and SVM classifier had an accuracy of 98.1%.

The computer with INTEL Core i7 processor and 16 GB memory was used for experimentation on MATLAB R2017b.

V. Conclusion

In this study, experimentation was carried out using RHOOF for extraction of features and six different ML techniques to choose the best performing model, which can be further used for abnormality identification. Before applying RHOOF, de-noising was carried out to remove the speckle noise present in an ultrasound video. As mentioned in Table I, the performance of all six techniques was analyzed based on the accuracy, sensitivity, specificity, precision, and AUC to choose the best performing model for abnormality identification. More importance was given to the sensitivity over specificity in the medical field. The following models were evaluated on accuracy and sensitivity: decision tree (91.3%, 94.2%), discriminant analysis (95.2%, 96.1%), logistic regression (98.1%, 100%), SVM (98.1%, 100%), ensemble (95.2%, 96.1%), and KNN (89.4%, 94.2%). From Table I, we can conclude that logistic regression (LR) and SVM had outperformed the other ML algorithms with an accuracy of 98.1% and sensitivity of 100%, respectively. However, during experimentation it is observed that SVM takes 1/3rd of time than LR for training, as a result we have chosen SVM for abnormality identification. Also, LR has a tendency of considering more outliers, which might lead to in-correct diagnostics. As this study is performed to evaluate the technique used in diagnosing cardiovascular problems, more emphasis is given to the correct identification of the disease. Using combination of proposed RHOOF algorithm with SVM we could correctly identify the ASD.

VI. Future Work

In other research fields, open databases are easily accessible; today, as a researcher in the biomedical engineering field, we faced a lot of challenges in obtaining medical images, and videos of required diseases, due to ethical and legal constraints. In the future, we aim to create a publicly/freely available database for researchers, with the help of practitioners and medical institutions.

VII. Declaration of Conflicting Interests

The author(s) declare(s) that there is no conflict of interest.

VIII. Funding

Acknowledgments

References

[1] Mrunal N. Annadate, Manoj S. Nagmode, "Tracking of Blood Movement in the heart using Elongated Horizontal Large Diamond Search Pattern and Optical Flow technique for enabling the detection of Atrial Septal Defect", IJEAT, Elsevier Scopus, Volume-8, Issue-6, Page 1862 to 1870, Aug 2019.

[2] Mrunal N. Annadate, Manoj S. Nagmode, "Automatic detection of Hypertrophic Cardiomyopathy using Segmentation Techniques in Heart Ultrasound Video", International Journal of Pure and Applied Mathematics, Volume 119 No., 12945-12954, 12 2018.

[3] Achmad Solichin , Agus Harjoko, Agfianto Eko, "Movement Direction Estimation on Video using Optical Flow Analysis on Multiple Frames", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018.

[4] Mrunal N. Annadate, Manoj S. Nagmode, "Efficient Speckle Reduction Techniques in Ultrasound Video Quality Enhancement– A Comparative Evaluation", IEEE, 978-1-5386-3243-7/17/2017.

[5] Rashmi, Ashwin Kumar, Nagaraja N S, "Abnormality Detection in Video Stream using Histogram of Optical Flow Orientation SVM Classifier", International Journal of Advanced Technology and Innovative Research (IJATIR), ISSN 2348–2370 Vol.09, Issue.07, Pages:1232-1238 , June-2017.

[6] Rensso Victor Hugo Mora Colque, Carlos Caetano, Matheus Toledo Lustosa de Andrade and William Robson Schwartz, "Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 27, NO. 3, MARCH 2017.

[7] Achmad Solichin, Agus Harjoko, Agfianto Eko Putra, "Grid-based Histogram of Oriented Optical Flow for Analysing Movements on Video Data", IEEE, 978-1-4673-8430-8/15/2015.

[8] Chetana D Patil, Bharathi V K, "Event Detection in Video Using Saliency Value and Histogram of Optical Flow", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE), DOI:10.15662/IJAREEIE.0505135 4201, 2015.

[9] G.N.Balaji, T.S.Subashinib, N.Chidambaramc, "Automatic classification of Cardiac Views in Echocardiogram using Histogram and Statistical Features", International Conference on Information and Communication Technologies (ICICT) 2014.

[10] J.R.R. Uijlings, I.C. Duta, N. Rostamzadeh, N. Sebe, "Real-time Video Classification using Dense HOF/HOG", ICMR, Glasgow, United Kingdom, '14 April 01 - 04 2014.

[11] Daniel Tenbrinck, Sönke Schmid, Xiaoyi Jiang, Klaus Schäfers and Jörg Stypmann, "Histogram-Based Optical Flow for Motion Estimation in Ultrasound Imaging", J Math Imaging 47:138–150, DOI 10.1007/s10851-012-0398, Springer Science+Business Media New York 2012.

[12] Tian Wang Hichem Snoussi, "Histograms of Optical Flow Orientation for Abnormal Events Detection", IEEE, 978-1-4673-56501-03/123/2012.

[13] Alessandro Becciu, Hans van Assen, Luc Florack, Sebastian Kozerke,

Vivian Roode, and Bart M., "A Multi-scale Feature Based Optic Flow Method for 3D Cardiac Motion Estimation", SSVM, LNCS 5567, pp. 588–599, 2009, Springer-Verlag Berlin Heidelberg 2009.

[14] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager and Ren´e Vidal, "Histograms of Oriented Optical Flow and Binet-Cauchy Kernels on Nonlinear Dynamical Systems for the Recognition of Human Actions" IEEE, 978-1-4244-3991-1/09/2009.

[15] Balza Achmad, Mohd. Marzuki Mustafa, and Aini Hussain, "Inter-frame Enhancement of Ultrasound Images Using Optical Flow" IVIC, LNCS 5857, pp. 191–201, Springer-Verlag Berlin Heidelberg 2009.

[16] F Wang, T Syeda-Mahmood, D Beymer, "Spatio-Temporal Motion Estimation for Disease Discrimination in Cardiac Echo Videos", IEEE, DOI: 10.1109/CIC.2008.4748992, October 2008.

[17] Aditi Roy, Shamik Sural, Jayanta Mukherjee, Arun K. Majumdar, "State-Based Modelling and Object Extraction from Echocardiogram Video", IEEE Transactions on Information Technology in Biomedicine, Vol. 12, No. 3, MAY 2008.

[18] Schlomo v. Aschkenasy, Christian Jansen, Remo Osterwalder, André Linka,Michael Unser, Stephan Marsch, and Patrick Hunziker, "Unsupervised image classification of medical ultrasound data by multiresolution elastic registration" Ultrasound in Med. & Biol., Vol. 32, No. 7, pp. 1047–1054, 2006.

[19] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection", Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, (CVPR'05) 1063-6919/2005.

[20] John L. Barron, David J. Fleet, Steven S. Beauchemin, "Performance of Optical Flow Techniques", International Journal of Computer Vision · February 1994.

[21] Berthold KP Horn and Brian G Schunck, "Determining Optical Flow", Artificial Intelligence, 185-203, 0004-3702, North Holland17(1981).

[22] Bruce D. Lucas, Takeo Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision", Proceedings of Imaging Understanding Workshop, pp. 121-130 (1981).

[23] Navneet Dalal, Bill Triggs and Cordelia Schmid, "Human Detection using Oriented Histograms of Flow and Appearance".

[24] J. H. Park, S. K. Zhou1, C. Simopoulos, J. Otsuki, and D. Comaniciu, "Automatic Cardiac View Classification of Echocardiogram".

[25] SS Beauchemin and JL Barron, "The Computation of Optical Flow".

[26] Matthew Eric Otey, Jinbo Bi, Sriram Krishnan, Bharat Rao, Jonathan Stoeckel, Alan Katz, Jing Han, and Srinivasan Parthasarathy, "Automatic View Recognition for Cardiac Ultrasound Images".

[27] Rensso Victor Hugo Mora Colque, Carlos Antˆonio Caetano J´unior and William Robson Schwartz, "Histograms of Optical Flow Orientation and Magnitude to Detect Anomalous Events in Videos".

[28] Jayawant N. Mandrekar, "Receiver Operating Characteristic Curve in Diagnostic Test Assessment Journal of Thoracic Oncology", Volume 5, Number 9, September 2010.

Impact of Circular Field in Underwater Wireless Sensor Networks

Syed Agha Hassnain Mohsan[1]*, Mushtaq Ali Khan[2], Arfan Mahmood[3], Muhammad Hammad Akhtar[4]
Hussain Amjad[5], Asad Islam[6], Alireza Mazinani[7], Syed Muhammad Tayyab Shah[8]

Department of Electrical Engineering, COMSATS University Islamabad (CUI), Pakistan[1,8]
Ocean College, Zhejiang University, Zhoushan, China[2,5]
Complex Networks and Control Lab, bhanghai Jiao Tong University, Shanghai, China[3]
National University of Science and Technology (NUST), Islamabad, Pakistan[4]
Beihang University (BUAA), Beijing China[6,7]

Abstract—**Underwater Wireless Sensor Networks (UWSNs) face challenges regarding high propagation delay, limited bandwidth, 3D topology and excessive energy consumptions. In this paper, a routing scheme with circular field is proposed for an efficient collection of data packets by using two mobile sinks in UWSNs. Results of proposed scheme are compared with previous implemented schemes which are used to measure the usage of mobile sink in the collection of data packets. In this study, we have compared the proposed scheme with current state-of-the-art routing protocols. The statistical significance of this work was analyzed in MATLAB. Marked observations to emerge from obtained results include an improvement in lifetime, increased throughput, increment in alive nodes and balanced energy consumption. In our view, these results strengthen the validity of proposed circular field. A significant increase in received packets is observed because maximum nodes are alive till 1500 rounds which provide maximum communication and less chance of creating void holes.**

Keywords—*Component; underwater wireless sensor networks; mobile sink; routing scheme; circlular field*

I. Introduction

A considerable research has been carried out on terrestrial sensor networks in different aspects but currently research community is attracted towards a new era of research in Underwater Wireless Sensor Networks (UWSNs). Human beings are unable to work at high pressure ocean environment for a long time. On the other hand, Terrestrial Wireless Sensor Networks (TWSNs) are feasible but it cannot be considerable for UWSNs because signals transmission is elusive due to higher attenuation of aquatic environment. Underwater Wireless Sensor Network (UWSN) consists of various components such as sensors and vehicles deployed in acoustic area for data collection and collaborative monitoring tasks. Generally, UWSNs differ from TWSNs in terms of sensors cost, dense deployment of these sensors, communication method and maximum power required for this communication. USWNs are being used in several applications of underwater environment such as observation of chemical pollution, seismic activities, submarine movements, monitoring of marine creatures, distributed tactical surveillance, disaster prevention and military applications. However, UWSNs face challenges regarding narrow bandwidth, 3D topology, long propagation delay, media access control, attenuation, harsh geographical

atmosphere, power constraints, losses of connectivity and high bit error rates. Though, research community has provided several techniques to cope with these challenges but still gap lies for research because of variable characteristics of ocean. Usually underwater communication is different from air due to varying conductivity, permittivity and permeability [1-2] parameters.

This technology has already achieved a great stature among researchers. With the progress in underwater technologies, a growing research inclination has been noticed by industry as well as academia in UWSNs [3-4]. Currently, acoustic signals are in great consideration for underwater applications as they are more feasible for underwater communications with low absorption rate [5]. Long transmission range can be achieved by acoustic communication. However, it faces time-varying signal distortion and transmission losses. In UWSNs, reliability and stability of network has paramount importance. The maintenance of network lifetime in UWSNs is big issue as it's difficult to replace those batteries which provide power to sensor nodes. Recently, researchers are working in underwater wireless power transfer to tackle these challenges. In addition, TWSNs routing protocols cannot be implemented to UWSNs because of limited bandwidth, long propagation delay, energy constraints, high mobility and non-uniform 3D topology. Researchers have suggested several routing protocols for UWSNs. Some routing protocols can enhance the stability duration of network on the basis of throughput while others increase throughput on a compromise over packet delay [5]. In these protocols clustering is considered as the pragmatic method for load balancing in UWSNs because sensor nodes are categorized in different groups. Among these sensor nodes, cluster head is used as leading head [6]. Some researchers proved that mobile sink is an effective way to enhance throughput [7]. In order to collect data, these mobile sinks can connect to individual sensor node or full cluster. In a research study discussed in [8], authors utilized two mobile sinks to gather data from sparse regions. It is noticed when both sinks are located at adjacent positions [9] then throughput is affected due to same sensor nodes.

Considering above discussion, we were inspired and implemented a routing protocol which uses two mobile sinks without covering same transmission area. The overriding objective of this study is to seek an intelligent use of mobile

*Corresponding Author

sinks. A previous routing scheme Towards Efficient Sink Mobility (TESM) [10] is implemented with a circular field while previously it was taken as rectangular field. A. Yahya *et al.* discussed that high MUR represents the effective use of mobile sinks and obtained MUR results of this work are compared with SEEC and TESM. The paper is organized as: Section 2 is based on relevant research discussion. Section 3 will focus on motivation behind this present study. The impacts of circular network field are evaluated in Section 4 while results discussion has been carried out in Section 5. Conclusion and future work is presented in last sections.

II. RELEVANT RESEARCH WORK

Since many years' researchers have been focusing to design several UWSNs protocols for qualitative and effective research analysis. Research community is working with the aim of paving a way for new era of underwater actuation and monitoring applications. UWSNs will help us to know ocean with its envisioned landscape of applications. Researchers have suggested various routing schemes to achieve better results as Goyal *et al.* [11] highlighted capabilities and limitations on the basis of a survey on UWSN with data aggregation. Khan *et al.* [7] designed a routing scheme named as weighting depth and forwarding area division depth base routing protocol (WDFAD-DBR). They achieved reliable results in sparse deployment by minimizing void holes issue. They proposed balanced energy consumption on the basis of division criteria of forwarding regions. A SEEC protocol is proposed in [9]. F. Pan *et al.* [12] presented an energy-balanced algorithm based on energy level partition and depth threshold. In this protocol, mobile sinks can collect data at sparse region and they implied clustering on dense regions. Wan *et al.* [13] suggested an energy efficient scheme based on multilevel cluster selection criteria which aims to choose cluster head according to high residual energy and less radius from sink. In [14], network lifetime of UWSNs is maximized by the assistance of sink mobility with geographic routing protocol. A reliable energy efficient routing protocol based on clustering is simulated [15] for UWSNs. Authors proposed that clustering technique is useful to reduce energy resources consumption for a sensor network. Researchers have proposed a routing protocol AEDG which involves effective data collection through AUV [16]. The usage of multiple mobile sinks to enhance network lifetime is discussed in previous studies. Authors validated that a better network lifetime can be achieved through assigning a predefined path to mobile sinks for movement. While research study in [17] focuses on relation between ocean currents and energy hole problems. This study introduces drift speed calculations. It mentions that drift speed can be calculated at any time if the current location of sensor nodes is known. Mobile sink changes its speed after sensing the ocean current accordingly. It is worth noting that throughput of this proposed scheme is also better. Many researchers have presented current research challenges of intrinsic properties, topology control and potentials in UWSNs through several topology control algorithms. M. Ma *et al.* [18] have used a hybrid approach to maximize network lifetime and minimize the energy holes. In [19], authors proposed a mechanism for data collection involving large scale multihop sensor networks. In a recent work, authors designed data uploading decision making on the basis of similar data collected [20]. In this research study, authors maximized the network lifetime in AEDG as AUV is sent to gather data from gateway nodes. Wang *et al.* deployed a selection algorithm based on Heuristic Adaptive Sink Sensor for prolonging network lifetime and energy saving purposes in this study. In another research study discussed in [11], Hop ID is given to specific sensor nodes according to hope count from sink. Habib M. Ammari [20] investigated sink mobility in a circular field. The field consists of concentric circular bands to find a possible solution for optimal sink mobility. In a Doctoral research study, Erik F. Golen [21] briefly discussed sensing and communication coverage of sensors in circular regions. In a recent research work, Chaima Zidi [22] illustrated energy efficient UASNs in which sensors were deployed in a circular sensor field. In a systematic review [23], modulation techniques and channel coding is briefly discussed for discrete underwater sensor nodes.

III. MOTIVATION

As discussed in [10], authors introduced a novel routing scheme TESM for efficient sink utility in which two mobile sinks are involved to gather required data. Authors introduced a novel metric Mobile Sink Utility Ratio (MUR) which is useful to measure the utility of mobile sinks while data gathering. They have used a rectangular field in their proposed scheme while current study focuses on circular field. Results have been compared with routing scheme for efficient sink utility and other previously implemented routing schemes with regard to network lifetime, throughput, residual energy, stability and instability period.

IV. SCHEME WITH CIRCULAR FIELD

An efficient use of mobile sinks in UWSNs appears as maximum throughput with minimum energy consumption. In this study, a routing scheme TESM discussed in [10] is considered with a different circular field. The achieved results will be discussed in next sections.

A. Network Model

This study includes a network circular field with a logical division of two equal parts. Circular field has a diameter of 100 meters. This logical division is adopted for an efficient movement of mobile sinks. It is worth noting, three sinks are used: one is static while other two are dynamic. Two mobile sinks are placed in each divided portion of circular field to gather data. The static sink is located above central position of circular network field. Network model can be seen in Fig. 1.

B. Energy Consumption Model

The transmission distance is represented as D and B represents the total size of packet. B_d and R_{disp} show bit duration and radio dissipation respectively. The total energy required [10] for sending B bits packets at transmission distance D can be calculated by following equation:

$$E_{trans}(B, D) = (B * R_{disp}) + (B * B_d) \qquad (1)$$

While the energy consumed during packet reception [11] can be calculated by using equation 2.

$$E_{rcv}(B, D) = B * R_{disp} \qquad (2)$$
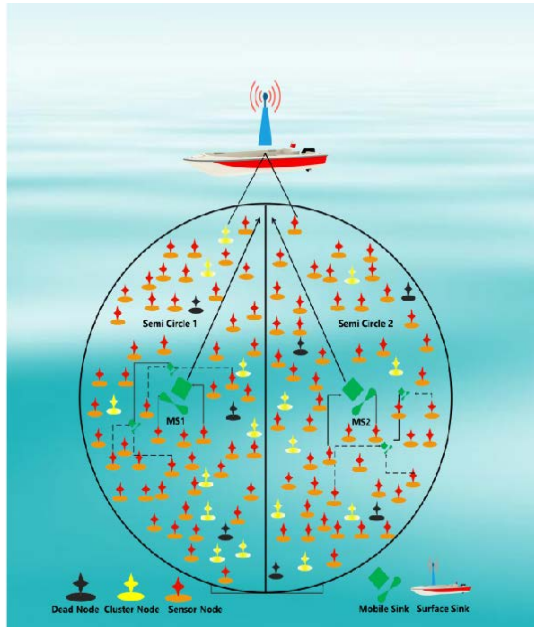
Fig. 1. Network Model.

The residual energy E_{ris} for each sensor node is obtained from equation 3.

$$E_{ris} = E_{int} - (E_{rcv} + E_{trans}) \qquad (3)$$

Here E_{int} shows the initial energy level of any sensor node.

C. Circular Network Field Configuration

Here, the procedure provided in literature [10] is followed to implement routing scheme with a circular network field. For this purposed work, the circular field is divided into two equal regions. Two mobile sinks MS1 and MS2 are deployed at center point of each region at same distance from center of circle. Nodes are randomly deployed in circular field. MS1 is present in semi-circle 1 while MS2 is present in semi-circle 2. Both sinks move at an angle of 45^0 directing upward towards base station deployed at terrestrial region. Corresponding sink receives data from nodes when sink appears within transmission range of these nodes. Transmission range and some others parameters are defined in next section. It is required to find sparse and dense areas after network deployment. Cluster is formed where nodes are dense. Random selection of cluster head is made according to highest energy. It is worth noting; in every round each member is provided opportunity to become cluster head. Cluster head can communicate to mobile sink and base station. Remaining nodes without cluster can communicate each other and can send packets to mobile sink as well. The coordinates of mobile sinks can be found from below equations.

$$X = r_i \cos \theta_i \qquad (4)$$

$$Y = r_i \sin \theta_i \qquad (5)$$

The circular network field is divided into two semi-circle regions where radius of each region is $r = 50$m. The origin of circle is represented as $O(x, y)$.

$$O(x, y) = (0, 0) \qquad (6)$$

While considering radius $r = 50$m, the area of two semi-circles can be calculated by following equation 7.

$$A_i = \frac{\pi r^2}{2} \qquad (7)$$

$i = 1, 2$ for two sub regions. As area calculation for both sub-regions is expressed in above equation, the total area of the whole circular network field is calculated. Equation 8 expresses the division criteria for each sector of circular network field.

$$S_n = \int_0^\pi \left(\frac{\theta}{360}\right) * \pi * r^2 \qquad (8)$$

Both mobile sinks as represented with MS1 and MS2. Mobile sinks move 45° in upward direction in a tilt way towards base station.

$$MS1 = \sum_{t=0}^n (\theta + 45^o/360) * \pi * r_i^2 \qquad (9)$$

Similarly

$$MS2 = \sum_{t=0}^n (\theta + 45^o/360) * \pi * r_i^2 \qquad (10)$$

Here n depicts total sectors while t=0 represents that MS1 and MS2 are at center position of first and second sub-region.

V. RESULTS DISCUSSION

To evaluate performance, the results are compared with previously implemented routing protocols SEEC, TESM and DBR. Simulation work was carried out by designing a complete UWSN environment through MATALB [11]. Table I summarizes simulation parameters.

To evaluate performance of routing protocols, proposed study considers primary metrics of network stability and instability period, packets received per round, network residual energy, network throughput, mobile sink utility ratio and packets received at sink.

A. Network Lifetime and Stability Period

By considering circular field, network lifetime is improved as it can be seen in Fig. 2. SEEC did not use two mobile sinks efficiently as their movement is in sparse regions only. Considering this transmission range, it is not suggested to use mobile sinks based on sparse area. Both MS1 and MS2 are restricted in each semi-circle. Both mobile sinks are moved upward in a tilt position towards base station. Due to this movement, a greater stability period performance is obtained as compared to SEEC, DBR and TESM. The restriction on movements of both mobile sinks in each semi-circle increases the stability period. It can be clearly seen that network lifetime results of proposed work are better comparatively with previous schemes. Fig. 2 validates more nodes are alive in this work than SEEC, DBR and TESM.

B. Network Residual Energy

In SEEC, clustering is achieved in top four dense regions while using only two mobile sinks in sparse regions. In DBR, energy consumption is more due to selection criteria because forward nodes are selected on lower depths only. In obtained results, both mobile sinks are restricted in each potion of circular field. If number of dense regions exceeds highly than sparse regions, mobile sinks movement will increase energy

consumption only. The distance between nodes is larger in sparse regions whereas it is smaller in case of dense regions. Residual energy of this proposed work compared with SEEC, TESM, DBR is plotted in Fig. 3.

TABLE I. SIMULATION PARAMETERS

Parameters	Values
Data rate	16 Kbps
Number of nodes	100
Packet size	50 Bytes
Center frequency	30 KHz
Initial energy of nodes	5 J
Receive power of packet	0.1 W
Transmission power of packet	2 W
Running rounds	3500
No. of Sinks	2
Radius of Circle	50 m
Transmission range	50 m
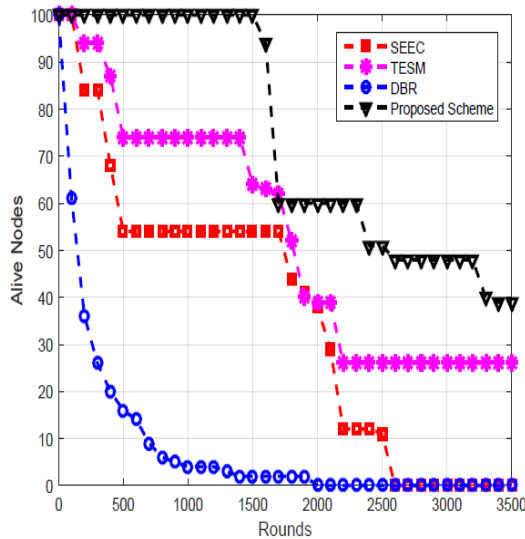
Fig. 2. Network Lifetime.
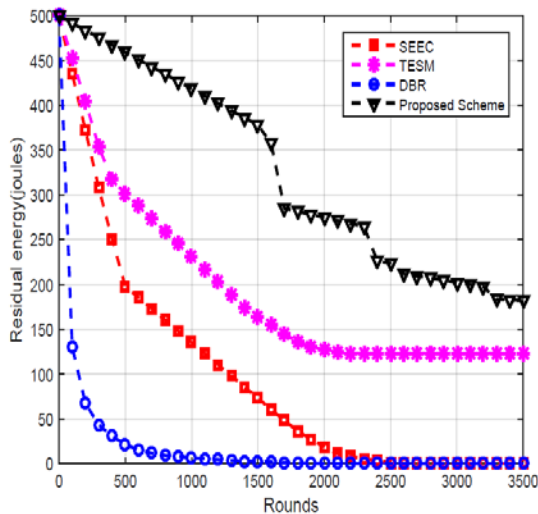
Fig. 3. Residual Energy.

C. Network Throughput

In Fig. 4, the received packets per round at any sink are more than SEEC, TESM and DBR. An increased throughput in this proposed work with circular field can be noticed as sensors are omni-directional as well. In SEEC, mobile sinks movement is according to sparse region and in TESM it is based on selection criteria opted for forwarding nodes.

Fig. 5 is showing the overall received packets at sink. Therefore, it gives the addition of received packets in both previous round and next round. Graph shows that more packets are received at sink than SEEC, TESM and DBR.

D. Mobile Sink Utility Ratio

Fig. 6 is showing the usage of mobile sink is high for data packets collection as compared to SEEC and lower than TESM.
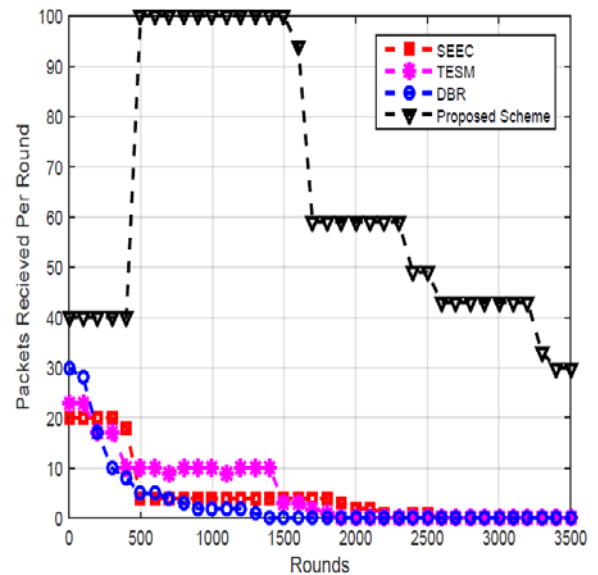
Fig. 4. Packets Received Per round.
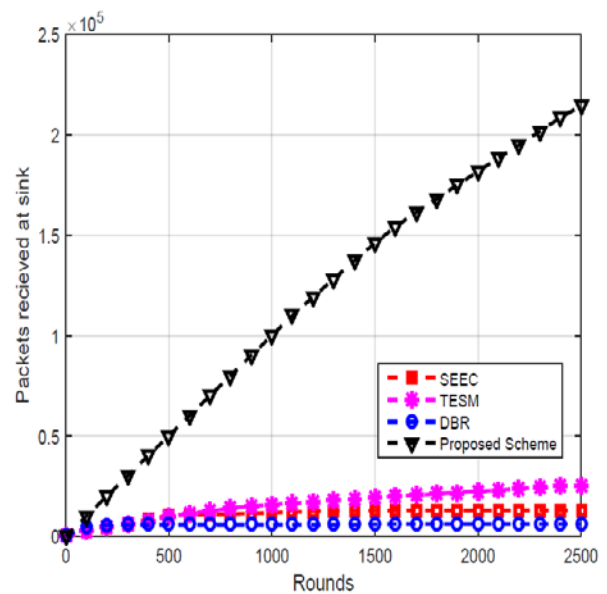
Fig. 5. Packets Received at Sink.
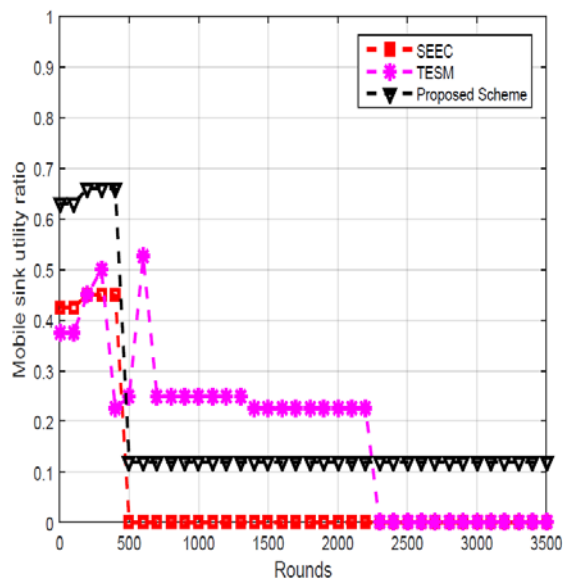
Fig. 6.   Packets Received at Sink.

Results of this scheme are not compared with other routing schemes as only SEEC and TESM use mobile sinks among all schemes. Therefore, proposed work is compared with SEEC and TESM to properly analyze the efficient utilization of mobile sinks. In these results, motion pattern of mobile sinks gives high MUR than SEEC but less than TESM because mobility pattern performance of TESM is better. In this case, high MUR results as high throughput while it is noticed that the overall packets collected in SEEC by mobile sinks are lower which results as low MUR.

E. Dead Nodes

Herein, stability period of implemented routing scheme TESM [8] with a circular field is shown. A good performance is noted in terms of stability period as compared to previous scheme of SEEC, DBR and TESM. In this work, network remains stabilize for a long time with a minimal consumption of energy. In proposed study, dead nodes are less than previous scheme as it is presented in Fig. 7.
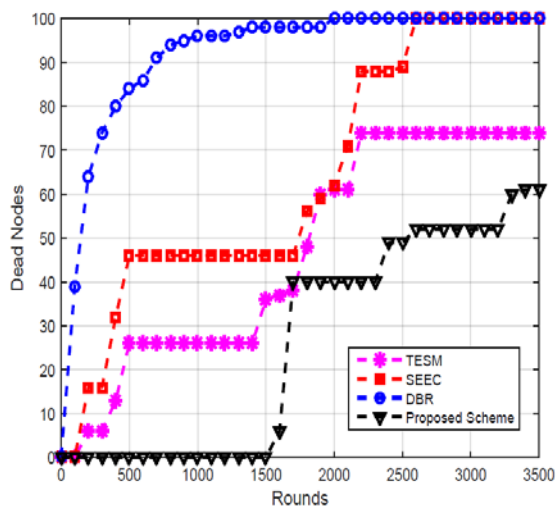
Fig. 7.   Stability and instability perdiod of network

VI. PERFORMANCE TRADEOFF

The proposed work gives better network lifetime, throughput and balanced energy consumption. It is worth noting that more nodes are alive while dead nodes are less. Thus, proposed work concludes a considerable progress of this work in aspects of stability period, residual energy and throughput.

VII. CONCLUSION

Results presented in this work illustrate that optimum and efficient use of mobile sinks is crucially important in enhancing network lifetime and throughput. This work has led us to conclude that a proper distribution of network field and shape is pivotal. Therefore, proposed work is compared with SEEC and TESM to properly analyze the efficient utilization of mobile sinks. In addition, the results of any routing scheme can be improved if proper use of resources in the protocol is maintained. In this study, the results while considering a circular field are compared with previous implemented schemes. To sum up our work, the proposed circular field is better in throughput, network lifetime and balanced energy consumption. A significant increase in received packets is observed because maximum nodes are alive till 1500 rounds which provide maximum communication and less chance of creating void holes. Thus, a good performance is noted in terms of stability period as compared to previous scheme of SEEC, DBR and TESM.

VIII. FUTURE WORK

In future, we are aiming to consider scarcity control in order to improve network lifetime. Though proposed work is energy efficient and throughput is also achieved but it will cause delay. Current work includes an investigation of possible techniques to reduce this delay. The delay in each round will be evaluated in future research work. Future work is to concentrate on minimizing interference in dense area which will make reliable delivery of data packets in any network. Holding time phenomenon towards reducing packet drop ratio is deferred to future work.

REFERENCES

[1]   Hassnain, S. A., M. J. Mughal, and Q. A. Naqvi. "Layered Chiral Spheres with Zero Backscattering." 2019 Photonics & Electromagnetics Research Symposium-Fall (PIERS-Fall). IEEE, 2019.

[2]   Hassnain, S. A., M. J. Mughal, and Q. A. Naqvi. "Analysis of Effective Medium Parameters on Polarizability of Homogeneous Chiral Sphere." 2019 Photonics & Electromagnetics Research Symposium-Fall (PIERS-Fall). IEEE, 2019.

[3]   Li, Ning, et al. "A probabilistic and highly efficient topology control algorithm for underwater cooperating AUV networks." Sensors 17.5 (2017): 1022.

[4]   Li, Ning, et al. "A survey on underwater acoustic sensor network routing protocols." Sensors 16.3 (2016): 414.

[5]   Yan, H.; Shi, Z.J.; Cui, J.H. DBR: Depth-based routing for underwater sensor networks. In International Conference on Research in Networking; Springer: Berlin/Heidelberg, Germany, 2008; pp. 72–86.

[6]   Han, G.; Jiang, J.; Bao, N.; Wan, L.; Guizani, M. Routing protocols for underwater wireless sensor networks. IEEE Commun. Mag. 2015, 53, 72–78.

[7]   Khan, Jawaad Ullah, and Ho-Shin Cho. "A distributed data-gathering protocol using AUV in underwater sensor networks." Sensors 15.8 (2015): 19331-19350.

[8] Coutinho, Rodolfo WL, et al. "Underwater wireless sensor networks: A new challenge for topology control–based systems." ACM Computing Surveys (CSUR) 51.1 (2018): 1-36.

[9] Sher, A.; Javaid, N.; Azam, I.; Ahmad, H.; Abdul, W.; Ghouzali, S.; Niaz, I.A.; Khan, F.A. Monitoring square and circular fields with sensors using energy-efficient cluster-based routing for underwater wireless sensor networks. Int. J. Distrib. Sens. Netw. 2017, 13, 1550147717717189.

[10] Yahya, Aqeb, et al. "Towards efficient sink mobility in underwater wireless sensor networks." Energies 11.6 (2018): 1471.

[11] Goyal, Nitin, Mayank Dave, and Anil K. Verma. "Data aggregation in underwater wireless sensor network: Recent approaches and issues." Journal of King Saud University-Computer and Information Sciences 31.3 (2019): 275-286.

[12] Feng, Pan, et al. "Improved energy-balanced algorithm for underwater wireless sensor network based on depth threshold and energy level partition." EURASIP Journal on Wireless Communications and Networking 2019.1 (2019): 1-15.

[13] Wan, Zhiping, et al. "An energy-efficient multi-level adaptive clustering routing algorithm for underwater wireless sensor networks." Cluster Computing 22.6 (2019): 14651-14660.

[14] Ahmed, Farwa, et al. "Mobile sinks assisted geographic and opportunistic routing based interference avoidance for underwater wireless sensor network." Sensors 18.4 (2018): 1062.

[15] Huang, Xiangdang, Shijie Sun, and Qiuling Yang. "Data Uploading Strategy for Underwater Wireless Sensor Networks." Sensors 19.23 (2019): 5265.

[16] Khan, Jawaad Ullah, and Ho-Shin Cho. "A distributed data-gathering protocol using AUV in underwater sensor networks." Sensors 15.8 (2015): 19331-19350.

[17] Chen, Y.S.; Lin, Y.W.Mobicast routing protocol for underwater sensor networks. IEEE Sens. J. 2013, 13, 737–749.

[18] Ma, Ming, and Yuanyuan Yang. "SenCar: An energy-efficient data gathering mechanism for large-scale multihop sensor networks." IEEE Transactions on Parallel and Distributed Systems 18.10 (2007): 1476-1488.

[19] Wang, Xingwang, et al. "Has4: A heuristic adaptive sink sensor set selection for underwater auv-aid data gathering algorithm." Sensors 18.12 (2018): 4110.

[20] Ammari, Habib M. "A unified framework for k-coverage and data collection in heterogeneous wireless sensor networks." Journal of Parallel and Distributed Computing 89 (2016): 37-49.

[21] Golen, Erik. "Intelligent deployment strategies for passive underwater sensor networks." (2009).

[22] Zidi, Chaima. Energy efficient underwater acoustic sensor networks. Diss. Université Sorbonne Paris Cité, 2018.

[23] Mehedi, Syed Agha Hassnain Mohsan Md, et al. "A Systematic Review on Practical Considerations, Recent Advances and Research Challenges in Underwater Optical Wireless Communication." IJACSA 11-7 (2020).

A Comparative Analysis of Data Mining Techniques on Breast Cancer Diagnosis Data using WEKA Toolbox

Majdah Alshammari[1], Mohammad Mezher[2]

Department of Computer Science

Fahad Bin Sultan University, Tabuk, KSA

Abstract—Breast cancer is considered the second most common cancer in women compared to all other cancers. It is fatal in less than half of all cases and is the main cause of mortality in women. It accounts for 16% of all cancer mortalities worldwide. Early diagnosis of breast cancer increases the chance of recovery. Data mining techniques can be utilized in the early diagnosis of breast cancer. In this paper, an academic experimental breast cancer dataset is used to perform a data mining practical experiment using the Waikato Environment for Knowledge Analysis (WEKA) tool. The WEKA Java application represents a rich resource for conducting performance metrics during the execution of experiments. Pre-processing and feature extraction are used to optimize the data. The classification process used in this study was summarized through thirteen experiments. Additionally, 10 experiments using various different classification algorithms were conducted. The introduced algorithms were: Naïve Bayes, Logistic Regression, Lazy IBK (Instance-Bases learning with parameter K), Lazy Kstar, Lazy Locally Weighted Learner, Rules ZeroR, Decision Stump, Decision Trees J48, Random Forest and Random Trees. The process of producing a predictive model was automated with the use of classification accuracy. Further, several experiments on classification of Wisconsin Diagnostic Breast Cancer and Wisconsin Breast Cancer, were conducted to compare the success rates of the different methods. Results conclude that Lazy IBK classifier k-NN can achieve 98% accuracy among other classifiers. The main advantages of the study were the compactness of using 13 different data mining models and 10 different performance measurements, and plotting figures of classifications errors.

Keywords—*Data mining; breast cancer; data mining techniques; classification; WEKA toolbox*

I. INTRODUCTION

Worldwide, breast cancer has become one of the most common cancers [1]. It originates in the area of the breast tissue that has a concentration of milk ducts. Although most cases occur in women, there have been reported cases in men as well. There are noticeable signs and symptoms of breast cancer. The first noticeable symptom is usually a different mass from the rest of the breast tissue. Most women, about 80%, discover these masses during self-examinations.

Breast cancer can be classified as benign or malignant; however, this classification is determined through diagnostic testing. Some criteria to consider are uniformity of cell size and shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, and normal nucleoli. By observing these criteria, doctors or scientists are able to make a diagnosis according to the patient's diagnostic test results.

Certain unhealthy lifestyle choices tend to put some people in danger of developing breast cancer: smoking, consuming fatty food and alcohol, lack of exercise, and stress. Although genetics plays a minor role in many breast cancer cases, unhealthy habits contribute more.

The rapid spread of breast cancer and the inability to accurately diagnose and recognize its presence represents a challenge for researchers and developers in biomedical engineering [2]. This challenge leads to deploying new data mining techniques. Data mining is the uprooting and recall of unknown data from the past that can be useful. Data mining also includes the acknowledge recovery and analysis of data that is saved in a data repository. Some of the important methods of data mining are classification, association, clustering and regression, etc. [3]. The focus of this paper is to drive the research towards new feasible solutions for mining breast cancer data. Thus, a data mining-based experiment for breast cancer classification mechanism is introduced with different types of classifiers. In addition to identifying the best classifier model that introduces higher classification accuracy for the predefined dataset used in this study, the data mining process is implemented by applying pre-processing operations and extracting features to the specified data records from the data set using WEKA.

The WEKA (Waikato Environment for Knowledge Analysis) is an open-source software that contains a set of algorithms for data mining tasks [4]. These algorithms can be applied to a data set either directly through the WEKA interface or via Java code. Then the different classifiers are implemented with different variables using several algorithms and multiple options to compute the best accuracy ratio.

In this study, the classification process was summarized through 13 experiments, including three experiments using the Bayes Net algorithm by three different search mechanisms and ten experiments using classification algorithms, Naïve Bayes (NB), Logistic Regression, Lazy IBK (Instance-Bases learning with parameter K), Lazy Kstar, Lazy Locally Weighted Learner (LWL), Rules ZeroR, Decision Stump, Decision Trees J48, Random Forest, and Random Trees to create a predictive

model that can be tested with new records and that can obtain classification accuracy, and compare the results obtained after implementing different algorithms compared to the slow algorithm IBK and k-NN. However, k-NN was the best in ranking for optimum time and accuracy values. The optimal classifier was determined to identify more new records for accurate breast cancer identification.

Moreover, this study aimed at utilizing data mining techniques to diagnose breast cancer using diabetic patients' datasets [5]. By looking at the literature, it is noticeable that there have been many efforts to use data mining for breast cancer datasets; however, previous studies lack in comparing WEKA with different parametric values and attributes. Experiments in this research used thirteen different data mining algorithms as well as the use of feature selection for data cleansing. This study shows competitive results compared to previous studies, mentioned in the literature.

The organization of this paper is as follows: Related works on breast cancer datasets using thirteen different classifiers are discussed in Section II, followed by Section III which provides an introduction about the classification algorithm used in the experiment. Section IV introduces the methodology used in this work; all the data mining techniques which are compared and analyzed are illustrated in Section V. Section VI concludes the proposed study and highlights the most accurate classifiers.

II. RELATED WORK

The comparisons made in [6], were based on the performance of four different machine learning algorithms: Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (C4.5) and k-Nearest Neighbors (k-NN) and were conducted on the Wisconsin Breast Cancer (WBC) datasets. The objective of the study and the experiments that were conducted were to determine the effectiveness of each algorithm in regards to precision, accuracy, specificity, and sensitivity. The results showed that SVM obtained 97.13% accuracy and outperformed Naïve Bayes, C4.5, and k-Nearest Neighbors algorithm (k-NN) that obtained accuracy variance between 95.12% and 95.28%.

In [7], Genetic Algorithm (GA) was used alongside different data mining techniques for WBC. GA was used to extract significant and informative features to reduce computational complexity and enhance the data mining processing speed. Data mining techniques used in this study were: Decision Trees (DT), Bayesian Network (BN), Logistic Regression (LR), Random Forest (RF), SVM, Rotation Forest, Radial Basis Function Networks (RBFN) and Multilayer Perceptron (MLP). Two WBC medical datasets (WBC and Wisconsin Diagnostic Breast Cancer (WDBC)) were used to test the performance of the algorithm models. The highest accuracy of 99.48% was obtained by Random Forest and GA feature selection.

The study conducted by [8] aimed at diagnosing breast cancer using three different techniques, namely: SVM, DT, and Artificial Neural Network (ANN). The study was applied on the WDBC dataset from the UCI. Feature selection was applied to increase the effectiveness of the methods. The ensemble method yielded the best results among the used methods. It gave 98.77% accuracy, 98.05% sensitivity and 100% specificity.

In [9], the authors used three well-known data mining methods, namely, Naïve Bayes (NB), J48, and RBF Network, to develop prediction models for the survivability of breast cancer. The data, that contains 683 instances, was acquired from the UCI [5]. To develop the prediction models, data selection, pre-processing, and transformation were applied. The results obtained from the experiment showed that the Naïve Bayes performed the best with a classification accuracy of 97.36%, RBF Network resulted in a classification accuracy of 96.77%, and the J48 resulted in classification accuracy of 93.41%.

The work by [10] used 12 different machine learning techniques for the diagnosis of breast cancer. The techniques that were used are namely; NB, Decision Table, Ada Boost M1, J48, J-Rip, Logistics Regression, Lazy IBK, Lazy K-star, Multiclass Classifier, Multilayer–Perceptron, RF, and RT. WBCD dataset was used to train the model. Most of the applied methods scored above 94%. Only NB underperformed, compared to the other models, with an accuracy of 73.21%. RT and Lazy classifier algorithms outperformed the others with an accuracy close to 99%.

In [11], evaluated six different data mining techniques, namely: SVM, Bayes Network (BN), ANN, k-NN, Decision Tree (C4.5) and Logistic Regression. The WEKA tool was used for the experiment on the WBC dataset. SVM and BN yielded the highest accuracy of 97.28%. However, the BN classifier produced minimal time compared to SVM, which makes the BN classifier better.

In [12], researchers employed eight different data mining techniques for breast cancer prediction. The dataset used for the experiment was WPBC [5]. The experiments were done on four classification algorithms: SVM, DT C5.0, NB and k-NN and on four clustering algorithms: EM, K means, PAM and Fuzzy c-means. The experiments were implemented using the R programming tool. The results showed that classification algorithms have better performance than the clustering where SVM and DT (C5.0) had the best accuracy of 81% and Fuzzy c-means resulted in the lowest accuracy of 37%, among the tested algorithms. The study conducted by [13] utilized three data mining techniques to classify breast cancer as either malignant or benign. The techniques conducted on the WDBC breast cancer dataset [5] are, namely: LR, NB and DT. Results showed that Logistic Regression (LR) got the highest classification accuracy of 97.90% among the other two tested classifiers.

The study conducted by [14] proposed nested ensemble methods to distinguish benign tumors from malignant breast cancers. Each ensemble method contains "Classifiers", as well as "Metaclassifiers" that can have more than two classification algorithms. Metaclassifiers were developed in the two-layer nested ensemble. The dataset used for the experiments was WBDC. The proposed method (used by [14]) was compared to the conventional single classifiers such as BN and NB. The results indicated that the two-layer nested ensemble method outperforms the single classifiers.

To analyze breast cancer data, [15] utilized four different DT classification algorithms, namely, Classification and Regression Trees (CART), J48, Best First Tree (BF Tree) and DT (AD Tree). The experiment employed the WEKA tool, and the results demonstrated that the J48 classifier reached the highest accuracy of 99% whereas the CART algorithms resulted in 96% accuracy; AD Tree algorithm resulted in 97%, and BF Tree algorithm resulted in 98%.

For the experiment in this research study, k-NN achieved the highest accuracy of 98% whereas in [7,4,1], RF achieved the highest accuracy. The highest accuracy in [8] was achieved by the ensemble methods of SVM, DT, and ANN. By analyzing the literature, it is noticeable that different techniques got the highest accuracy in each study as follows: in the study conducted by [11], the highest accuracy was achieved by Bayes Network (BN), and SVM and DT (C5.0) got the best accuracy in [12]. Similarly, by looking at [13], it is noticeable that the highest accuracy was yielded by LR and in [15] the highest accuracy was achieved by the J48 classifier.

By analyzing the literature, it was also noticed that the proposed research yields competitive results in terms of accuracy. However, not all mentioned previous works that used WEKA tools for data mining, the data mining using WEKA tools, achieved the same task. For example, [10], [13] and [14] used data mining methods to classify cases of breast cancer into malignant and benign. Moreover, the other studies did experiments on a few techniques while this study tested thirteen different algorithms.

Table I shows all the previous studies where a different methodology for utilizing data mining techniques to diagnose was used.

TABLE I.   SHOWS COMPARISONS BETWEEN PREVIOUS STUDIES

Ref.	Model	Dataset	Highest performance	Accuracy result
[6]	NB, SVM, C4.5, k-NN	WBC	SVM	97.13%
[7]	GA, DT, BN, LR, RF, SVM, RF, RBFN, MLP	WBC, WDBC	RF	99.48%
[8]	SVM, DT, ANN	WDBC	SVM, DT, ANN	98.05 %
[9]	NB, J48, RBF Network	WDBC	NB	97.36%
[10]	NB, Decision Table, Ada Boost M1, J48, J-Rip (LR), Lazy IBK, Lazy K-star, NN, RF, RT	WDBC	Lazy IBK, Lazy K-star, RF, RT	99.14%
[11]	SVM, BN, ANN, k-NN, DT(C4.5) and LR	WBC	SVM, BN	97.28%
[12]	SVM, Decision Trees C5.0, NB, k-NN, EM, K means, PAM, Fuzzy c-means	WPBC	SVM, DT (C5.0)	81%
[13]	LR, NB, DT	WPBC	LR	97.90%
[14]	**BayesNet, NB**	**WDBC**	**BN**	**98.07%**
[15]	CART, J48, BF Tree, AD Tree	Congressional Voting Records Data Set	J48 classifier	99%

III. CLASSIFICATION ALGORITHM

The experiment for this study ran the k-NN algorithm on the dataset. However, this algorithm is known as IBK in WEKA toolbox. The IBK classification algorithm, for each test instance, measured the distance to identify the nearest k instances to that instance from the training data. Then all selected instances were used for explication of prediction results. Such mechanism is referred to as k-NN algorithm. Fig. 1 shows the pseudo code of the k-NN algorithm that is described as follows [16,17]:

Consider (X_i, C_i) where $i = 1, 2…, n$ are the points of data. X_i is the feature values & C_i are the labels for X_i for each i. Considering 'c' is the number of classes then for all values of i, classes are represented as $C_i \in \{1, 2, 3, ……, c\}$.

Consider x is a point of an unknown label, and finding that unknown label class using k-NN algorithms is performed by:

Step 1: D (x, xi) = ECULIIAN (x, xi), i = 1,2, 3,…n
Step 2: Ds = SORT (D (x, xi)) in Ascending
Step 3: GET k elements from top of Ds, K is +int
Step 4: Get Indices of the k elements
Step 5: while k≥0, ki refers to k points belongs to ith class
Step 6: If ki >kj ∀ i ≠ j then put x in class i

Fig. 1.   Pseudo-Code of the IBK Algorithm.

In the IBK algorithm, a model was not manufactured but generated a just-in-time prediction for a test case. In each instance, the IBK algorithm reached a distance measure for locating k "near" cases in training data and used those instances to predict in order to determine which classifier in the WEKA toolbox, using the diabetic patient's dataset, had the highest accuracy. Experiments in this research analyzed the comparison of thirteen algorithms in various accuracy measurements.

Each algorithm conducted is represented briefly in terms of how it operates with the key parameters of the respective algorithm. The parameters of the algorithms conducted in this study are highlighted in Table III. The region size of k-NN is verified by the k-parameter. The distance metric used in k-NN is another important parameter. In the k-NN algorithm, the distance metric in default is Euclidean distance, which is ideal for quantitative data of the same size to determine the distance between instances.

The parameter C called the complexity parameter in WEKA governs the versatility with which the line can be drawn to separate groups. There is no margin violation with a value of 0, while the default is 1. The type of kernel to be used is a key parameter in SVM. The easiest kernel is a linear kernel that separates data from a direct line or hyperplane. The default kernels for the WEKA is the polynomial, where the higher the polynomial kernel, the wigglier of the exponent value. The classes are separated by means of a curved or angled line. In LR algorithm, the ridge parameter determines how much the algorithm needs to be forced to decrease the coefficient value. This regularization is deactivated by setting it to 0.

In order to implement the NB algorithm, the Kernel Estimator argument was used which might more accurately suit the actual distribution of attributes in the dataset. With DT, the researchers of this study chose to transform the no Pruning parameter to Real value. The minimum number of the tree instances in a leaf node when constructing the tree out of the training data was set to 7.

IV. METHODOLOGY

The methodology followed in this paper first started with data collection and data preparation from the BC dataset. Then data mining techniques were applied to generate a classification model that was used for evaluation and deployment. The diagrammatic representation of the proposed study framework is shown in Fig. 2. This methodology framework represents the standard data mining framework that should be followed in various applications of data mining, and the framework states that before data processing, selection and pre-processing should be performed to get clean and complete data records for processing. Then transformation of data represents a data extension conversion or compatibility conversion of data shape to an equivalent data mining tool format to perform data mining algorithm implementation and generate the desired knowledge outcome.
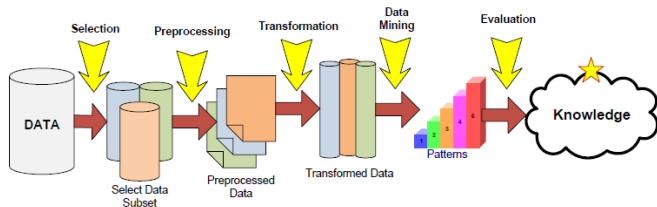
Fig. 2.   Study Framework.

Analysis of the breast cancer dataset using WEKA machine learning software tool aims for mining the relationship in breast cancer data for efficient classification. A dataset is an aggregation of data which refers to the contents of one statistical data matrix or one database table. This dataset is then processed using WEKA toolbox. In this sense, Table I provides an overview of the dataset used in the experiments. The dataset values for each of the variables, such as the height and weight of an object, are then listed and each member of the dataset is indicated by a datum [17-18]. The dataset may include data for single or multiple members, with respect to the number of rows [19-20].

For this study, diabetic patients' dataset was collected and consists of 699 patients records with 10 different attributes and nine nominal attributes of them were selected as shown in Table II. The data mining algorithms were explored to identify efficient classification of diabetes dataset [5]. Accuracy metric was used as the main comparison base while other metrics were also considered such as the Precision Recall (PRC), corresponding sensitivity (recall), the Receiver Operating Characteristics (ROC) and Matthews Correlation Coefficient (MCC).

TABLE II.   DATASET ATTRIBUTES DESCRIPTION

Description	Invalid records	Range	Mean Dev.	Standard Dev.
Clump thickness	45	10-1	4.442	2.821
Uniformity of cell size	35	10-1	3.151	3.065
Uniformity of cell shape	-	10-1	3.215	2.989
Marginal adhesion	70	10-1	2.83	2.865
Single epithelial cell size	-	10-1	3.234	2.223
Bare nuclei	30	10-1	3.545	3.644
Bland chromatin	-	10-1	3.445	2.45
Normal nucleoli	25	10-1	2.87	3.053
Mitoses	-	10-1	1.603	1.733

V. COMPARISON WITH OTHER ALGORITHMS

In this section, an experimental analysis of the effectiveness of the proposed methodologies for the thirteen different classifiers using the same dataset was completed. The experiments were implemented using WEKA toolbox. The experiments were done on a Toshiba desktop computer with Intel(R) Core (TM)i7-4710MQ CPU with @2.50GHz, 2.5GHz, and 8192MB RAM. The result of the evaluation represents the classification model results used to evaluate 25% of the selected dataset records, while the remaining 75% was used for training.

Table III shows the different algorithms used in the experiments' setup. The experimental methodology of the study, used 11 different classification algorithms and 3 other search methods.

TABLE III.   ALGORITHMS USED IN THE EXPERIMENTS' SETUP

No.	Algorithm Variant
1	BayesNet - Tabu Search
2	BayesNet - K2 Search
3	BayesNet - TAN Search
4	Naive Bayes
5	Logistic
6	Lazy – IBK
7	Lazy – Kstar
8	Lazy – LWL
9	Rules ZeroR
10	Trees DecisionStump
11	Trees J48
12	Trees RandomForest
13	Trees RandomTree

Fig. 3 represents the execution of the performance records registered by WEKA. The figure shows evaluation of classification for the PRC, ROC area and MCC results. Where Precision Recall (PR) is the precision values for corresponding sensitivity (recall) values and the ROC area refers to general performance of the classifier. The MCC is a measure of the quality of the binary classifications. However, the general combined model measures the precision and recall calculated in the F-Measure as in (1).

$$F = 2 * Precision * Recall / (Precision + Recall) \qquad (1)$$

So, alternative performance measurement was the confusion matrix specifically using the ROC curve shown in Fig. 3. However, the accuracy of the Algorithm 9 (Rules-ZeroR algorithm) was very weak; in contrast, PRC is more useful which shows how the classifier was behaving on one class.

The classifiers from strong and weak classifications (Fig. 3) were reviewed and analysed in order to investigate these results further (Table IV). Groups of classifiers that were very similar in their performances were found. Thereby, similar performance from the similar classifiers are highlighted in bold in Table IV.
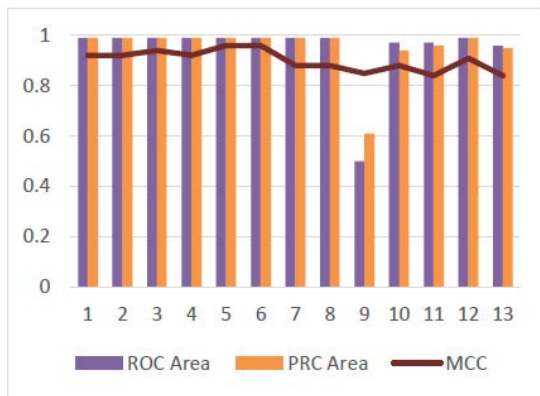
Fig. 3.   Performance Results for All Experiments.

TABLE IV.   EXPERIMENTS INDICES USED IN EVALUATION

No	F-Measure	Recall	Precision
1	0.97	0.97	0.97
2	0.97	0.97	0.97
3	0.98	0.98	0.98
4	0.97	0.97	0.97
5	**0.99**	**0.99**	**0.99**
6	**0.99**	**0.99**	**0.99**
7	0.95	0.95	0.96
8	0.95	0.95	0.96
9	0.85	0.73	0.73
10	0.95	0.95	0.96
11	0.94	0.94	0.94
12	0.97	0.97	0.97
13	0.94	0.94	0.94

In Fig. 4, Roles ZeroR had the weakest classifier which ultimately lead to a lower accuracy rate while other investigated algorithms produced more than 90% accuracy rate. Fig. 4 also shows that the best performed classifier in the three measurements was for both Algorithm 5 (logistic algorithm) and Algorithm 6 (lazy-IBK algorithm). In Fig. 4, the thirteen classifiers' results show both the weak (shown in orange color) and strong classifiers' (shown in blue color) performance. The classifiers were experimented for the dataset illustrated in Table II.

The statistical records for the execution of the experiments shown in Fig. 5, are the kappa statistics, absolute error, and the mean error. The failure classifier in the experiment was for Algorithm 5 (logistic algorithm). All statistics correspond to the classifiers' results where the type of errors is computed as a part of Kappa statistics plots.

Fig. 5 shows the final results of all experiments which are introduced in Table V. Table V shows the accuracy and elapsed time during the setup experiment time of each algorithm and its configurations. The worst performance was by the Rules ZeroR algorithm, and the best by the Lazy IBK and K-star algorithms while all other tests introduced more than 90% accuracy. The highest performance results are grouped and highlighted in bold.
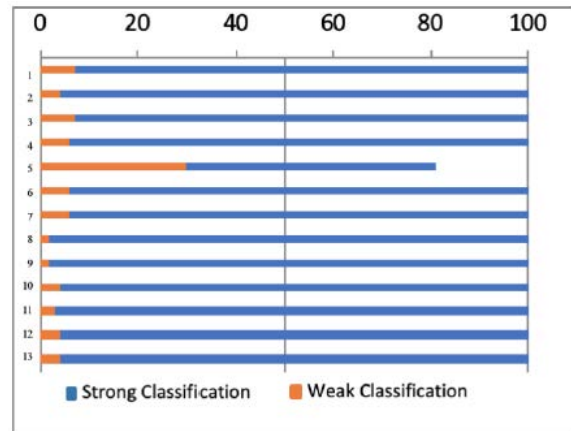
Fig. 4.   Strong and Weak Classifications.
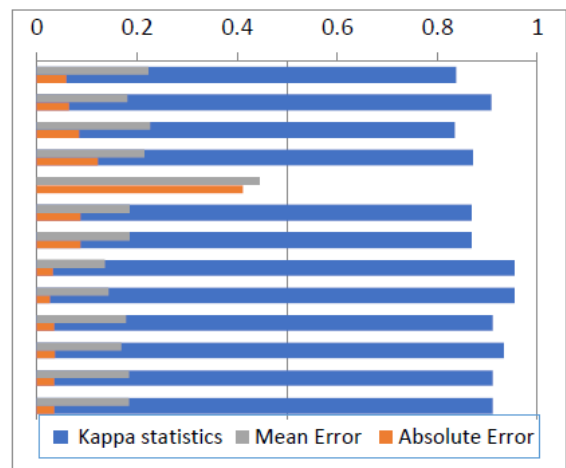
Fig. 5.   Error and Kappa Statistics.

TABLE V.   FINAL ACCURACY AND ELAPSED TIME RESULTS

No	Classifier	Accuracy	Elapsed Time
1	BayesNet- TabuSearch	96.4%	0.01
2	BayesNet- K2 search	96.4%	0
3	BayesNet TAN search	97.3%	0
4	Naive Bayes	96.4%	0
5	Logistic	92.8%	0.15
6	Lazy – IBK	**98.2%**	0
7	Lazy – Kstar	**98.2%**	0.92
8	Lazy – LWL	94.6%	0.17
9	Rules. ZeroR	73 %	0
10	Trees. DecisionStump	94.6%	0
11	Trees.J48	93.7%	0.02
12	Trees. RandomForest	96.4%	0.04
13	Trees. RandomTree	93.7%	0

VI. CONCLUSIONS

This paper studied a common practical problem in the detection or recognition of data patterns using data mining techniques. The comparative analysis proposed here for the Breast Cancer dataset using different pre-processing techniques was conducted using the WEKA data mining tool. The final evaluation of classification processes was done by extracting accuracy ratios for all experiments, and the results showed high rates ranging between 72% and 98%. The other ten experiments used the different classification algorithms to obtain the highest accuracy ratios; however, the IBK and K-star algorithms from Lazy algorithms showed the best performance up to 98.2% in optimum time and accuracy values. Those records can be further used in real-world applications such as any development models introduced for biomedical laboratory technology.

In the future, this study will be extended by utilizing deep learning techniques in order to get the highest accuracy. Moreover, the proposed technique will be applied on datasets for different cancer types.

REFERENCES

[1] Otoom, Ahmed Fawzi, Emad E. Abdallah, and Maen Hammad. Breast Cancer Classification: Comparative Performance Analysis of Image Shape-Based Features and Microarray Gene Expression Data. International Journal of Bio-Science & Bio-Technology, Vol.7, No.2 ISSN: 2233-7849, pp.37 – 46, 2015.

[2] A. Thomas, A. Rhoads, E. Pinkerton, M. C. Schroeder, K. M. Conway, W. G. Hundley, et al., Incidence and Survival Among Young Women with Stage I-III Breast Cancer: SEER 2000-2015. JNCI cancer spectrum, vol. 3, pp. pkz040-pkz040, 2019.

[3] Sayedeh Somayeh Hosaini and Mehran Emadi Breast Cancer Tumor Diagnosis from Mammography Images Using Wavelet Transform and Hidden Markov Model. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 4, Issue 8, ISSN: 2278 – 8875, pp. 6815-6823, 2015.

[4] Hajar Saoud, Abderrahim Ghadi, Mohamed Ghailani, Boudhir Abdelhakim.2018. Application of Data Mining Classification

[5] Algorithms for Breast Cancer Diagnosis: SCA '18: 3rd International Conference on Smart City Applications Tetouan Morocco October, 2018 ISBN:978-1-4503-6562-8.

[5] Breast Cancer Wisconsin Data Set UCI. 1992.

[6] Chaurasia, V., Pal, S. & Tiwari, B. B. 2018. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms and Computational Technology 12(2): 119–126. doi:10.1177/1748301818756225.

[7] Aličković, E. & Subasi, A. 2017. Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications 28(4): 753–763. doi:10.1007/s00521-015-2103-9.

[8] Zorluoglu, G. & Agaoglu, M. 2017. Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods 2 Related Works 2 Classification Methods The classification models of Clementine used in 2: 24–27.

[9] Chaurasia, V., Pal, S. & Tiwari, B. B. 2018. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms and Computational Technology 12(2): 119–126. doi:10.1177/1748301818756225.

[10] Kumar, V., Mishra, B. K., Mazzara, M., Thanh, D. N. H. & Verma, A. 2020. Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. Lecture Notes on Data Engineering and Communications Technologies, hlm. Vol. 37. Springer Singapore. doi:10.1007/978-981-15-0978-0_43.

[11] Saoud, H., Ghadi, A., Ghailani, M. & Boudhir, A. A. 2018. Application of data mining classification algorithms for breast cancer diagnosis. ACM International Conference Proceeding Series. doi:10.1145/3286606.3286861.

[12] Ojha, U. & Goel, S. 2017. A study on prediction of breast cancer recurrence using data mining techniques. Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data ScienceandEngineering527–530. doi:10.1109/CONFLUENCE.2017.7943207.

[13] Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree. International Journal Of Engineering And Computer Science 6(2): 20388–20391. doi:10.18535/ijecs/v6i2.40.

[14] Barua, P. D. & Gururajan, R. 2020. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters 132: 123–131. doi: 10.1016/j.patrec.2018.11.004.

[15] Rodríguez-Jiménez, J. M., Cordero, P., Enciso, M. & Mora, A. 2016. Data mining algorithms to compute mixed concepts with negative attributes: an application to breast cancer data analysis. Mathematical Methods in the Applied Sciences 39(16): 4829–4845. doi:10.1002/mma.3814.

[16] Ibrahim, Amal S., and Nabiel NH Mikhail. "The Evolution of Cancer Registration in Egypt: From proportions to population-based incidence rates." SECI Oncology 2015 DOI: 10.18056/seci, pp.1-21, 2015.

[17] Shraddha Soni. "A Literature Review on Data Mining and its Techniques "Indian journal of applied research, Volume 05- issue 06, ISSN: 2249-5555, pp.730-731, June. 2015.

[18] Ibrahim, Amal S., and Nabiel NH Mikhail. The Evolution of Cancer Registration in Egypt: From proportions to population-based incidence rates. SECI Oncology 2015 DOI: 10.18056/seci, pp.1-21, 2015.

[19] Fariba Mirbaha, Gloria Shalviri, Bahareh Yazdizadeh, Kheirollah Gholami and Reza Majdzadeh. Perceived barriers to reporting adverse drug events in hospitals: a qualitative study using theoretical domains framework approach. Implementation Science, 10: 110 DOI: 10.1186/s13012-015-0302-5, pp.1-10, 2015.

[20] Sayedeh Somayeh Hosaini and Mehran Emadi Breast Cancer Tumor Diagnosis from Mammography Images Using Wavelet Transform and Hidden Markov Model. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 4, Issue 8, ISSN: 2278 – 8875, pp. 6815-6823, 2015.

Performance Analysis of Efficient Pre-trained Networks based on Transfer Learning for Tomato Leaf Diseases Classification

Sawsan Morkos Gharghory

Computers and Systems Department
Electronics Research Institute
Giza, Egypt

Abstract—**Early diagnosis and accurate identification to tomato leaf diseases contribute on controlling the diffusion of infection and guarantee healthy to the plant which in role result in increasing the crop harvest. Nine common types of tomato leaf diseases have a great effect on the quality and quantity of tomato crop yield. The tradition approaches of features extraction and image classification cannot ensure a high accuracy rate for leaf diseases identification. This paper suggests an automatic detection approach for tomato leaf diseases based on the fine tuning and transfer learning to the pre-trained of deep Convolutional Neural Networks. Three pre-trained deep networks based on transfer learning: AlexNet, VGG-16 Net and SqueezeNet are suggested for their performances analysis in tomato leaf diseases classification. The proposed networks are carried out on two different dataset, one of them is a small dataset using only four different diseases while the other is a large dataset of leaves accompanied with symptoms of nine diseases and healthy leaves. The performance of the suggested networks is evaluated in terms of classification accuracy and the elapsed time during their training. The performance of the suggested networks using the small dataset are also compared with that of the-state-of-the-art technique in literature. The experimental results with the small dataset demonstrate that the accuracy of classification of the suggested networks outperform by 8.1% and 15% over the classification accuracy of the technique in literature. On other side when using the large dataset, the proposed pre-trained AlexNet achieves high classification accuracy by 97.4% and the consuming time during its training is lower than those of the other pre-trained networks. Generally, it can be concluded that AlexNet has outstanding performance for diagnosing the tomato leaf diseases in terms of accuracy and execution time compared to the other networks. On contrary, the performance of VGG-16 Net in metric of classification accuracy is the best yet the largest consuming time among other networks.**

Keywords—Deep learning; Alex; squeeze; VGG16 networks; tomato leaf diseases diagnosis and classification

I. INTRODUCTION

In the past decades, the plant diseases identification were mostly performed through the optical observation by farmer. Unfortunately, the process of detection and diagnosis of crop diseases by this method were error-prone, expensive and time consuming. In addition, there is no local experience for dealing with any new diseases maybe occur in places that were previously unidentified [1]. Machine learning techniques has been emerged as an intelligent technique to be used in large scale of this field. They were applied in early stage of plant diseases diagnosis and classification. Li et al. [2] suggested the K-means clustering segmentation method to the grape disease images. The authors in this work proposed a SVM classifier which was designed using thirty one of significant features that were selected to identify both of grape downy mildew disease and grape powdery mildew disease. The classification rates in testing phase were respectively 90% and 93.33%. Athanikar and Badar [3] implemented Neural Network to classify the potato leaf image into category of healthy and diseased. Their results demonstrated that BPNN could effectively detect the spots leaf disease and could particularly categorized the disease type with accuracy 92%.

Recently, the deep learning is getting more interest, particularly the Convolutional Neural Network (CNN). CNN is a type of deep learning structure that was designed for the classification purposes especially for digital image classification and it has been regarded as one of the best approaches for pattern recognition tasks. The Manual extraction of features in the conventional techniques is a tedious task and researches need mostly a lot of time to test and extract the suitable features for the classification. On contrast, CNN can automatically extract features of image by tuning the parameters in both of convolutional and pooling layers. CNN also has another advantage which is the ability of learning from big data set. On contrary, the algorithms of the conventional machine learning usually needs only hundreds samples that are used for training but when larger training sets are used, these algorithms converge slowly or maybe cannot converge. Deep learning method has the robustness and the ability of generalization so that it outperforms in many fields such as: signal processing [4], pedestrian detection [5], face recognition [6], road crack detection [7], and biomedical image analysis [8]. Deep learning techniques have also accomplished impressive outcome in the agriculture field and were benefit for horticultural workers and smallholders including: recognition of weeds [9] selection of fine seeds [10], pest identification [11], fruit counting [12], and research on land cover [13]. The wide spread of deep CNNs in the agriculture field has lead to a big progress especially in plant diseases classification, in which they can find high variance of pathological symptoms in visual appearance. In addition to this, CNN can find the high dissimilarity in intra-class and

even the low similarity between inter-classes that perhaps are noticed only by the botanists [14]. More studies of using CNN in the field of crop disease recognition and identification as a new hot spot research in agricultural field were presented in [15-23]. These studies demonstrated that CNNs have not only reduced the requirements of image preprocessing, but also improved the accuracy of diseases recognition. Lee et al. [24] proposed a CNN approach to identify leaf images and reported an average accuracy of 99.7% on a dataset covering 44 species, but the scale of datasets was very small.

In this paper three pre-trained deep networks based on the transfer learning and fine tuning are suggested for tomato leaf diseases classification. The performance analysis of the three networks in existence of different number of images and with variation to the values of learning parameters are evaluated through the comparison and results analysis. The rest of the paper is organized as follows. Section 2 reviews the related works of the application of transfer learning to CNN in a different fields. Section 3 introduces the structures of the three pre-trained networks: AlexNet, VGG-16Net and SqueezeNet in addition to the used data set of tomato leaf diseases. The experimental results of the fine tuning and the transfer learning -based pre-trained networks are presented in Section 5. Finally Section 6 concludes the analysis to the suggested networks performances and the results in addition to the future work.

II. RELATED WORKS

Deep learning networks can implemented the transfer learning either by using the pre-trained network to extract attributed features that can be applied to a new field or via fine tuning the weights of network through its training with a new data set. The transfer learning has been applied to deep networks in different domains with different applications. In the work [25], the transfer learning was utilized in the biometrics domain in which the joint probabilistic was exploited for face recognition to cope the problem of insufficient images of the wanted identities. Also, among the applications of applying transfer learning-based deep leaning in the biometrics are the ear recognition as they were introduced in the works [26-27]. In [28], they also compared AlexNet architecture, the 16-layer VGG model architecture, and the latest SqueezeNet architecture for ear recognition using limited training data. The outcomes of this work showed that the architecture of SqueezeNet trained by using the learned parameters with ImageNet data and its fine tuned through utilizing 1383 of ear images was the best model. In [29], the authors applied transfer learning to the well-known AlexNet Convolution Neural Network (AlexNet CNN) for human recognition based on ear images. The work in [30] compared the performance of the pre-trained CNN AlexNet with the same network but with its fine tuning for the application of Arabic characters recognition. Their results proved that the transfer learning based on fine tuning to AlexNet produced a higher accuracy compared to the same AlexNet model without tuning as a fixed feature extractor. Meanwhile, the plants identification using the 2015 LifeCLEF dataset based on the transfer learning through the fine tuning to the pre-trained deep networks GoogleNet, VGGNet and AlexNet as proposed in [31]. Their output results showed that the most affecting factor on the performance of the transfer learning based-fine tuning

was the number of iterations. Mohanty et al. [17] made a fine tuning to deep learning models that pre-trained on ImageNet to be used in identifying 14 crop species with 26 leaf diseases. The models were tested on the available a public dataset including 54,306 images of healthy and diseased plant leaves collected under controlled conditions. They achieved the best accuracy of 99.35% on a hold-out test dataset**.** Zhang et al. [32] addressed the detection issue of cherry leaf powdery mildew disease using GoogLeNet which achieved accuracy of 99.6%. Their results also demonstrated that the performance of deep learning model can be boosted by the transfer learning in crop disease identification. In [33], a united convolutional neural networks (CNNs) architecture based on an integrated method is suggested. The proposed United Model is designed to distinguish leaves with the common grape diseases, it achieves an average accuracy of validation 99.17% and accuracy of test 98.57%. Also the work in [34] used the pre-trained models and multiple classifiers for detecting the potato leaf diseases. The logistic regression classifier with VGG19 outperformed the other classifiers by a classification accuracy 97.8% with the test dataset.

CNN and the three pre-trained deep networks based on the transfer learning and fine tuning are explained in the next section.

III. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Network (CNN) is emerged inspiring from the researches in human brain cortex. It is developed to extract significant features by sequential operations of convolution and pooling [35]. Convolutional layers, pooling layer, activation function layers, dropout layers, and fully-connected layers are the main layers in CNN architecture. **Convolutional layers** carry the outputs of convolution filters or kernels with preceding layer. The main parameters of these filters or kernels are the weights and biases which can be learned in each iteration through optimization function. Purpose of the optimization function is generating kernels that are a good data representation without error. **Pooling layers** are used for the down sampling to lower size of neuron and reduce the performance issue of over-fitting. Max pooling operation is the most type used in pooling layers which captures the maximum value of the pooling window. **Activation function layers** are used to add non-linearity to the network. In the literature, there are a lot of activation functions such as sigmoid, tanh and ReLU that is the most one used [12]. **Dropout layers** are used to overcome the problem of over-fitting by randomly shut down the neurons in the network. **Fully connected layers** are utilized to calculate the scores or probabilities of classes. The classifier inputs are the results of the fully connected layers, the most well-known classifier is the softmax classifier. Since CNN is a supervised learning, the loss between the ground truth data and the network output is calculated and this loss is an input to the optimization algorithm. The most common optimization algorithm is the Stochastic Gradient Descent (SGD) algorithm in which this algorithm updates the weights according to the loss value calculated in each iteration. Both of the loss function and SGD are depicted in equations 1 and 2 as follows:

$$J(\omega) = \frac{1}{n}\sum_{i=1}^{n} y_i \log(h_\omega(X_i)) \qquad (1)$$

$$\omega_{t+1} = \mu\omega_t - \alpha\nabla J(\omega_t) \qquad\qquad (2)$$

where: x is the training sample with number of input data n, y_i is the true label data and $h_\omega(X_i)$ is the predicted label of CNN network in a given current weights ω. Also, μ is the momentum weight for both of current weights ω_t and learning rate α. The most common CNN deep learning architectures such as AlexNet, SequezzeNet, and VGG-16 Net are briefly explained in the next section.

A. AlexNet

The first well known CNN was AlexNet and it was among the early successful architectures of deep learning developed by author in [36] and it consists of several layers of convolution layers, Rectified Linear Unit (ReLU) and may be with batch normalization and Max-pooling in some layers. Each layer has many kernels and each kernel is initialized randomly at the beginning of training and through the optimization function the kernels are learned. The number of kernels at the preceding layer determines the dimension depth of the convolutional layers. Through the convolution, each kernel maps the preceding layer to new space. In this paper, the pre-trained model used in our study consists of five convolutional (conv) layers and three fully connected layers as shown in Fig. 1. The first convolution layer consists of 96 filters, each one with dimension of 11 x 11 x 3 which is the height, width and depth, respectively and it is applied on an input image of size 227 x 227 x 3. Thus, the Rectified Linear Unit (ReLU) from the first convolution layer generates 96 activation map. In the same fashion, the four remaining convolutions layers for performing the convolutional operations are respectively as follows: conv2 includes 256 filters each one with dimension 5x5x48, conv3 includes 384 filters each one with dimension 3x3x256, conv4 contains 384 filters each one with dimension of 3x3x192 and finally conv5 includes 256 filters each of them with dimension 3x3x192. Findings from these layers are activation maps with various neuron which activated in each map. The convolutional layers followed by ReLU, Max-pooling and normalization layers. ReLU is a nonlinear and a non-saturating activation function which is applied to the output of both of the 5 convolution layers and the last two fully connected layers. The function of Max-pooling layers are reducing the dimension of the previous convolution layer output through finding and saving the maximum value in the concerning field. The last two fully connected layers 6 and 7 have 4096 neurons where all of them are linked to each other, while the fully connected layer 8 (fc8) has 1000 output classes as trained with ImageNet data. The objective of Dropout layer is to randomly prevent the number of a network connections for training and this showed its ability to improve the network performance over test phase [37]. The final fc8 layer is followed by the softmax and classifier with 1000 output categories in which the loss function used is the cross entropy.

B. VGG-16Net

VGG-16Net is deeply learning series network and it consists of thirteen convolutional (conv) layers [38], each layer followed by ReLU layer and its architecture is shown in Fig. 2 in which all conv layers are with green color.
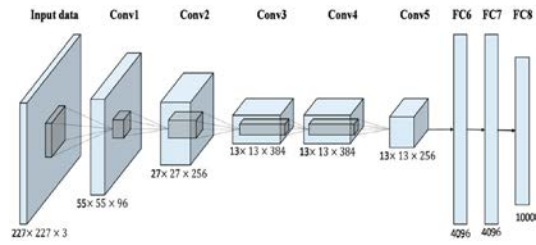
Fig. 1. The AlexNet Structure.
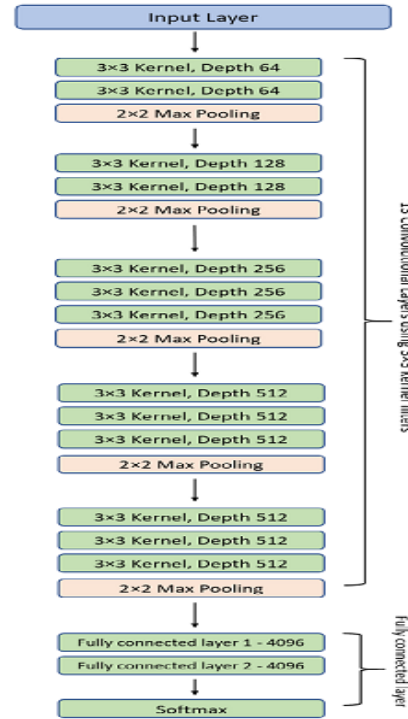
Fig. 2. VGG-16 Net Structure.

The first conv1 layer receives an input image with size equal to 224 x 224x3. The input image propagated through a set of conv layers having filters with receptive field of 3×3. Also the net architecture contains five max-pooling layers which are used for down-sampling with stride equals two. Max-pooling layers are implemented over a window of 2×2-pixels and they follow some of the conv layers. In addition, there are three fully connected layers (fc) following the conv layers with channel size equal 4096, 4096, and 1000 respectively. Each neuron in fc layer accepts the input from the activations of the previous neuron layer. The output size of 1000 in final fc layer represents the number of ImageNet categories used in training the global classifier. The final layers in the VGG-16 Net are the soft-max layer and the classification layer. The rectification non-linearity layer (ReLU) equips all the hidden layers [38]. The main advantage of using VGG-16 architecture is that it can be generalized well with any new datasets. From experimental results of VGG-16 Net applications, it was concluded that the features of the previous layers of a pre-trained network usually include information about the edge and color. With other meaning, the later layers hold features more specific to the classes' details. Hence, the earlier layers parameters of VGG-16 network does not need for

the fine-tuning as explained in [39]. Motivation from this, only the fine tuning to the last layers of the network has been proposed in the present literatures [40-41]. VGG-16 was trained by one million images or more, thus it has the capability to categorize the input images into 1000 classes.

C. SqueezeNet

It consists of 68 layers; a squeeze convolution layer comprise nine fire modules that has filters with only of size 1x1 and it feeds an expand layer that has convolution filters with size 1x1 or 3x3 [42]. Also each fire-expand layer feeds a fire-ReLU-expand layer. The final conv10 layer has 1000 output categories and followed by a classification layer. The full architecture of SqueezeNet is shown in Fig. 3 and the structure with two fire modules is shown in Fig. 4.
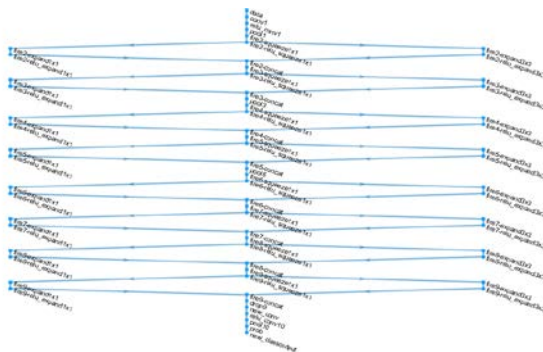
Fig. 3. SqueezeNet Structure with All Layers.
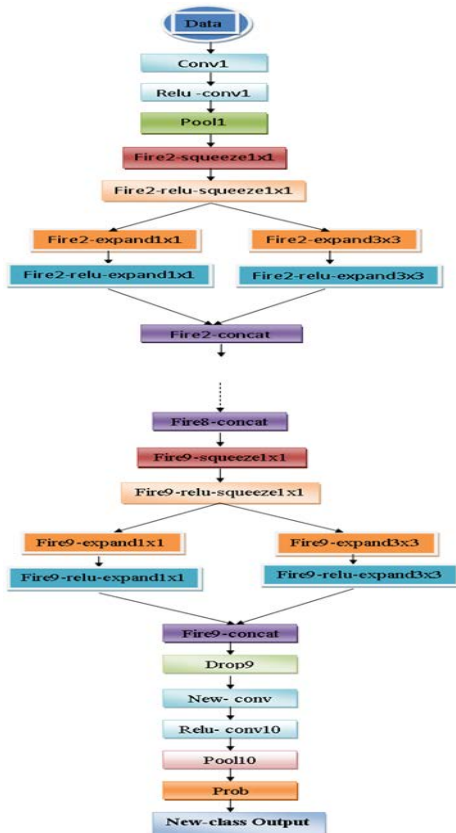
Fig. 4. SqueezeNet Structure with Exposition to the Fire Modules 2 and 9.

IV. THE PROPOSED NETWORKS MODELS AND DATASET

A brief explanation to the architecture of the pre-trained networks based on transfer learning and the fine tuning to the learning parameters in addition to the utilized dataset are introduced in this section.

A. The Pre-trained Networks based on Transfer Learning

In this paper, transfer learning is suggested for the pre-trained deep networks AlexNet, VGG-16Net and SqueezeNet with application to tomato leaf diseases diagnosis. Concerning the transfer learning to the pre-trained AlexNet, the two last fully connected layers have been modified with the desired number of categories under consideration in the field of tomato leaf diseases. Also, with regard to VGG-16 Net, the transfer learning is implemented on it through excluding only the three last layers in VGG-16 Net architecture and retaining the remaining layers of the network structure. The last three layers are substituted by a new layer of fc layer, a softmax layer, and a new classification layer so that its classification output should be suitable to the new classification task. The transfer learning is also applied to the pre-trained SqueezeNet by modifying both of the final convolution layer named conv10 and the classification layer to be convenient with the desired number of classes assigned in our study cases. Also, fine-tuned is suggested to all the pre-trained networks through assigning both factors of the learning rate of weight and learning rate of bias of the fully connected layer to are 10. In case of the fine-tuned AlexNet, the learning rate of both weight and bias are 10 and 5 times the learning rate of the fully connected layer in the global AlexNet. While these learning parameters are respectively 10 and 10 times thos of learning rate of the fully connected layer (fc8) in the global VGG-16Net. Regarding the fine-tuned SqueezeNet, the learning rate of both weight and bias are 10 and 10 times those of learning rate of the final convolutional layer (conv10) of the global SqeezeNet.

B. Data Set

In this paper, the pre-trained networks based on transfer learning and the fine tuning of the learning rate parameters and adjusting the Mini-Batch Size to be at a suitable value are carried out to resolve the diagnosis issue of tomato leaf diseases. Nine different diseases in addition to healthy leaves of tomato crop from Plant Village dataset [43] are utilized in our study. The Nine tomato leaf diseases are Late Blight (LB), Leaf Mold (LM), Septoria leaf Spot (SS), Mosaic Virus (MV), Bacteria Spot (BS), Early Blight (EB), Yellow Leaf Curl Virus (YC), Spider mites Two-spotted Spider mite (SM), Target Spot (TS). The tomato leaf images of Plant Village data set are in three channels of red, green and blue (RGB) with dimension of 256x256 with total numbers equal 18160 images. All the used images are resized to be convenient with the input data size to each network. Images of six samples from each class are shown in Fig. 5. Four diseases and healthy leaves with a few number of images for each class are only used in the first test for evaluating the performance of the suggested networks and the comparison with the work in literature. While all the above mentioned nine leaf diseased and healthy leaves with the aforementioned of total number images are used in the second test for the performance analysis and the evaluation of the suggested networks.

(a) Late Blight Samples

(b) Septoria leaf Spot Samples

(c) Bacterial Spot Samples

(d) Mosaic Virus Samples

(e) Healthy Samples

(f) LeafMold Samples

(g) Yellow LeafCurl Virus

(h) Early Blight Samples

(i) Target Spot

(j) Spider Mites- Two-Spotted
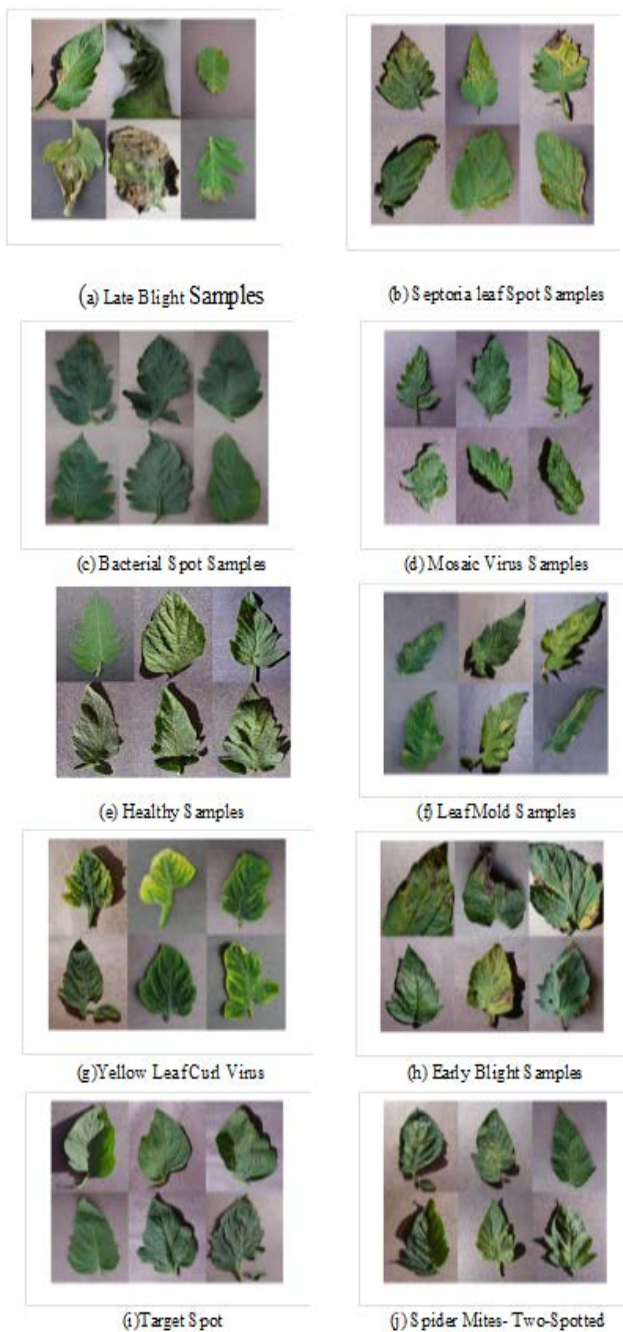
Fig. 5.   (a-j) Nine tomato Leaf Diseases and Healthy Tomato Leaf.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of the pre-trained networks based on transfer learning and the fine tuning are verified through their diagnosis results and classification for tomato leaf diseases. Two set of data in which each of them has different numbers of images with different numbers of diseases are used for evaluating the suggested networks performance. In the **first part of our study** to the networks performance evaluation, the suggested fine-tuned networks are carried out on four tomato leaf diseased and healthy leaves. The suggested networks Alex, Squeeze and VGG-16 are trained with the following tomato leaf diseases: Bacteria Spot (BS), Late Blight (LB), Spetoria

Spot (SS), Yellow leaf Curl (YC) and healthy leaves in which 100 images are used from each class. The tomato leaf images in each category are split to 80 images to be used for the training phase and the remaining images are used to the test phase. The learning parameters for training the suggested networks are tuned and selected to be as follows: the Initial Learning Rate is set to be as 0.0001, the maximum number of epochs is chosen to be 15 and finally the Mini-Batch Size parameter to the three networks is tested with the following values 5, 15, 22 and 30 respectively. The performance of the suggested networks in terms of the overall classification accuracy, the accuracy of classification for each category besides the comparison of their performance with that of the state-of-the-art technique are evaluated. The classification accuracy of the suggested networks in test phase and the elapsed time during the training of the three networks using the above mentioned values of Mini-Batch Size are depicted in Tables I and II. Also, the comparison of the three pre-trained networks in terms of the accuracy and the elapsed time during their training using the different Mini-batch size are shown in Fig. 6 and 7. The confusion matrix to the classification outputs of the suggested networks using the Mini-Batch Size values at 30, 15 and 5 are depicted in Tables III, V and VII. The Confusion matrix shows the number of samples of the corresponding detected class for each true class. Also, the classification accuracy of each class, in addition to both of the true and false samples of each class using Alex, Squeeze and VGG-16 networks at different Mini-Batch sizes are depicted in Tables IV, VI and VIII. In addition, the performance of the suggested networks is evaluated through their comparison with the classifier used in literature [15]. The authors in this paper presented a Convolutional Neural Network model and Learning Vector Quantization (LVQ) algorithm based method for tomato leaf disease detection and classification. The dataset used in their work contains 500 RGB images of tomato leaves with four symptoms of diseases in which 20 images from each class are used in the test phase of the classifier. The LVQ classifier had been fed with the output feature vector of convolution part for training the network in which the maximum number of epochs was 300. The classifier performance using the state-of-the-art technique [15] in terms of the confusion matrix, classification accuracy to each category and the average accuracy are depicted in Table IX.

From the results given in tables, the overall accuracy of classification using the three suggested networks ranged from 93% to 99% according to the Min-Batch sizes. The accuracy of classification to recognize the tomato healthy leaves was the best one and reached to 100% with all the suggested networks and at different values of Mini-Batch sizes. The accuracy of classification of all leaf diseases except the Spetoria Spot disease was ranging from 90% to 100% according to the type of used network and the size of Mini-Batch parameter. It was found that the classification accuracy of Spetoria Spot disease using SeqeezeNet with Mini-Batch sizes at 30 and 15 was poor compared to that of the other tomato diseases when using the other two suggested networks. The main reason to this low accuracy in diagnosing the Spetoria Spot disease perhaps due to the similarity of its symptom with the other symptoms of diseases and this led to difficult discrimination using squeezeNet at large size of min-batch parameter. The

classification accuracy increased to 100% in diagnosing the Spetoria Spot disease with the three suggested networks when using Mini-Batch size at value 5. Also concerning the elapsed time during training the networks, it was ranging from 8 minutes with AlexNet to almost 160 minutes with VGG-16Net which was the longest time among the other networks. Also from the results given in tables, it was verified that with the small value of Mini-Batch size, the classification accuracy rate of AlexNet is low, while with increasing the Mini-Batch size the accuracy rate of AlexNet is also increasing. On contrast, the elapsed time in training the AlexNet at a small value of Min-Batch size was larger than that time when the Mini-Batch size was large. The accuracy rate of SquezeeNet classification reduces with increasing the value of Min-Batch size, while it increases with reducing the value of Min-Batch size. On the other side, the elapsed time in training the SqueezeNet is inversely proportional to the value of Mini-Batch size. Its training execution time is big when the value of Mini-Batch size is small and vice versa. SqueezeNet takes a smallest execution time during its training when using a small value of Mini-Batch size compared to the two other networks. As VGG-16 Net is network with a deep structure, it generally takes the longest time during its training among other networks. Therefore, the execution process of training VGG-16 Net may be failed as a result of error of out of memory when increasing the size of Mini-Batch parameter to value greater than 30. On other side, at a small Mini-Batch size its classification accuracy rate was larger than its accuracy at a large value of Min-Batch size.

Regarding the comparison of our suggested networks with the classifier in literature [15], our pre-trained networks outperformed the classifier in literature. The average rate of the suggested networks accuracy with the same dataset used in literature was ranging from 93% to 99% either with a small or large Mini- Batch size and with maximum number of iterations equals 15 epochs. While, the average accuracy of this work in literature was 86% at maximum number of iterations equals 300 epochs. Hence, the accuracy rate of the suggested networks improved by 8.1% to 15% over the accuracy rate of the classifier introduced in [15].

TABLE I. THE ACCURACY OF CLASSIFICATION USING THE THREE PRE-TRAINED NETWORKS WITH DIFFERENT MINI BATCH SIZE

Pre-Trained Networks	Accuracy of classification in test phase			
	Mini Batch Size 5	Mini Batch Size 15	Mini Batch Size 22	Mini Batch Size 30
AlexNet	96%	97%	97%	99%
SqueezeNet	98%	94%	94%	93%
VGG-16 Net	99%	97%	97%	96%

TABLE II. THE ELAPSED TIME TAKEN DURING THE TRAINING OF THE THREE PRE-TRAINED NETWORKS WITH DIFFERENT MINI BATCH SIZE

Pre-Trained Networks	Elapsed Time using different Batch size in minutes			
	Mini Batch Size 5	Mini Batch Size 15	Mini Batch Size 22	Mini Batch Size 30
AlexNet	22 min	11min,29sec	12min, 45sec	8 min, 44sec
SqueezeNet	13min, 57 sec	10min,50sec	10min,50sec	9min, 51 sec
VGG-16 Net	157min	123 min	160min	155 min
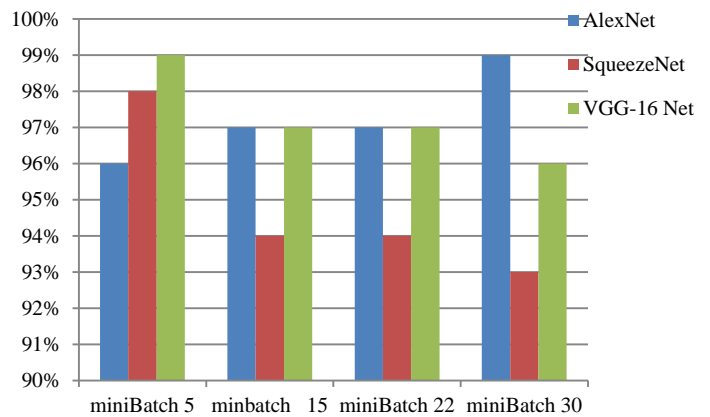
Fig. 6. The Classification Accuracy of the Pre-Trained Networks with Different Sizes of the Mini- Batch Size Parameters.
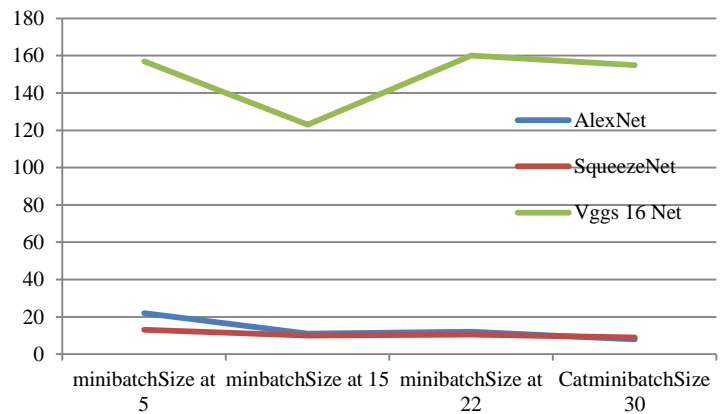
Fig. 7. The Elapsed Time of Training the Pre-Trained Networks with Different Sizes of the Mini- Batch Parameters.

TABLE III.   THE CONFUSION MATRIX TO THE CLASSIFICATION OUTPUTS USING THE ALEX, SQUEEZE AND VGG-16 NETWORKS AT MINI-BATCH SIZE=30

Confusion Matrix of the classification with Mini Batch Size=30

AlexNet — confusion matrix (True Class rows BS, H, LB, SS, YC vs Predicted Class BS, H, LB, SS, YC):

True \ Pred	BS	H	LB	SS	YC
BS	20				
H		20			
LB			20		
SS	1			19	
YC					20

SqueezeNet:

True \ Pred	BS	H	LB	SS	YC
BS	18		1	1	
H		20			
LB	1		19		
SS	2	1	1	16	
YC					20

VGG-16 Net:

True \ Pred	BS	H	LB	SS	YC
BS	18			2	
H		20			
LB			19	1	
SS				20	
YC	1				19

TABLE IV.   THE ACCURACY OF CLASSIFICATION FOR EACH LABEL LEAF OF TOMATO DISEASES WITH ITS TRUE AND FALSE SAMPLES USING ALEXNET, SQUEEZENET AND VGG-16 NET AT MINI-BATCH SIZE=30

Leaf Diseases	Accuracy at Batch Size 30								
	AlexNet			SqueezeNet			VGG-16 Net		
	Accuracy	True Sample	False Sample	Accuracy	True Sample	False Sample	Accuracy	True Sample	False Sample
B S	100%	20	0	90%	18	2	90%	18	2
H	100%	20	0	100%	20	0	100%	20	0
L B	100%	20	0	95%	19	1	95%	19	1
S S	95%	19	1	80%	16	4	100%	20	0
Y C	100%	20	0	100%	20	0	95%	19	1

TABLE V.   THE CONFUSION MATRIX TO THE CLASSIFICATION OUTPUTS USING THE ALEX, SQUEEZE AND VGG-16 NETWORKS AT MINI-BATCH SIZE=15

Confusion Matrix of the classification with Mini Batch Size=15

AlexNet:

True \ Pred	BS	H	LB	SS	YC
BS	18		1		1
H		20			
LB			20		
SS	1			19	
YC					20

SqueezeNet:

True \ Pred	BS	H	LB	SS	YC
BS	19		1		
H		20			
LB			19	1	
SS	1		2	17	
YC	1				19

VGG-16 Net:

True \ Pred	BS	H	LB	SS	YC
BS	19			1	
H		20			
LB			19	1	
SS			1	19	
YC					20

TABLE VI.   THE ACCURACY OF CLASSIFICATION FOR EACH LABEL LEAF OF TOMATO DISEASES WITH ITS TRUE AND FALSE SAMPLES USING THE ALEX, SQUEEZE AND VGG-16 NETWORKS AT MINI-BATCH SIZE=15

Leaf Diseases	Accuracy at Batch Size 15								
	AlexNet			SqueezeNet			VGG-16 Net		
	Accuracy	True Sample	False Sample	Accuracy	True Sample	False Sample	Accuracy	True Sample	False Sample
B S	90%	18	2	95%	19	1	95%	19	1
H	100%	20	0	100%	1	20	100%	20	0
L B	100%	20	0	95%	19	1	95%	19	1
S S	95%	19	1	85%	17	3	95%	19	1
Y C	100%	20	0	95%	19	1	100%	20	0

TABLE VII.   THE CONFUSION MATRIX TO THE CLASSIFICATION OUTPUTS USING THE ALEX, SQUEEZE AND VGG-16 NETWORKS AT MINI-BATCH SIZE=5

Confusion Matrix of the classification with MiniBatch Size=5

TABLE VIII.   THE ACCURACY OF CLASSIFICATION FOR EACH LABEL OF LEAF TOMATO DISEASES WITH ITS TRUE AND FALSE SAMPLES USING THE ALEX, SQUEEZE AND VGG-16 NETWORKS AT MINI-BATCH SIZE=5

Leaf Diseases	Accuracy at Batch Size 5								
	AlexNet			SqueezeNet			VGG-16 Net		
	Accuracy	True Sample	False Sample	Accuracy	True Sample	False Sample	Accuracy	True Sample	False Sample
B S	90%	18	2	90%	18	2	95%	19	1
H	100%	20	0	100%	20	0	100%	20	0
L B	90%	18	2	100%	20	0	100%	20	0
S S	100%	20	0	100%	20	0	100%	20	0
Y C	100%	20	0	100%	20	0	100%	20	0

TABLE IX.   THE CONFUSION MATRIX AND THE AVERAGE OF CLASSIFICATION ACCURACY OF THE CLASSIFIER USED IN LITERATURE[15]

Leaf diseases	B S	L B	S S	Y C	H	Accuracy of each class
B S	18	0	0	2	0	90%
L B	0	17	0	3	0	85%
S S	0	0	16	3	1	80%
Y C	0	0	3	17	0	85%
H	0	0	0	2	18	90%
Average Accuracy						86%

In the **second part of our study** to the networks performance evaluation for tomato leaf diseases classification, the fine-tuned networks are applied on tomato leaves with the aforementioned nine diseases and healthy leaves. Large numbers of the given images for each class are used in this part, the number of images range from 373 images of Mosaic Virus disease to 5357 images of Yellow Curl disease. The dataset is split randomly into 0.8 that is utilized for the training phase and the remaining of them is used for the test phase. In order to adjust the last three layers of the pre-trained Alex and Squeeze networks for the new classification task with 10 categories, both the learning rate of weight and learning rate of bias of the fully connected layer are set to be 10. Also both of AlexNet and SqueezeNet are trained by setting the following parameters values: maximum number of epochs at 15, the learning rate at 0.0001 and the Mini-Batch size at 30. The confusion matrix to the classification results by the fine-tuned AlexNet is depicted in Table X. Also, the accuracy rate of classification for each class of tomato leaves, the number of true and false samples and the average accuracy of overall classification are depicted in Table XI. The training progress and the loss values during training AlexNet against the number of epochs are shown in Fig. 8. Also, the confusion matrix of classification results by the fine-tuned SqueezeNet is depicted in Table XII. Furthermore, the accuracy rate of classification for each class of tomato leaves, the number of true and false samples and the average accuracy of overall classification are depicted in Table XIII. Fig. 9 shows the training progress and the loss values of training SqueezeNet against the number of epochs.

TABLE X.   THE CONFUSION MATRIX OF CLASSIFICATION USING ALEXNET WITH THE LARGE DATASET OF 9 TOMATO LEAF DISEASES AND HEALTHY LEAVES

Predicted Classes

True \ Pred	BS	EB	H	LB	LM	MV	SM	SS	TS	YC
BS	423	1							1	
EB	10	182		3				1	4	
H		1	315					1	1	
LB		9	1	365	2		1	2	2	
LM	1	1			180	1	1	4	1	1
MV					1	73	1			
SM					1	1	327		6	
SS	2	2		1	2	1	1	343	2	
TS	1				2	2	13	3	259	1
YC										1071

TABLE XII.   THE CONFUSION MATRIX OF CLASSIFICATION USING THE SQUEEZENET WITH THE LARGE DATASET OF 9 TOMATO LEAF DISEASES AND HEALTHY LEAVES

Predicted Classes

True \ Pred	BS	EB	H	LB	LM	MV	SM	SS	TS	YC
BS	417	3		1			2	1		1
EB		181		9			1	6	3	
H	1		316					1		
LB	1	9	1	366	1					4
LM				1	185	2	1		1	
MV					1	73	1			
SM					1		331		2	1
SS	4	5		3	1			339	2	
TS	5	3		1	1		9	3	259	
YC	8						1			1062

TABLE XI.   ACCURACY OF CLASSIFICATION WITH THE TRUE AND FALSE SAMPLES USING ALEXNET WITH LARGE DATASET OF 9 TOMATO LEAF DISEASES AND HEALTHY LEAVES

Leaf Diseases	Total Samples	Test samples	True samples	False sample	accuracy
B S	2127	425	423	2	99.5%
E B	1000	200	182	18	91%
Healthy	1591	318	315	3	99.1%
L B	1909	382	365	17	95.5%
L M	952	190	180	10	94.7%
M V	373	75	73	2	97.3%
S M	1676	335	327	8	97.6%
S S	1771	354	343	11	96.8%
T S	1404	281	259	22	92.2%
Y C	5357	1071	1071	0	100%
Accuracy					97.4%

TABLE XIII.   THE ACCURACY OF CLASSIFICATION WITH THE TRUE AND FALSE SAMPLES USING THE SQUEEZENET AND THE LARGE DATA SET OF 9 TOMATO LEAF DISEASES AND HEALTHY LEAVES

Leaf Diseases	Total Samples	Test samples	True samples	False sample	Accuracy %
B S	2127	425	417	8	98
E B	1000	200	181	19	91
H	1591	318	316	2	99
L B	1909	382	366	16	95.8
L M	952	190	185	5	97
M V	373	75	73	2	97
S M	1676	335	331	4	98.8
S S	1771	354	339	15	95.7
T S	1404	281	259	22	92
Y C	5357	1071	1062	9	99
Accuracy					97.2%

Fig. 8.   The Training Progress of the Tuned AlexNet Versus the Number of Epochs.
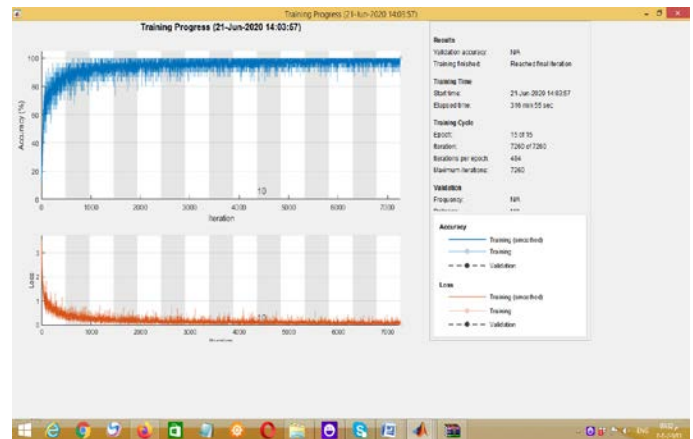
Fig. 9.   The Training Progress of the Tuned SqueezeNet Versus the Number of Epochs.

From the results given in tables, the accuracy of AlexNet of tomato leaf diseases classification in test phase was 97.4% while the elapsed time taken in its training was 296 minutes and 50sec. On the other side from the results given in tables with SqueezeNet, the accuracy of tomato leaf diseases classification in test phase was 97.2% and the elapsed time taken in training the SqueezeNet was almost 316 minutes and 50sec (5hour and 17minutes). The classification accuracy of tomato leaves with symptoms of Early Blight either using AlexNet or SqueezeNet was the lowest among the diagnosis accuracy of other tomato diseases. The reason may be back to that the Early Blight disease appears first on old and mature leaves near the base at stem end of fruits with a spot of ring shape and this made a difficulty in its discrimination. AlexNet achieves high accuracy of diagnosing YC disease that reaches to 100% and was larger than SqueezeNet that achieves accuracy reaching to 99%. The accuracy of classification in diagnosing Target Spot disease with both of the fine-tuned Alex and Squeeze networks has an acceptable and was almost 92.2% due to its similar symptoms with the Spider Mites symptoms. Also, because the similarity of symptoms of both Spider mites and target spots, there are false numbers of them with both of their attributed categories. On the other side, AlexNet and SqueezeNet prove their ability in diagnosing the other diseases categories with high classification accuracy reached up to 99%. Also due to the deep learning structure of VGG-16 Net comparing to the structure of other networks, it was verified that this network was costly computationally. It was found that the time needed for training VGG-16 Net reached to 69 hours in which it was the largest among other networks. Therefore, only the results of AlexNet and SqueezeNet were enough for exposition.

VI. CONCLUSION

In this paper, the classification of tomato leaf diseases utilizing the images from Plant Village dataset was performed by the suggested pre-trained deep networks AlexNet, squeezeNet and VGG-16 Net. The main challenge in tomato diseases diagnosis and classification in our study was that the symptoms of tomato leaf diseases are very similar to each other which results in some leaves may be embedded and classified into wrong classes. The accuracy of classification of AlexNet, SqueezeNet and VGG-16Net using 500 images of tomato leaves as assigned to **the first part** of the work and with Mini-batch size at 30 were 99%, 93% and 96% respectively. Whereas, the classification accuracy of AlexNet, SqueezeNet and VGG-16 Net using the same number of images and with Mini-batch size at 5 were 96%, 98% and 99%, respectively. Furthermore, the performance of the three fine-tuned networks for tomato leaf diseases diagnosis is evaluated through the comparison with that of the-state-of-the-art technique. The accuracy rate of our pre-trained networks increased by 8.1% to 15% over the value of accuracy rate of the classifier introduced in literature. The execution time of training AlexNet using small dataset and with Mini-Batch size at 30 was the shortest among training times for other networks. Also, the performance of AlexNet in terms of both classification accuracy and elapsed time using dataset of 18160 images as assigned to **the second part** of the work was efficient network and outperformed over to other networks. It achieves

classification accuracy of 97.4% with elapsed time in training of almost 296 minutes. On contrary, VGG-16 Net has large execution time during its training either using small Mini-Batch size or large Mini-Batch size compared to that of other networks.

In the future work, Internet of Things and mobile applications are suggested with the deep learning CNN to identify and classify the plant diseases type.

REFERENCES

[1] Srdjan S, Marko A, Andras A, Dubravko C, Darko S. "Deep neural networks based recognition of plant diseases by leaf image classification". Comput Intell Neurosci 2016;2016:1–11

[2] Li G, Ma Z, Wang H. Image recognition of grape downy mildew and grape powdery mildew based on support vector machine. In: Proc international conference on computer and computing technologies in agriculture. Beijing, China. pp. 151–162, 2012

[3] Athanikar G, Badar P. Potato leaf diseases detection and classification system. Int J Comp Sci Mob Comput 2016;5 (2):76–88.

[4] Xie D, Zhang L, Bai L. Deep learning in visual computing and signal processing. Appl Comput Intell Soft Comput 2017;2017:1–13.

[5] Ouyang W, Wang X. Joint deep learning for pedestrian detection. In: Proc. IEEE international conference on computer vision. Sydney, Australia. p. 2056–63.

[6] Sun Y, Wang X, Tang X. Hybrid deep learning for face verification. IEEE Trans Pattern Anal. Mach Intell 2016;38 (10):1997–2009.

[7] Zhang L, Yang F, Zhang Y, Zhu Y. Road crack detection using deep convolutional neural network. In: Proc. IEEE international conference on image processing. Arizona, USA. p. 3708–12, 2016.

[8] Zhou Z, Shin J, Zhang L, Gurudu S, Liang J. Fine-Tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In: Proc. IEEE conference on computer vision and pattern recognition. Honolulu, Hawaii. p. 7340–9.

[9] Lottes P, Behley J, Milioto A, Cyrill Stachniss C. Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. IEEE Robot Autom Lett 2018;3(4):2870–7.

[10] Wang X, Cai C. Weed seeds classification based on PCANet deep learning baseline. In: Proc IEEE Asia-Pacific signal and information processing association annual summit and conference. Hong Kong, China. p. 408–15.

[11] Cheng X, Zhang Y, Chen Y,Wu Y, Yue Y. Pest identification via deep residual learning in complex background. Comput Electron Agric 2017;141:351–6.

[12] Rahnemoonfar M, Sheppard C. Deep count: fruit counting based on deep simulated learning. Sensors 2017;17(4):905.

[13] Zhu H, Liu Q, Qi Y, Huang X, Jiang F, Zhang S. Plant identification based on very deep convolutional neural networks. Multimed Tools Appl 2018;77(22):29779–97.

[14] Kussul N, Lavreniuk M, Skakun S, Shelestov A. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci Remote S 2017;14(5):778–82..

[15] Melike Sardogan, Adem Tuncer and Yunus Ozen, "Plant Leaf Diseases Detection and Classification Based on CNN with LVQ Algorithm" pp. 382-385, 3rd International Conference on Computer Science and Engineering, 2018 IEEE.

[16] Mohanty S, Hughes D, Salathe´ M. "Using deep learning for image-based plant disease detection". Front Plant Sci 2016;7:1419.

[17] Mohanty, S.P.; Hughes, D.; Salathe,M." Inference of Plant Diseases from Leaf Images through Deep Learning", arXiv. 2016. Available online: https://www.semanticscholar.org/paper/Inference-of-Plant-Diseases-from-Leaf-Imagesthrou-Mohanty-Hughes/62163ff3cb2fbbf5361e340f042b6c288d3b8e6a (accessed on 28 December 2017).

[18] Sladojevic, S.; Arsenovic, M.; Anderla, A.; Culibrk, D.; Stefanovic, D. Deep neural networks based recognition of plant diseases by leaf image

classification. Comput. Intell. Neurosci. **2016**, 2016. [CrossRef] [PubMed]

[19] Hanson, A.M.J.; Joy, A.; Francis, J. "Plant leaf disease detection using deep learning and convolutional neural network", Int. J. Eng. Sci. Comput. **2017**, 7, 5324–5328.

[20] Lu, Y.; Yi, S.J.; Zeng, N.Y.; Liu, Y.; Zhang, Y. "Identification of rice diseases using deep convolutional neural networks". Neurocomputing **2017**, 267, 378–384. [CrossRef]

[21] Tan, W.X.; Zhao, C.J.; Wu, H.R. "CNN intelligent early warning for apple skin lesion image acquired by infrared video sensors. High Technol. Lett. **2016**, 22, 67–74.

[22] Kawasaki, Y.; Uga, H.; Kagiwada, S.; Iyatomi, H. "Basic study of automated diagnosis of viral plant disease using convolutional neural networks. In Proceedings of the 12th International Symposium on Visual Computing, Las Vegas, NV, USA, 12–14 December 2015; pp. 638–645.

[23] Fuentes, A.; Yoon, S.; Kim, S.C.; Park, D.S. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors **2017**, 17, 2022. [CrossRef] [PubMed].

[24] Lee SH, Chan CS, Wilkin P, Remagnino P. Deep-plant: plant identification with convolutional neural networks. In: Proc IEEE international conference on image processing. Quebec, Canada. p. 452–6.

[25] N. Nuo, W. Tay, J. Botzheim, and N. Kubota, "Joint probabilistic approach for real-time face recognition with transfer learning," Rob. Auton. Syst., vol. 75, pp. 409–421, 2016.

[26] Ž. Emeršic, D. Štepec, V. Štruc, and P. Peer, "Training convolutional neural networks with limited training data for ear recognition in the wild," arXiv preprint arXiv:1711.09952, 2017.

[27] Y. Zhang, and Z. Mu, (2017). "Ear detection under uncontrolled conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks," Symmetry, vol. 9, no. 4, pp. 1-19, 2017.

[28] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360, 2016.

[29] Ali Abd Almisreb, Nursuriati Jamil and N. Md Din, "Utilizing AlexNet Deep Transfer Learning for Ear Recognition". 2018 Fourth International Conference on Information Retrieval and Knowledge Management, March, 2018 IEEE, pp. 8-12.

[30] C. Boufenar, A. Kerboua, and M. Batouche, "Investigation on deep learning for off-line handwritten Arabic character recognition," Cogn. Syst. Res., 2017. high-level semantics," Pattern Recognit. Lett., vol. 72, pp. 82–90, 2016.

[31] M. M. Ghazi, B. Yanikoglu, and E. Aptoula, "Plant identification using deep neural networks via optimization of transfer learning parameters," Neurocomputing, vol. 235, pp. 228–235, August 2017.

[32] Zhang K, Zhang L, Wu Q. Identification of cherry leaf disease infected by Podosphaera Pannosa via convolutional neural network. Int J Agric Environ Inf Syst 2019;10(2):98–110.

[33] Miaomiao Ji, Lei Zhang , Qiufeng Wu, "Automatic grape leaf diseases identification via UnitedModel based on multiple convolutional neural networks, INFORMATION PROCESSING IN AGRICULTURE, October 2019,pp. 1-9, journal homepage: www.elsevier.com/locate/inpa.

[34] Divyansh Tiwari Mritunjay AshishNitish Gangwar, "Potato Leaf Diseases Detection Using Deep Learning", Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2.

[35] Y. LeCun and Y. Bengio, Convolutional networks for images, speech, and time series, in The Handbook of Brain Theory and Neural Networks, Vol. 3361 (MIT Press, 1995), pp. 255–258.

[36] A. Krizhevsky, I. Sutskever and G. H. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances* in Neural Information Processing Systems, 2012.

[37] Srivastava, N., Hinton, G., Krizhevsky, A, Sutskever, I., Salakhutdinov, R. (2015) "Dropout: a simple way to prevent neural network from overfitting". *Journal of Machine Learning Research,* 15, 1929-1958**.**

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Prepr. arXiv1409.1556*, 2014.

[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[40] Hughes, D.P., Salathe, M. (2015) "An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics", arXiv:1511.08060.

[41] P. K. Sonawane and S. Shelke, "Handwritten Devanagari Character Classification using Deep Learning," in *2018 International Conference on Information, Communication, Engineering and Technology (ICICET)*, 2018, pp. 1–4.

[42] Forrest N. Iandola et al. \SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size". In: CoRR abs/1602.07360,(2016) , url: http://arxiv.org/abs/1602.07360.

[43] S. Lu, Z. Lu, and Y.-D. Zhang, "Pathological brain detection based on AlexNet and transfer learning," J. Comput. Sci., vol. 30, pp. 41– 47, 2019.

A New Big Data Architecture for Real-Time Student Attention Detection and Analysis

Tarik Hachad[1], Abdelalim Sadiq[2]
Laboratory of Information Modelling and Communication
Systems, Faculty of Sciences
Ibn Toufail University
Kenitra, Morocco

Fadoua Ghanimi[3]
Laboratory of Technological Information and Modeling
Faculty of sciences Ben M'sick
University Hassan II Casablanca
Morocco

Abstract—**Big Data technologies and their analytical methods can help improve the quality of education. They can be used to process and analyze classroom video streams to predict student attention, this would greatly improve the learning-teaching experience. With the increasing number of students and the expansion of educational institutions, processing and analyzing video streams in real-time become a complicated issue. In this paper, we have reviewed the existing systems of student attention detection, open-source real-time data stream processing technologies, and the two major data stream processing architectures. We also proposed a new Big Data architecture for real-time student attention detection.**

Keywords—Attention detection; big data analysis; stream processing; real-time processing; Apache Flink; Apache Spark; Apache Storm; Lambda architecture; Kappa architecture

I. INTRODUCTION

Student attention plays a significant role in the teaching-learning operation. It allows the student to focus on information and ignore any disturbing or distracting factor. The teacher can easily and in a natural way know if a student is in a state of attention or not. In small classrooms, students are naturally more engaged than in the largest one with a large student population or in an amphitheater. Indeed, a small classroom allows an environment that promotes student engagement as long as it is easier to monitor. In contrast, the more the classroom is large it influences the students' attention. The teacher will have to spend more time to draw students' attention and ended up losing control over part of the students. The automation of the continuous detection and evaluation of the student's attention during the lecture is the optimal solution for large audiences. In fact, it offers the teacher the possibility of knowing the attention level of the students at any time during the course. It can also notify the teacher of students with a very low level of engagement or those who are lost during the course session. Like that, the teacher can send corrective messages to less engaged students or review the course so that can be more attractive. In a previous work, we set up an architecture for the detection and the analysis of the student's attention through the use of different technologies of facial and body expressions detection. The system has been designed for monitoring a classroom with a limited number of students. It is based on the analysis of the video stream generated by a high definition camera placed in front of a classroom. The axes of analysis used are facial expressions, gaze direction, and body

gestures. The analysis results of these axes are merged to deduce the level of attention of each student. The results must be obtained in real-time, for this, we opted for a parallelized computation. The analysis tasks are time-consuming, especially for a high definition image stream with a high frequency of 30 images/second. The generalization of this system on an entire school or a university will explode the number of images received by the system, that must be processed simultaneously and the output results must be provided in real-time. This high scaling requires the use of Big Data technologies in order to overcome these issues.

The main objective of this article is to present a state of the art of existing student attention detection systems and some concepts of Big Data. We also present in detail and with a comparison the different tools and architectures allowing real-time stream processing. Finally, we propose our own architecture based on the comparisons made.

II. STATE OF THE ART

A. Existing Systems

Most of the works on detecting student attention has focused on the concept of attention and its relationship to different facial and body features.

Whitehill et al. [1] based their study on the analysis of facial expressions. The goal of their work is the development of an automated system for real-time recognition of student engagement.

Krithika et al. [2] have worked on the analysis of student concentration in an online learning environment. The main idea of their work is to be able to predict the level of concentration of the student from two measurements namely the rotation of the face and the eyes' movements.

Zaletelj et al. [3] admit that the problem of detecting student attention was to establish the correct correlation between the student's attention and the teacher's observations. For this, they have designed a system that uses the capabilities of the Kinect One sensor. This device makes it possible to collect behavioral data from students in a non-intrusive way. They proposed a method to match the features of the data collected to the students' facial and body expressions. Then, they applied machine learning methods to build models for predicting student's attention.

Vettivel et al. [4] tried to establish the relationship between attention and human parameters such as heart rate variation, facial expressions, and brain waves. They used appropriate sensors to collect information on the three student parameters during the course. The student is alerted whenever he loses his concentration. This solution combines the three parameters to increase the accuracy of the system. The features are extracted from the collected data then they are classified to predict the attentive and non-attentive states.

Goldberg et al. [5] developed a manual rating instrument, to continuously measure the observable behavior of students. They used then computer vision techniques to perform automated analysis of video recordings to extract features of the students' head pose, gaze direction, and facial expressions. Using these extracted features, they tried to estimate manually annotated attention levels for each student. As they opted for continuous labeling, a regressor is trained to relate the visible features to the manual labels. For more precision, they took into account the synchronous behavior of the neighboring students.

B. Big Data

In education, Big Data technologies can be used to collect and analyze huge amounts of information about students in order to develop more effective learning. This makes the experience more practical, especially for large establishments with classes whose size is constantly increasing.

Big Data is an industrial term that was coined to describe huge volumes of data that we have never had or processed before. It also brings together techniques for storing, analyzing, and visualizing the results obtained from more varied and complex massive data structures.

The International Data Corporation (IDC) defines the Big Data as a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high velocity capture, discovery, and/or analysis [6].

According to Jason Bloomberg [7] Big Data is a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.

In accordance with the Gartner definition of Big Data, which articulates its definition in three parts: Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [8][7].

Big Data is a very broad concept that includes three essential dimensions: volume, variety, and velocity. This requires a real revolution in the methods of storage and data processing. Fig. 1 highlights the constraints raised by the Big data.

The volume which designates the size of the data is now greater than terabytes and petabytes. The rapid transition to these scales greatly exceeds the traditional storage and processing capacities[9][10][11].
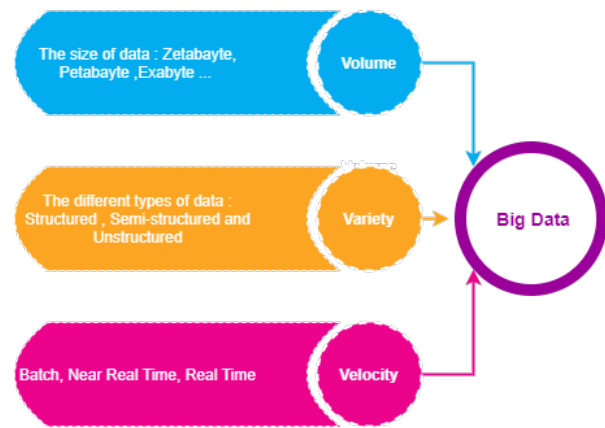
Fig. 1. The Three vs of Big Data.

Variety: This characteristic amplifies the challenge of Big Data. Since we must not only manage structured data but also semi-structured and mainly unstructured data. The vast majority of Big Data is in unstructured or semi-structured formats, such as text files, log files, audio, image, video files, social media updates, machine data, and sensors signals, etc. [9][11].

Velocity refers mainly to the speed at which data is being generated, produced, created, or refreshed. It is required not only for Big Data, but also for data processing. For time-limited processes, Big Data should be used as it streams into the organization in order to maximize its value [9][10][11].

Big Data does not primarily mean a huge amount of data or a database, they constitute the core set of technologies and components for large-scale data processing and analysis. Regardless of the type of data (structured, unstructured, or semi-structured), the data can be in one of the following three states: Data at Rest, Data in Motion, and Data in Use. Each of the formats previously mentioned requires specific processing methods.

In this paper, we deal with a case of processing data in motion, since the data is a stream of images from multiple high definition cameras. In what follows, we will present certain data stream processing tools with a comparison in order to choose the tool that best satisfies our need and that performs real-time processing.

III. STREAM PROCESSING PLATFORMS

There are many architectures and platforms to choose from. However, selecting the right architecture with the right implementation is often difficult. In what follows we will present the most efficient stream processing tools with the details of their features as well as the most used Big Data architectures.

A. Apache Flink

Apache Flink [7] is an open-source platform that came from Berlin TU University. It supports both batch and stream processing and can guarantee an exactly-once-processing. The Flink cluster architecture is illustrated in Fig. 2. Flink is scalable, has an in-memory option, and provides input APIs in

Scala and Java. It can be integrated into the Hadoop ecosystem (HDFS, YARN), or run in a completely independent way since it has its own runtime. The core of Flink is a distributed streaming dataflow, accepting programs structured as graphs (JobGraph) of activities that produce and consume data. Each JobGraph can be executed using one of the different distribution options available for Flink (like single JVM, YARN, or cloud) [8]. Flink's processing model applies transformations to parallel data collections [9][10][11].

DataStreams represent the abstraction of Flink streams. They are similar to Storm tuples, in the form of partially ordered recording sequences. DataStreams are fed by data from different external sources such as log files, message queues, etc. DataStreams support multiple operators such as map, filtering, and reduction in a parallelized way.

FlinkML has its own machine learning library. Moreover, it provides an adapter for the SCALABLE ADVANCED MASSIVE ONLINE ANALYSIS (SAMOA) library, which offers a variety of machine learning algorithms for stream processing.

Flink fault-tolerance approach is based on snapshots over distributed checkpoints that maintain the status of jobs. Snapshots act as consistency checkpoints to which the system can return in case of failure [12].

B. Storm

Apache Storm [13] is an open-source platform for real-time stream processing, written in Java and Clojure. The origin of this project goes to the Back type company specialized in the analysis of social media data and which was sold later to Twitter. The project became an Apache top-level project in September 2014 [14]. The main data structure in Storm is the tuple, a list of serializable values that are user-defined types. Fig. 3 explains the architecture of Storm.
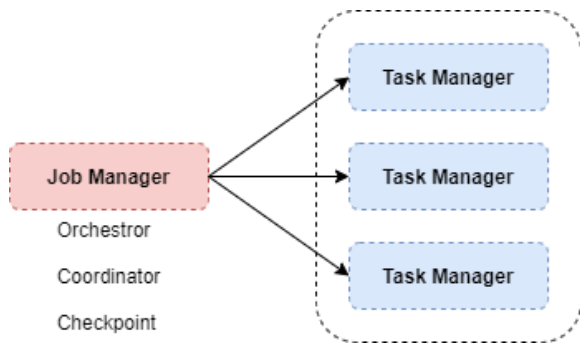
Fig. 2. Architecture of Flink Cluster.
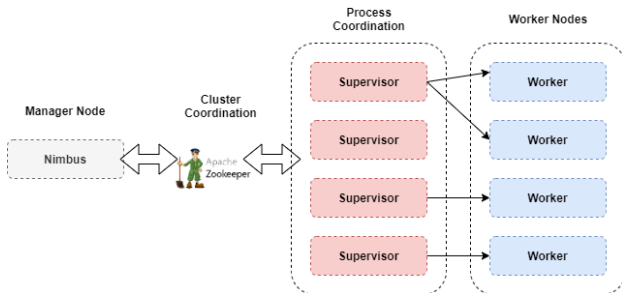
Fig. 3. Storm Cluster Architecture.

The backbone of Storm Architecture is spouts and bolts. A spout is the entry point to the stream, it connects to an external data source such as a message queue and retrieves the data in tuple format. Stream processing is performed by bolt nodes. Each bolt performs a transformation of limited complexity, in this way, several bolt nodes are grouped in a coordinated manner to perform the entire computation. A Storm application can be defined through a topology of spout and bolt nodes forming a directed acyclic graph (DAG), with arcs representing streams of tuples flowing from one node to another [15].

Apache Storm handles fault tolerance through its backup and acknowledgment (ACK) mechanism imposed by its topology. To ensure that the tuples are re-processed after failure, the spouts will always keep the tuples in their output queues until the bolts acknowledge them. When a tuple enters the topology, the spout which receives it adds an identifier, then sends this identifier to the acker bolt. The aker bolt is as a register for all tuples Ids that enter the Storm topology. Once a bolt processes a tuple, it sends an ACK to the acker bolt. While the tuples leave the topology, the acker bolt removes their identifiers. In case of failure, all tuples that have not received an acknowledgment will be re-processed.

Although Storm does not come with a machine learning library, it interfaces perfectly with the SAMOA tool offering implementations for classification and clustering algorithms for Big Data streams.

C. Spark

Spark is a project supported by the Apache community. It was initiated by the Berkeley University of California, it is a distributed data processing platform, written in Java and Scala. Fig. 4 shows the architecture of Spark cluster. Spark has four libraries running on top of the Spark engine: Spark SQL, MLlib for machine learning, GraphX, and Spark Streaming for stream processing.

Spark stores data in a distributed data set called resilient distributed dataset (RDD), which represents an immutable collection of read-only fault-tolerant objects (Python, Java, or Scala). Data will be stored in memory in a partitioned manner across multiple machines in the cluster. The sequence of operations to obtain a particular RDD is stored in the RDD itself, so in case of failures, the RDD can be rebuilt.

The operations that can be performed on RDDs are: The transformation which allows the construction of a new RDD from one or more input RDDs and the action which is an operation producing a single value or writing of the output on the disk.
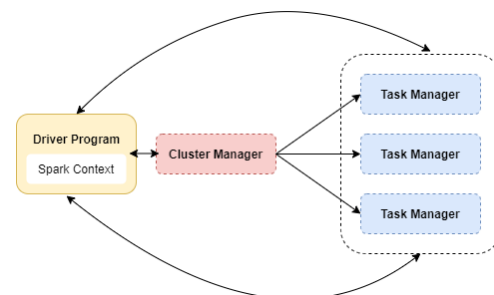
Fig. 4. Architecture of Spark Cluster.

The stream abstraction is called Discrete Stream (D-Stream) defined as a set of short, stateless, deterministic tasks. In Spark, streaming computation is treated as a series of deterministic batch computations on small time intervals [12]. In this approach, an incoming stream is packaged into sequences of small chunks of data, which can then be processed by a batch system [16]. While this may be adequate for many projects, it is not a true real-time system [14].

Table I summarizes the main features of stream processing systems. Flink and Storm perform real streaming while Spark is able to do micro-batching. The three systems present two different programming models: the compositional model which provides basic building blocks, such as spouts and bolts while the declarative model offers operators to generate functional code capable of automatically creating the topology.

TABLE I.   DISTRIBUTED STREAM PROCESSING ENGINES

	Spark	Flink	Storm
Source Model	Open source	Open source	Open source
Architecture	Master-Slave	master-slave	master-slave
Coordination	Zookeeper	Zookeeper	Zookeeper
Execution model	Batch, micro-batch	Batch, streaming	Streaming
Stream abstraction	DStream	DataStream	Tuple
Supported languages	Java, Python, R, Scala	Java, Scala	Any
Associated ML tools	MLlib, Mahoot, H2O	FlinkML SAMOA	SAMOA
In-memory processing	yes	yes	yes
Low latency	<= Time of micro-batching	yes	yes
Fault tolerance	yes	yes	yes

IV. BIG DATA ARCHITECTURES BENCHMARKING

In this section, we will review the two main stream processing architectures, Kappa and Lambda. We will then compare these two architectures in order to propose the architecture which best suits our use case.

A. Lambda Architecture

Lambda architecture was introduced by Marz and Warren [17] and provides a set of architectural principles to ingest and process both stream and batch data, in a single Big Data architecture [18]. The main idea behind the Lambda architecture is to combine real-time and batch processing in a single technology stack. This ensures low latency and provide better results. Fig. 5 shows the main components of the architecture. In fact, it combines several paradigms into a single system that breaks down processing into three layers: Batch, serving, and speed. The batch layer is generally based on Hadoop technology, it stores the raw data as soon as it arrives and calculates the views that will be sent to the serving layer for indexing and tracking results. Batch processing is performed regularly at a defined interval on all data. The speed

layer processes new data that are not already delivered in the batch view for rapid consumption. Incoming data are sent to both the batch layer and the speed layer for processing. When the system is queried, the results of the two layers will be merged to calculate the response. The views from the batch layer and those from the speed layer in near real-time are sent to the serving layer. This later indexes the views transmitted by the batch and speed layers so they can be queried in Ad-Hoc with low latency.
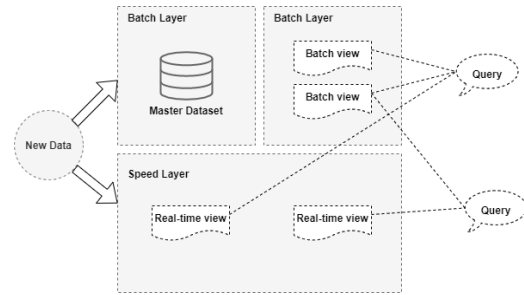
Fig. 5.   Lambda Architecture.

B. Kappa Architecture

Unlike Lambda, Kappa is only focused on the processing of data streams. Described for the first time by Kreps [19], the key idea for its implementation is not to replace the Lambda architecture but to use a single layer in real-time to process both stream and batch data. Fig. 6 shows the two layers of the Kappa architecture: The streaming layer responsible for data processing and the serving layer for querying results. Only one code is used for data processing, unlike lambda where it is necessary to generate a code version for each layer. Similarly, queries only search in one location instead of two for Lambda (speed and batch views). The Kappa architecture is more suitable for cases where permanent data storage is not necessary.

Table II presents a brief comparison of the features of the Lambda and Kappa architectures.

The Lambda and Kappa architectures provide both real-time and historical data analysis in a single environment. Lambda uses two separate technology stacks to manage batch and stream processing. However, Kappa offers the possibility of building streaming and batch processing system based on the same technology. This is one major advantage of the Kappa architecture compared to the Lambda architecture. On the other hand, the Lambda architecture wins over Kappa when it comes to storing very large data sets (in terabyte range) and which can be processed more efficiently in Hadoop for large-scale historical analysis.
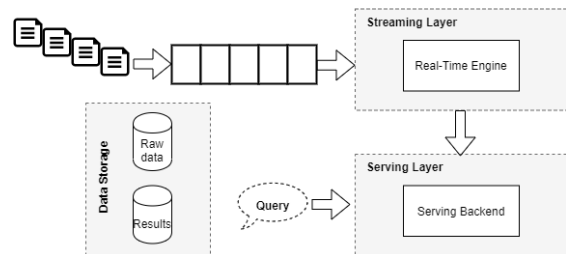
Fig. 6.   Kappa Architecture.

TABLE II.   COMPARISON OF THE DIFFERENT FEATURES OF LAMBDA AND
KAPPA ARCHITECTURES

	Layers	Processing	Basic technologies	Scalability	Real-time
Lambda	Batch, serving and real-time layer	Batch and real-time	Immutable	Yes	Isn't accurate
Kappa	Stream processing and serving layer	Real-time	Immutable	Yes	Accurate

V. OUR PROPOSED BIG DATA ANALYTICS ARCHITECTURE FOR REAL-TIME STUDENT ATTENTION DETECTION

Traditional Computer Vision (CV) systems have limitations in terms of reliability, scalability and fault tolerance. In fact, in this type of system, the CV processing unit collects and processes the data at the same time. Thus, in case of failure, the data of the video stream will be lost and the processing will be interrupted. Detecting a node failure and switching the processing to another node will fragment the data and increase the error rate on the output result. In order to ensure a reliable and efficient data processing of a large-scale video stream, it is necessary to have a highly scalable, fault-tolerant and loosely coupled distributed system. Video stream analysis requires large-scale parallel processing, fast extraction of features from each frame, deployment of multiple machine-learning libraries, and returns processing results in different formats. In our system, we propose a Big Data architecture that best meets the requirements mentioned above. It will be based on the Kappa architecture principle for real-time processing of the video stream and will also allow permanent storage of this data. As shown in Fig. 8, it has five main layers, the data producer, data ingestion, flow processing, storage and presentation.

A. Design and Implementation

*1) Data collection:* The video stream is produced by a cluster of a high definition IP (HD 1080P: 1920 x 1080 (16/9) video at 30Fps) cameras designed to provide real-time video streams. The volume of data video is so important that it can easily reach the terabyte scale.

The cameras that make up the cluster provide data with precise specifications such as codec, resolution or the number of frames per second. For this, each camera has an ID that identifies it and keeps a mapping between the camera and its specifications. In this way, the stream processor can easily create images and video sequences from the video stream.

2) Data ingestion: For large scale stream acquisition, we choose to use Apache Kafka, that is, basically an open-source messaging system consisting of three components: Message Producer, Message Broker and Message Consumer. Kafka exchanges data between applications. Using the OpenCV library, we convert the video stream coming from cameras to images. Each image is stored in a JSON object. The content of this object consists of six fields: Id(ID of the camera), cols(number of columns), rows(number of rows in a frame), type (are OpenCV Mat-specific details), data(is a base-64

encoded string for the byte array of the frame), and the timestamp (the time at which the frame was generated). The JSON object obtained is a record which will constitute the body of a message to send to Kafka broker. Given the size of the frame that will be transported in the JSON objects, compression is needed to reduce the transfer time and ensure real-time processing. The compressed message is then transmitted to the producer, then to the Kafka Broker who is responsible for delivering it to consumers (stream processing unit).

3) Stream processing: Stream processing is built on Apache Storm, a distributed computing infrastructure for data stream processing and allows programming in python. Storm does not have an API to handle the low-level image processing. To overcome this, we use the OpenCV library which provides low-level image processing services on the top of Storm. As shown in Fig. 7, the stream processing unit consists of three main layers: The image pre-processing, mainly responsible for performing basic image processing tasks such as restoration (JSON to frame), encoding images/video, image enhancement, grayscale images, etc. The features extraction is responsible for extracting the features required to the analysis. These features will be provided as input to the third layer which is the distributed data mining. To implement this layer, we will use the Apache SAMOA a platform for mining Big Data streams [20]. APACHE SAMOA is both a framework and a library. As a framework, it allows the algorithm developer to abstract from the underlying execution engine, and therefore, reuse their code on different engines such as Storm, Flink, Samza, and Apex [21]. As a library, Apache SAMOA contains implementations of state-of-the-art algorithms for distributed machine learning on streams. For classification, Apache SAMOA provides a Vertical Hoeffding Tree (VHT), a distributed streaming version of a decision tree. For clustering, it includes an algorithm based on CluStream. For regression, HAMR, a distributed implementation of Adaptive Model Rules. The library also includes meta-algorithms such as bagging and boosting [21][22].
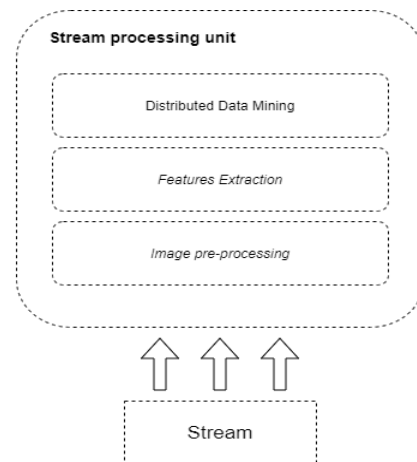

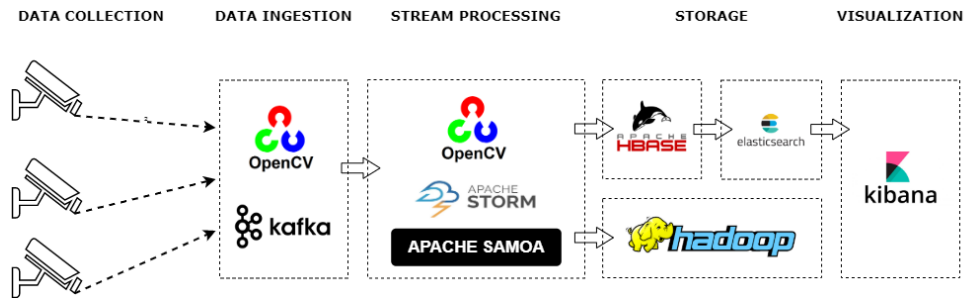Fig. 7.   The Stream Processing unit's Layers.

Fig. 8. Video Stream Analytics System Architecture.

4) Storage: The size of the video streams coming from multiple cameras installed in the different classrooms of a university can easily reach the Exabyte. So, scaling up, must be supported. Hence, storing such amounts of data is a real challenge. To solve this issue, video streams received from different cameras, through Kafka, encapsulated in JSON objects, are saved as MP4 files on the Hadoop Distributed File System storage (HDFS) of the Hadoop cluster. The storage will be distributed to ensure a fast flow recording. This also will ensure high data availability and the ability to perform distributed processing. Metadata for video streams and the result of video analysis are stored in an Hbase database, a data result duplication will be done in Elasticsearch for a possible visualization.

The received video streams are stored as 90 seconds video files. The size of a video file is strongly related to its length and it impacts the speed of storage, transmission, and analysis. The duration of a 90-second video file is chosen after taking into account the intra-cluster bandwidth, processing, machine performance, and fault tolerance. A small video file is transmitted faster than a large file. Furthermore, in case of failure, retransmission is faster.

The average size of a 90 second video stream is 16.2 GB, considering the average frame size which is 6M. Single-camera recording video (equivalent to 8 hours) requires a capacity of 5,184 TB. Table III summarizes the storage capacity required to store video stream for different durations.

5) Presentation: Visualization of analysis results is provided by Kibana, the official GUI visualization platform for Elasticsearch. Kibana has 4 main components: Discover, visualize, dashboards, and management. It provides access to the data stored in the Elasticsearch cluster through a Web interface that supports search, visualizes and analyzes functionalities.

TABLE III. DISK SPACE REQUIREMENTS FOR ONE WEEK OF RECORDED VIDEO STREAM

Duration	Size of a frame	Size of a video
90 seconds	6MB	16,2 GB
5 minutes	6MB	54GB
1 Hour	6MB	648GB
1 Day	6MB	5,184 TB
1 Week	6MB	31,104TB

VI. DISCUSSION

We have drawn up a comparison of the two most known architectures in Section IV. We discuss here the criteria for choosing our architecture and to what extent it meets the requirements of our problem.

The purpose of this study is to set up a real-time Big Data architecture capable of responding to the constraints of detecting and monitoring the student's attention at large scale. In other words, it must handle and process continuously and in real-time a large stream of images coming from the HD cameras placed in the classrooms. The video stream must also be stored for later analysis, to serve as a support for creating attention detection datasets or for training and validation tasks.

Our architecture is inspired by the concept of Kappa architecture since we only involve the processing of image data streams. The high capabilities of Apache Kafka make it possible to handle large incoming data streams. To overcome the real-time processing problem, Storm is used to pre-process the image streams, as well as the features extraction tasks. This is guaranteed, as Storm can be plugged-in as a pre-processor of the stream for each Kafka topic. In order to achieve real-time processing, the Storm topology must be created inside of it. For this purpose, we have coupled the use of Storm with SAMOA to benefit from its distributed machine learning libraries. The business logic of our architecture implies the need to store the data received for possible uses. Storage is provided by Hadoop with the possibility of processing which gives our architecture a hybrid form and makes it benefit from the advantages of the two traditional architectures.

The constraints regarding scalability and reliability were respected. Our architecture can scale-out on-demand to adapt to the increased load resulting from an expansion of university establishments and the number of students. Storm and Kafka clusters can scale horizontally. In addition, in case of machine failures, the data replication mechanisms allow fault tolerance, without loss of data.

VII. CONCLUSION

The main concern of this paper was to study the different technologies and architectures of Big data to propose an architecture capable of dealing with the problems linked to the high scale transition for student's attention detection systems. We compared the two main architectures on the market, namely, Lambda and Kappa, in order to choose the one that best meets our needs. Given the constraint of real-time

processing, we chose to base our architecture on the Kappa architecture which is simple and requires only a single implementation of the business logic. The choice of Apache Storm was not arbitrary but deduced from a comparison of stream processing tools. Storm is real-time processing with no restrictions on the implementation language. Furthermore, to implement machine learning algorithms in a distributed way, Storm can interface perfectly with Apache SAMOA.

The main limitation of this study is the lack of real-world data, which we take into consideration in future research. The next step of our work is to create a data set for the student's attention detection and to analyze it with several machine learning algorithms. Finally, we intend to perform the test and performance evaluation of our architecture.

REFERENCES

[1] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," IEEE Trans. Affect. Comput., vol. 5, no. 1, pp. 86–98, 2014.

[2] Krithika L.B and Lakshmi Priya GG, "Student Emotion Recognition System (SERS) for e-learning Improvement Based on Learner Concentration Metric," Procedia Comput. Sci., vol. 85, pp. 767–776, 2016.

[3] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features," EURASIP J. Image Video Process., vol. 2017, no. 1, p. 80, Dec. 2017.

[4] N. Vettivel, N. Jeyaratnam, V. Ravindran, S. Sumathipala, and S. Amarakecrthi, "System for Detecting Student Attention Pertaining and Alerting," in 2018 3rd International Conference on Information Technology Research (ICITR), 2018, pp. 1–6.

[5] P. Goldberg et al., "Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction," Educ. Psychol. Rev., Dec. 2019.

[6] J. Gantz and D. Reinsel, "Extracting value from chaos," IDC iview, vol. 1142, no. 2011, pp. 1–12, 2011.

[7] S. Ali, A. Rauf, and J. Ahmad, "Protecting Unauthorized Big Data Analysis using Attribute (Data) Relationship."

[8] G. Glossary, "Big Data defintion." 2013.

[9] P. Zikopoulos, C. Eaton, and others, Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[10] S. Madden, "From databases to big data," IEEE Internet Comput., vol. 16, no. 3, pp. 4–6, 2012.

[11] S. Sagiroglu and D. Sinanc, "Big data: A review," in 2013 international conference on collaboration technologies and systems (CTS), 2013, pp. 42–47.

[12] M. A. Lopez, A. G. P. Lobato, and O. C. M. B. Duarte, "A Performance Comparison of Open-Source Stream Processing Platforms," in 2016 IEEE Global Communications Conference (GLOBECOM), 2016, pp. 1–6.

[13] A. Fundation, "storm.apache.org," 1394. [Online]. Available: https://storm.apache.org.

[14] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," J. Big Data, vol. 2, no. 1, p. 24, 2015.

[15] I. Bartolini and M. Patella, "Real-Time Stream Processing in Social Networks with RAM3S," Futur. Internet, vol. 11, no. 12, p. 249, 2019.

[16] S. Shahrivari, "Beyond batch processing: towards real-time and streaming big data," Computers, vol. 3, no. 4, pp. 117–129, 2014.

[17] N. Marz and J. Warren, Big Data: Principles and best practices of scalable real-time data systems. New York; Manning Publications Co., 2015.

[18] V. Persico, A. Pescapé, A. Picariello, and G. Sperl\'\i, "Benchmarking big data architectures for social networks data processing using public cloud platforms," Futur. Gener. Comput. Syst., vol. 89, pp. 98–109, 2018.

[19] J. Kreps, "Questioning the lambda architecture," Online Artic. July, p. 205, 2014.

[20] G. De Francisci Morales, "SAMOA: A platform for mining big data streams," in Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 777–778.

[21] N. Kourtellis, G. D. F. Morales, and A. Bifet, "Large-scale learning from data streams with Apache SAMOA," in Learning from Data Streams in Evolving Environments, Springer, 2019, pp. 177–207.

[22] T. Vasiloudis, F. Beligianni, and G. De Francisci Morales, "BoostVHT: Boosting distributed streaming decision trees," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 899–908.

Analysis of K-means, DBSCAN and OPTICS Cluster Algorithms on Al-Quran Verses

Mohammed A. Ahmed[1], Hanif Baharin[2], Puteri N.E. Nohuddin[3]
Institute of IR 4.0, Universiti Kebangsaan Malaysia
Bangi, Selangor, 43600, Malaysia

Abstract—**Chapter Al-Baqarah is the longest chapter in the Holy Quran, and it covers various topics. Al-Quran is the primary text of Islamic faith and practice. Millions of Muslims worldwide use Al - Quran as their reference book, and it, therefore, helps Muslims and Islamic scholars as guidance of the law life. Text clustering (unsupervised learning) is a process of separation that has to be divided text into the same section of similar documents. There are many text clustering algorithms and techniques used to make clusters, such as partitioning and density-based methods. In this paper, k-means preferred as a partitioning method and DBSCAN, OPTICS as a density-based method. This study aims to investigate and find which algorithm produced as the best accurate performance cluster for Al-Baqarah's English Tafseer chapter. Data preprocessing and feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF) have applied for the dataset. The result shows k-means outperformed even has the smallest of Silhouette Coefficient (SC) score compared to others due to less implementation time with no noise production for seven clusters of Al-Baqarah chapter. OPTICS has no noise with the medium of SC score but has the longest implementation time due to its complexity.**

Keywords—K-means; DBSCAN; OPTICS; Al-Baqarah clustering; Silhouette Coefficient; Tafseer; text clustering

I. INTRODUCTION

The Quran is a significant religious text written in Quranic Arabic, followed by believers of the Islamic faith. The Quran means "perfect reading," in terms of language, which Muslims believed to be revealed to people as a guide in all aspects of life. Al - Quran is the message of Allah that was revealed and spread to the prophet Mohammed From the Prophet SAW 's time until now protected by Allah. Al-Quran wrote in Arabic but has translated into numerous languages around the world, as well as English. It is text data that may be further analyzed. Chapter Al-Baqarah is Al-Quran's longest chapter, so there are various themes in Al-Baqarah's chapter. Such themes are not written sequentially but depend on asbabunnuzul ayat (verses) while revealed. Hence, grouping verses of similar characteristics of a text they compose will form a cluster that could reflect any theme in Surah (chapter) Al–Baqarah [1].

Data mining began in the 1980s, made a lot of progress in the 1990s, and is growing in the early 2000s. Data mining can transform large-scale data set into knowledge to help meet significant issues; it can meet this need by providing data knowledge-based tools [2]. Discovering useful information from groups of text documents is known as text mining. Text mining has a meaningful effect on different applications, for example, social media data, opinion mining, and recommendation systems. Text mining is a common approach to uncover meaningful information from text collections, including clusters, outliers, and the evolution of clusters. The lack of ground-truths in real-world samples creates a demand to perform such analyzes in an unsupervised context [3].

Text clustering (unsupervised learning) is the process for the mining of text, which divides the similar text of documents into groups or clusters. It is typically done by finding patterns and trends through the use of several text manipulation techniques and specific algorithms. Text clustering is a part of data mining. Documentation loaded in the vector weight term become cluster objects. Therefore, there are many clustering algorithms and techniques used to make clusters, such as the partitioning method (k-means and k-medoid) [4], hierarchical method (Agglomerative and Divisive) [5], density-based method (DBSCAN and OPTICS) [6], and grid-based method (STING and CLIQUE) [2].

Most of the articles adapted one method to cluster the text translated Al-Quran (Tafseer) as the dataset for its experiments, such as [5] [7-10]. This paper has adapted two clustering methods (partitioning and density-based) and compared and analyzed Al-Baqarah's chapter text of English Tafseer as the experiment dataset. K-means preferred as a partitioning method and DBSCAN, OPTICS as a density-based method. This study aims to investigate and find which algorithm produced the best performance cluster for Al-Baqarah's chapter.

The rest of this paper is structured as follows: Section 2 discusses the related work of this research; Section 3 discusses the research methodology. Section 4 describes the experimental procedure and results. Finally, Section 5 presents the conclusions.

II. RELATED WORK

There are many papers related to this article. The clustering experiment of [1] utilizes a mixture of k-means clustering techniques, k-medoid, and bisecting k-means, together with Jaccard similarity, correlation coefficients, and cosine similarity produce different validity values. The ideal cluster results, however, in chapter al - Baqarah clustering process given by cosine similarity of k-medoid. This research [11] tried to group Hadith texts of Indonesian text language to compare Fuzzy c-means and k-means algorithms with determined parameters and experiments. F-measure and Silhouette Coefficient calculations used as measure calculations to

evaluate clusters performances. Findings demonstrate that the Fuzzy c-means algorithm is more useful to group the hadith text based on consistency with the chapter and original data. Additionally, this paper suggested using DBSCAN and compare with k-means as future work, and this is one of the motivations to do the investigation related to this article.

This study [7] resulted in an initial practical move to understand the concepts of verses in the Holy Quran. The algorithm used to cluster 6236 total verses, using partitioning (k- means) for unsteamed, steamed words that formed three clusters. The paper [5] uses a hybrid of TF-IDF (Term Frequency-Inverse Document Frequency) and Network Analysis (map) approach to extract keywords and identify relationships between keywords and Tafseer chapters. Six short chapters of 130 keywords are taken from Malay translated Tafseer of Al-Quran. The proposed method was called KCRA. Reference [12] developed a semantic-based question answering (QAS) for Indonesian Quran translation, which asked the users three questions, then created a TF-IDE for each term belonging to the respective expected response category (also called entity group) to feed or provide a semantic interpreter on user request. The author has organized 222 ontology principles into 6, 24, and 77 of Time, Position, and Individual concepts, respectively. The research [10] aims to develop a web-based verse search (information retrieval) system for Al-Quran, which is integrated with the clustering algorithm (SPC) to facilitate Muslim discovery of relevant information in the Quran verse by grouping the Quran verse within their its similar group. This paper [13] discusses a study on generating weighted vectors for each concept in Indonesian Quran Translation (ITQ) and applying QAS such as [12]. Still, here the author provides more information on TF-IDE findings and has different work procedures.

In [14], various text clustering algorithms were studied. The main objective of this study is comparing different clustering algorithms and finding out which algorithm is most suitable for users using WEKA free open source software, with the advantages and disadvantages of each algorithm like DBSCAN and k-means. The author founds that the k-means clustering algorithm is the most straightforward compared to other algorithms. The author in [15] suggested an algorithm to estimate the optimum value of ϵ-neighbourhood and *MinPts* for the DBSCSN algorithm, based or used k-means algorithm.

The author in [16] has applied an OPTICS clustering on text data and provided valuable insight into the operation of OPTICS and its applicability to text information. The SCI algorithm introduced in this paper to create clusters from the OPTICS plot can be used as a benchmark to check OPTICS efficiency based on measurements of purity and coverage. The author in [17] suggested an ICA incremental clustering algorithm based on the OPTICS. Like OPTICS, the ICA also generates a dataset's cluster-ordering structure. The ICA is, however, smarter because no parameters are needed, which are very difficult to define for users who do not know the properties of datasets and delete some of these complex definitions. ICA is ideal for processing dynamic datasets. The author in [18] introduced a new similarity measure for sequential data and suggested an improved density-based clustering technique to find meaningful clusters in different web databases. Experiments compared DBSCAN clustering characteristics, OPTICS algorithm with the new advanced SSM-DBSCAN algorithm, and SSM-OPTICS on web sessions thought various similarity indicators such as Euclidean, Jaccard, Fuzzy, and Cosine.

III. RESEARCH METHODOLOGY

A. Text Preprocessing

Before the clustering algorithm is applied, the standard preprocessing procedures are used to preprocess text documents, which includes:

- Tokenization: Text data is divided into the basic sequence of independent units.

- POS tagging: The process of marking up a word in a text (corpus) as corresponding to a particular part-of-speech.

- Case folding or normalization: The process of converting all the characters in a document into the same case, either all upper case or lower.

- Stop word removal: Deleting particular common words that happened most often, such as 'is',' are',' that', 'an'….

- Stemming: Convert verbs to their origins for convenience.

- Term weighting steps: The texts are processed in numerical format or matrix in the preprocessing steps.

B. Feature Extraction

Term weighting used to extract terms or features after the preprocessing completed. Term weighting aims to convert text data into a numeric format. The literature contains many schemes for term weighting. For text document representation, the vector space model (VSM) calculates the term frequency-inverse document frequency scheme (TF-IDF) [19]. The VSM shows each document with a vector and weighs the cell values as follows:

$$d_i = \{F_{i,1}, F_{i,2}, \dots, F_{i,j}, \dots, F_{i,t}\} \tag{1}$$

where t refers to the features number and F_{ij} refers to the feature j weight in document i [20]. The following weighting scheme is used to measure the feature's weighting:

$$F_{i,j} = TF - IDF(i,j) = TF(i,j) \times \left(\log \frac{d}{DF(j)}\right) \tag{2}$$

where $TF_{(i,j)}$ the feature j frequency in document i, and $DF(j)$ is all the documents containing feature j. Matrix of size $m \times n$ is used to represent the VSM as follows:

$$VSM = \begin{pmatrix} F_{1,1} & F_{1,2} & F_{1,(t-1)} & F_{1,t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots \\ F_{(m-1),1} & \cdots & \cdots & F_{(m-1),t} \\ F_{m,1} & F_{m,2} & \cdots & F_{m,t} \end{pmatrix} \tag{3}$$

C. Cluster Techniques

This study adopted the three most popular clustering algorithms, described as follows:

1) k-means: The k-means algorithm is the most common clustering algorithm. It is a popular clustering technique introduced 50 years ago. Due to its simplicity, the k-means algorithm was commonly used to handle huge datasets. It is also easily implemented, enjoying low computational complexity and rapid convergence [21]. The process includes two main iterative steps. This process classifies the whole dataset into heterogeneous clusters. Across the years, many versions of this algorithm were developed to enhance its performance, such as k-medoids [22], kernel k-means [21], and k-harmonic-means [23].

k-means is the simplest and most fundamental version of partitioning cluster analysis. The k-means algorithm defines a cluster's centroid as the cluster's mean point value as follows. First, it selects a random kl of Dc items, each initially representing a mean or centre cluster. The k-means algorithm procedure is summarized as follows [2]:

Algorithm: k-means

Input: Dc: a dataset comprising nu items,
kl: the number of clusters.
Output: A set of kl clusters.

Steps:
Step1: arbitrarily select kl items from Dc as the initial cluster centres;
Step2: **repeat**
Step3:   re)assign each item to the cluster with which the item is most related to the basis that the items in the cluster has such a mean value;
Step4:   Updating a clusters' means, requires computing the mean value of each cluster item;
Step5: **until** there is no change

---

2) DBSCAN: Hierarchical and partitioning clustering approaches are useful for spherical-shaped of clusters only. Density cluster methods are designed to solve the problem of locating arbitrary shape clusters like "S" shape and oval clusters. Such data would likely misidentify convex regions where noise or outliers are included in clusters, and this is the principal strategy behind clustering approaches based on density, which can find non-spherical form clusters. Density-Based Spatial Clustering of Applications with Noise referred to (DBSCAN) identifies central objects that have neighbourhoods of dense. It links centre objects and their neighbourhoods to construct dense areas as clusters. This algorithm requires two user-specified parameters as inputs, which are ϵ and MinPts [2]. The DBSCAN algorithm procedure is summarized as follows [2]:

Algorithm: DBSCAN

Input: *Data* = a data set including *nu* objects, *MinPts* = the threshold of neighbourhood density, ϵ = the radius parameter.
Output: A establishment of density-based clusters.

Steps:
Step1: sign all objects as **unchecked**;
Step2:**do**
Step3:   at random choose an **unchecked** object *b*;
Step4:   sign *b* as **checked**;
Step5:   **if** the ϵ-neighbourhood of *b* includes at least *MinPts*
Step6:   establish a new cluster *Cl*, then add together *b* to *Cl*;
Step7:   suppose *M* be the set of objects appearing in the ϵ-neighborhood of *b*;
Step8:   **for** every point *b'* in *M*
Step9:   **if** *b'* is equal **unchecked**
Step10:   sign *b'* as **check**;
Step11:   **if** the ϵ-neighbourhood of *b'* includes at least *MinPts* points, add the *M* points;
Step12:   **if** *b'* isn't a cluster member yet, add *b'* to *Cl*;
Step13:   **ending for**
Step14:   output *Cl;*
Step15:   **else** sign *b* as **noise;**
Step16:**until** no object is **unchecked**;

---

3) OPTICS: This method proposed to remove the problem of using a set of required variables in clustering analysis. Ordering Points To Identify the Clustering Structure referred to OPTICS, explicitly does not produce a dataset clustering but produces an ordering cluster. It is a continuous array of all objects within analysis which describes data clustering structure dependent on density, then the objects in a denser cluster listed in the ordering of the cluster. This order refers to the density-based clustering acquired from a large variety of parameters. OPTICS, therefore, does not require the users to give a certain density threshold. The cluster order can be applied to obtain necessary clustering data (e.g., arbitrary-shaped clusters or cluster centres), to derive and display the clustering composition. The basic idea is to define a specific database cluster and noise like DBSCAN. OPTICS identifies the cluster based on density [2].

OPTICS requires two essential aspects per item. Firstly, the (core distance) of item **b** is the smallest value ϵ' so that ϵ'-neighbourhood has at least MinPts items. That's the minimum distance threshold, which makes **b** a core item is ϵ'. If b isn't a core item, the core distance of **b** is indeterminate. Secondly, the minimum radius value (reachability-distance) to item **b** from **c** makes **b** density-reachable from **c**. Under the concept of density-reachability, **c** must be a core point, and b must be in **b**'s neighbourhood. And hence, the reachability distance is max{core-distance(**b**), dist(**b**, **c**)} from **c** to **b**. If **c** is not a core item, the reachability-distance from **c** to b is unspecified.

Several core items can directly reach an item **b**. Accordingly, **b** could have numerous reach-distances for different core items. The smallest reachability-distance of **b** is special importance as it provides the shortest route through which **b** links to a dense cluster [2].

D. Cluster Evaluation

Cluster validation or evaluation of the effects of clustering is useful for measuring the accuracy of the clustering or grouping. Many cluster validation methods can be used. If the ground truth of a dataset is not available such as this paper's experiments, an intrinsic method must be used to evaluate clustering accuracy. Generally, Intrinsic approaches measure clusters by evaluating well how-isolated clusters are and how dense clusters are. Most intrinsic approaches benefit a metric similarity between items in the dataset, such as the Silhouette Coefficient [2] [24].

1) Silhouette Coefficient (SC): A higher Silhouette Coefficient score refers to a more structured cluster model. The Silhouette Coefficient is defined and consists of two scores for each sample. If x is the mean distance between a sample and all other points in the same class, and y is the mean distance between a sample and all other points in the next closest cluster. Then, the Silhouette Coefficient sc for a single sample is given as [24]:

$$sc = \frac{y-x}{max(x,y)} \tag{4}$$

The score is set to -1 for inappropriate clustering and +1 for very dense clustering. Zero scores show overlapping clusters. The scoring is higher when the clusters are dense and well separated, which is a standard cluster concept [24] [2] [11].

2) Execution time: The algorithms execution time of this study is limited or related to use the hardware platforms of Intel Core i7-8550U CPU with @ 1.80 GHz and RAM of 8 GB and software platforms of MS Windows 10, Python 3.7.7.

IV. EXPERIMENTS AND RESULT DISCUSSIONS

Fig. 1 describes the flowchart experiments of this study, the first part of this section discusses the source of the experiment's dataset, the statistics result before and after text prepossessing, output terms after the feature extraction process, and the parameters used to implement the algorithms. In contrast, the second part of this section discusses the results of three cluster algorithms and which is the optimal cluster algorithm for the experiment's dataset.

A. The Experiments

1) Used dataset: http:/tanzil.net/trans/ a website provides documentation about the Quran translation in various languages, for many translators from various Tafseer. Many such researchers used this website to collect data and adapted for its experiments [5] [7] [8] [13] [25-27]. Al-Baqarah chapter used in this research that consisted of 286 verses written by (Ahmed Ali) of English Tafseer Al-Quran, the text document contains 286 lines, each line represents Tafseer of one verse. The total words of this document are 11478, including all terms, names, numbers, symbols, and marks.

2) The preprocessing: Preprocessing of data was done in many stages, namely the document reading of 11478 words, tokenizing these words to become a sequence of independent units, stop word removal and normalize from becoming 1508 terms of features, and stemming from becoming 1221 features. Now the data is ready for the next step.

3) Features Extraction: After the text is preprocessed, the dictionary or corpus containing all the words in the document is assembled. Therefore, weight is calculated (generated) for each document feature (time weighting), or (equivalent weighting). TF-IDF used to assign a weight to each term or feature. All these weights will combine in a matrix to construct the corpus ready for the clustering process. Fig. 2 shows the bar chart for the most first 15 features of the Al-Baqarah chapter.
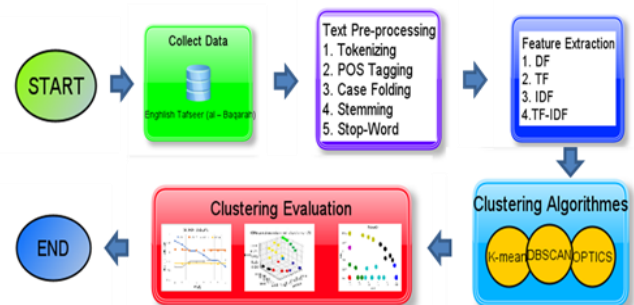
Fig 1.   Flowchart of the Study Experiments.
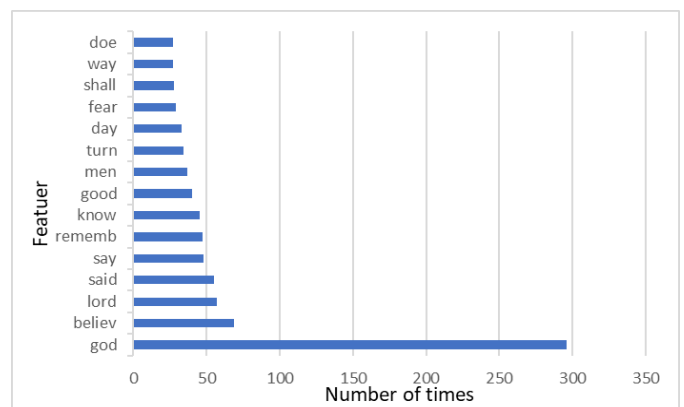
Fig 2.   Features of Al-Baqarah Chapter.

4) Cluster algorithm parameters: The algorithm k-means clustering adapted from the partitioning method for this study. The value of k needs to be determined before implementation, which should be more than one. Based on the experiments of Choiruddin [28] and Huda [1], the optimal value of k to cluster the Al-Baqarah chapter using k-means is seven. Chapter Al-Baqarah has 53 subjects, although some still have the same subject as others. The verses with the same subject are then grouped into seven major themes or topics.

Based on the above and in this study experiments, the values MinPts and ϵ must be 7, 0.2, respectively to make the DBSCAN algorithm produced seven clusters with some noise, while the value of MinPts should be 9 for seven cluster output of the OPTICS algorithm.

B. Results and Discussion

The finding and discussion of this study can be summarized as follows:

- Depends on [1] and [28], k=7 for k-means to cluster chapter Al-Baqarah.

- Fig. 3 shows the visual clustering for the two/three most common features ('god', 'believ', and 'lord') applied by the k-means algorithm. These legends (seven cluster colors) represent the cluster for the two/three features.
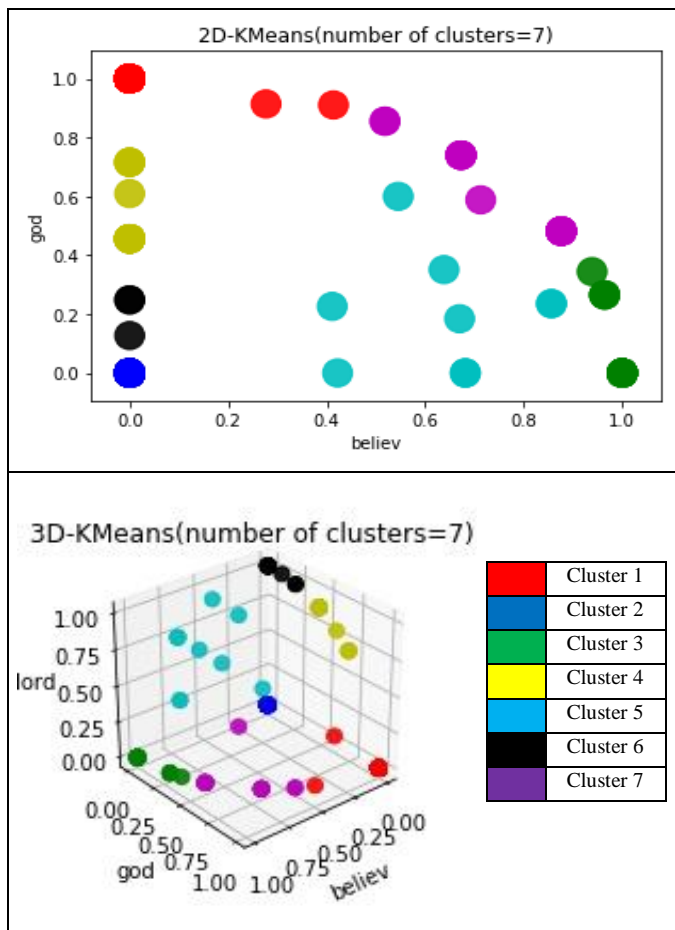
Fig 3.   Cluster Visualization for the 2D/3D k-Means Algorithm.

- Fig. 4 gives the SC results for each cluster (seven) of the k-means, where the average for these results is equal to 0.8948 with 0.319 seconds of implementation time.
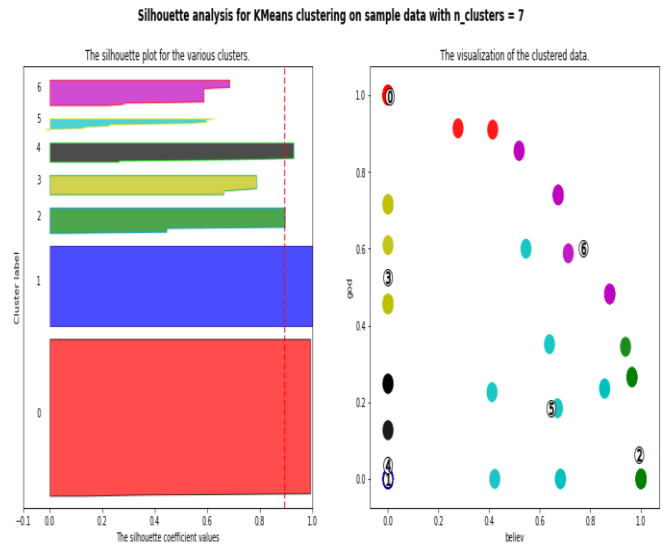
Fig 4.   Silhouette Coefficient for Each Cluster with the Average for the k-Means Algorithm.

- Fig. 5 visuals the output of the OPTICS cluster algorithm for the two most common features ('god', 'believ'), where $MinPts = 9$ to make OPTICS produced seven clusters with no need to calculate ϵ due to algorithm property. The SC is equal to 0.9098, with 0.935 seconds of implementation time.
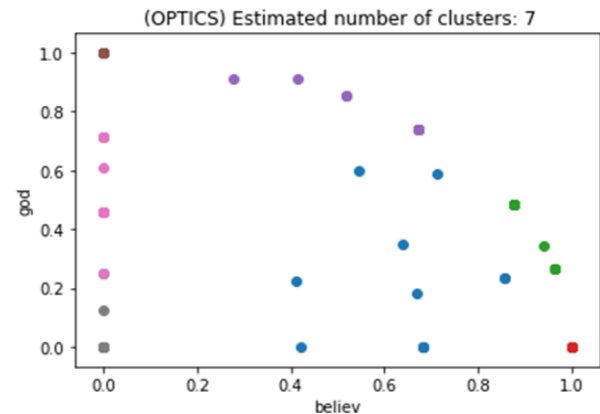
Fig 5.   Cluster Visualization for OPTICS Algorithm.

- Fig. 6(A) shows the visual clustering for the two most common features ('god', 'believ') applied by the DBSCAN algorithm with some noise of black circles. From this study experiments, $MinPts = 7$, and $\epsilon = 0.2$ to make DBSCAN produce seven clusters (Fig. 6(B) describes this cluster production number with regards to $MinPts$ and ϵ). DBSCAN produces some noises, Fig. 6(C) illustrates these noises with regards to $MinPts$ and ϵ. Fig. 6(D) shows the SC results obtained from this study trials. The SC is equal to 0.9129 with 0.327 second implementation time if DBSCAN produced the seven clusters.

(A)

(B)

(C)

(D)
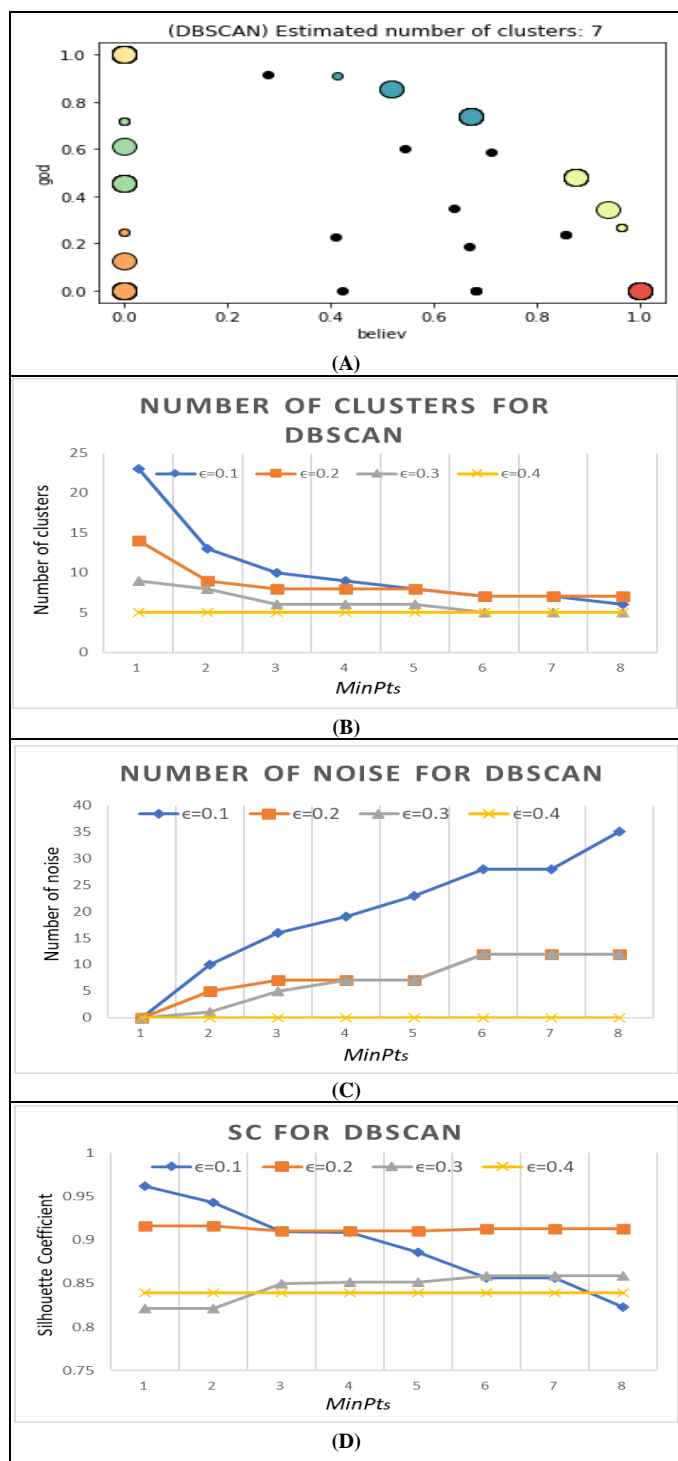
Fig 6.   Cluster Visualization for the DBSCAN Algorithm.

- Table I and Fig. 7 illustrate the analysis between these three cluster algorithms regards to (time, SC, number of noises, *Minpts,* and є). From these results, k-means outperforms even has the smallest value of SC compared to others due to less of implementation time with no noise production for seven cluster of Al-Baqarah chapter. OPTICS has no noise with the

medium of SC value but has the longest implementation time due to its complexity.

TABLE I.   CLUSTERING ANALYSIS BETWEEN THE THREE ALGORITHMS FOR SEVEN CLUSTERS

Matrices	K-means	DBSCAN	OPTICS
Time (seconds)	0.319	0.327	0.935
Silhouette Coefficient (SC)	0.8948	0.9129	0.9098
Number of noise	-	12	-
MinPts	-	7	9
є	-	0.2	-
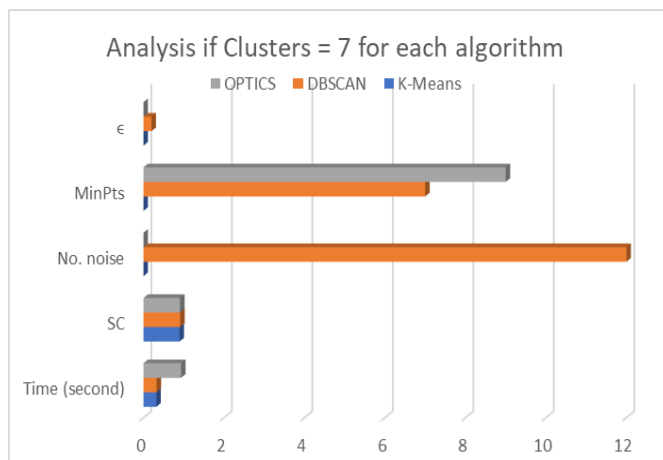
Fig 7.   Clustering Analysis between the Three Algorithms for Seven Clusters.

V. CONCLUSIONS

The clustering experiments of this study adapted k-means, DBSCAN, and OPTICS text cluster algorithms to cluster English Tafseer of Al-Baqarah chapter to a seven cluster. Chapter Al-Baqarah has 53 subjects, although some still have the same subject as others. The verses with the same subject are then grouped into seven major themes or topics. This study aims to investigate and figure out which algorithm was the best performance cluster for Al-Baqarah's chapter. The visual and statistics result proves k-means outperforms even has the lowest score of SC compared to others due to less of implementation time with no noise production for seven cluster of Al-Baqarah chapter. DBSCAN has the highest SC score but produced some noises. OPTICS has no noise with the medium of SC value but has the longest implementation time due to its complexity.

This research contributed to researchers and Muslims because the Al-Quran is the reference book. Thus, it is beneficial for Muslims in general, and this research investigated document clustering using many available current techniques to achieve high accuracy and performance. For future work, the authors hope to implement more text cluster algorithms such as the hierarchical method (Agglomerative) and grid-based method (STING) and make more comparisons.

REFERENCES

[1]   F. Huda, M. R. Deyana, Q. U. Safitri, W. Darmalaksana, U. Rahmani, and Mahmud, "Analysis Partition Clustering and Similarity Measure on

Al-Quran Verses," in 2019 IEEE 5th International Conference on Wireless and Telematics (ICWT), 2019, pp. 1–5.

[2] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques, 3rd ed. Elsevier, 2012.

[3] W. A. Mohotti, "Unsupervised text mining: Effective similarity calculation with ranking and matrix factorization," Queensland University of Technology, 2020.

[4] C. Luo, Y. Li, and S. M. Chung, "Text document clustering based on neighbors," Data Knowl. Eng., vol. 68, no. 11, pp. 1271–1288, 2009.

[5] S. Chua and P. N. E. Nohuddin, "Relationship Analysis of Keyword and Chapter in Malay-Translated Tafseer of Al-Quran," J. Telecommun. Electron. Comput. Eng., vol. 9, no. 2–10, pp. 185–189, 2017.

[6] Indah, N. G. N. Reza,K. Rice, B. V. R. S. Okta, A. Suwanto, S. W. P. N. Tuti, R. Yulia, and Robbi, "DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru," in Journal of Physics: Conference Series, 2019, vol. 1363, no. 1, p. 12001.

[7] C. Slamet, A. Rahman, M. A. Ramdhani, and W. Darmalaksana, "Clustering the verses of the Holy Qur'an using K-means algorithm," Asian J. Inf. Technol., vol. 15, no. 24, pp. 5159–5162, 2016.

[8] S. Chua and P. N. E. binti Nohuddin, "Frequent pattern extraction in the Tafseer of Al-Quran," in The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M), 2014, pp. 1–5, doi: 10.1109/ICT4M.2014.7020667.

[9] B. Hamoud and E. Atwell, "Quran question and answer corpus for data mining with WEKA," in 2016 Conference of Basic Sciences and Engineering Studies (SGCAC), 2016, pp. 211–216.

[10] Z. Indra, A. Adnan, and R. Salambue, "A Hybrid Information Retrieval for Indonesian Translation of Quran by Using Single Pass Clustering Algorithm," in 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019, pp. 1–5.

[11] R. S. Pratama, A. F. Huda, A. Wahana, W. Darmalaksana, Q. U. Safitri, and A. Rahman, "Analysis of Fuzzy C-Means Algorithm on Indonesian Translation of Hadits Text," in 2019 IEEE 5th International Conference on Wireless and Telematics (ICWT), 2019, pp. 1–5.

[12] S. J. Putra, R. H. Gusmita, K. Hulliyah, and H. T. Sukmana, "A semantic-based question answering system for indonesian translation of Quran," in Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services, 2016, pp. 504–507, doi: 10.1145/3011141.3011219.

[13] S. J. Putra, K. Hulliyah, N. Hakiem, R. P. Iswara, and A. F. Firmansyah, "Generating weighted vector for concepts in indonesian translation of Quran," in Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services, 2016, pp. 293–297.

[14] N. Sharma, A. Bajpai, and M. R. Litoriya, "Comparison the various clustering algorithms of WEKA tools," facilities, vol. 4, no. 7, pp. 78–80, 2012.

[15] C. Qi, L. Jianfeng, and Z. Hao, "A text mining model based on improved density clustering algorithm," in 2013 IEEE 4th International Conference on Electronics Information and Emergency Communication, 2013, pp. 337–339.

[16] P. Deepak and S. Roy, "Optics on text data: Experiments and test results," IBM India Res. Lab, 2006.

[17] J.-S. Fu, Y. Liu, and H.-C. Chao, "ICA: An incremental clustering algorithm based on OPTICS," Wirel. Pers. Commun., vol. 84, no. 3, pp. 2151–2170, 2015.

[18] K. Santhisree and A. Damodaram, "SSM-DBSCAN and SSM-OPTICS: Incorporating a new similarity measure for Density based Clustering of Web usage data," Int. J. Comput. Sci. Eng., vol. 3, no. 9, p. 3170, 2011.

[19] B. Bansal and S. Srivastava, "Hybrid attribute based sentiment classification of online reviews for consumer intelligence," Appl. Intell., vol. 49, no. 1, pp. 137–149, 2019.

[20] R. Douglass, "A cluster-based approach to browsing large document collections," in Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1992, pp. 318–329.

[21] S. J. Nanda and G. Panda, "A survey on nature inspired metaheuristic algorithms for partitional clustering," Swarm Evol. Comput., vol. 16, pp. 1–18, 2014.

[22] K. Subhadra, M. Shashi, and A. Das, "Extended ACO based document clustering with hybrid distance metric," in 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2015, pp. 1–6.

[23] Y. Kumar and G. Sahoo, "A hybrid data clustering approach based on improved cat swarm optimization and K-harmonic mean algorithm," AI Commun., vol. 28, no. 4, pp. 751–764, 2015.

[24] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, pp. 53–65, 1987.

[25] H. T. Sukmana, R. H. Gusminta, Y. Durachman, and A. F. Firmansyah, "Semantically annotated corpus model of Indonesian Translation of Quran: An effort in increasing question answering system performance," in 2016 4th International Conference on Cyber and IT Service Management, 2016, pp. 1–5.

[26] M. Z. Husin, S. Saad, and S. A. M. Noah, "Syntactic rule-based approach for extracting concepts from quranic translation text," in 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), 2017, pp. 1–6.

[27] I. Zeroual and A. Lakhouaja, "A new Quranic Corpus rich in morphosyntactical information," Int. J. Speech Technol., vol. 19, no. 2, pp. 339–346, 2016.

[28] H. Choiruddin, "Klasifikasi Kandungan Al-Qur'an." Jakarta: Gema Insani, 2005.

Image Restoration based on Maximum Entropy Method with Parameter Estimation by Means of Annealing Method

Kohei Arai [1]

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Abstract—**Image restoration based on Maximum Entropy Method (MEM) with parameter estimation by means of annealing method is proposed. The proposed method allows spatial resolution enhancement. Using overlap sampling with a low resolution of sensor, high spatial resolution (corresponding to the sampling interval) can be achieved through ground data processing with image restoration methods. Through the experiments with simulation imagery data derived from Advanced Very High Resolution Radiometer (AVHRR) data, it was found that spatial resolution enhancement can be achieved, MEM is superior to the others when S/N ratio is poor (less than 33) while Conjugate Gradient Method (CGM) is superior when the S/N ratio is higher than 33. It was also found that the CGM is superior to the proposed method for the existing sampling jitter.**

Keywords—*Image restoration; Maximum Entropy Method (MEM); annealing; Advanced Very High Resolution Radiometer (AVHRR); Conjugate Gradient Method (CGM)*

I. INTRODUCTION

Image restoration methods can be roughly classified into linear restoration filters and nonlinear restoration filters [1]. The former starts with classical Wiener filters and parametric Wiener filters that only restore the best approximation image on average, and evaluates the difference between the restored image and the original image on the observed image, not on the space of the original image. A general inverse filter, a least-squares filter with constraints, and a projection filter and a partial projection filter that may be greatly affected by noise in the restored image have been proposed [2]. However, the former is insufficient for optimization of evaluation criteria, etc. and is under study.

On the other hand, since the latter is essentially a method for finding a nonlinear solution, it can only take a method based on an iterative method, and various methods based on iterative methods have been tried. Although there are various iterative methods, there are a stationary iterative method represented by the successive over-relaxation method (SOR method) and a non-stationary iterative method represented by the conjugate gradient method [3].

In general, the former requires a large number of iterations, but the accuracy is high, and the latter has excellent convergence, but accumulation of rounding errors is a problem. When applied to image restoration, it is necessary to pay attention to noise resistance.

On the other hand, the maximum entropy method has been proposed as an image restoration method because it can take into account constraints (or prior knowledge) and resistance to noise [4]. However, this parameter estimation requires solving transcendental equations, and here again, iterative methods are needed [5]. As a method for estimating this parameter, methods using stationary iterative methods such as Newton's method and quasi-Newton's method and non-stationary iterative methods such as the conjugate gradient method have already been proposed. However, these methods are essentially local depending on the initial value. The solution may fall or the solution may diverge depending on the noise.

The method proposed in this article is based on image restoration based on the maximum entropy method, and in its parameter estimation, an annealing method was introduced to avoid the local solutions and divergence described above and obtain a stable minimum solution. Annealing is an iterative method of searching for the minimum solution of an objective function. In this case, it has been theoretically proved that the solution converges to a constant value regardless of the initial value [6]. This image restoration based on the maximum entropy method is a type of nonlinear restoration filter, and is the only image restoration that can be applied when the observation process is nonlinear, the constraints on the original image are nonlinear, and the evaluation criterion indicating the quality of the restored image is nonlinear. The problem of restoring high spatial resolution images from densely sampled data can be reduced to the problem of image restoration [7].

The author takes up the conjugate gradient method and the maximum entropy method, which are famous for reaching a solution in a finite number of times, and try to clarify the adaptive limits of image restoration based on these methods.

The superiority of the proposed method is compared with the image restoration based on the conjugate gradient method, taking the case of restoring a high spatial resolution image from a high-density sampled sequence as an example. In addition, the author has confirmed the robustness against noise compared with the most general least-squares method.

The following section describes research background followed by theoretical background. Then the proposed method is described followed by experiment. After that conclusion is described together with some discussions.

II. RESEARCH BACKGROUND

Inverse Problem in Super-Resolution and Its Solution is well introduced [8]. The image resolution using subpixel motion is proposed [9]. Also, a method for image compression with a cosmetic restoration is proposed [10]. On the other hand, DEM estimation with Simulated Annealing: SA based on surface reconstruction method is proposed [11]. Meanwhile, a method for estimation of Sea Surface Temperature: SST, wind speed and water vapor with microwave radiometer data based on simulated annealing is proposed [12].

Simultaneous estimation of geophysical parameters with microwave radiometer data based on accelerated Simulated Annealing: SA is proposed [13]. On the other hand, category decomposition method for un-mixing of Mixed Pixel: Mixels acquired with spaceborne based visible and near infrared radiometers by means of Maximum Entropy Method: MEM with parameter estimation based on Simulated Annealing is proposed [14]. Moreover, method for estimation of refractive index and size distribution of aerosol using direct, diffuse and aureole by means of simulated annealing is proposed [15]. Also, estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing is proposed [16].

III. PROPOSED METHOD

A. Image Restoration Method

In image restoration, when evaluating the quality of the estimated image f' of the original image f, f cannot be directly used for evaluation because f is unknown. Then, the observed image g that observed f' was compared with the estimated image Hf', and minimize the following L2 norm.

$$\| Hf'\text{-}g \|^2 \rightarrow \min. \tag{1}$$

It is assumed that the most likely original image f has been estimated when. Here, H is the transfer function of the image degradation operator. Such a method is called a least squares method.

Taking the variation of f' in equation (1) and setting it to 0

$$\delta f' H^t (Hf'\text{-}g) = 0 \tag{2}$$

$$H^t (Hf'\text{-}g) = 0 \tag{3}$$

Solving equation (3) for f' gives:

$$f' = (H^t H)^{-1} Hg \tag{4}$$

(Moor-Penrose generalized inverse matrix).

For equation (4), if H is regular and there is an inverse matrix,

$$f' = (H^t H)^{-1} Hg = H^{-1}g \tag{5}$$

That is, when an inverse matrix exists in H, the image restoration filter is the inverse matrix itself. However, such cases are rare. In addition to these, starting with the Wiener filter and the parametric Wiener filter that assume that the frequency spectrum of the signal-to-noise ratio is known, the difference between the restored image and the original image is evaluated and approximated on the observed image, not in the space of the original image.

Generalized inverse matrix filters, least-squares filters with constraints, and projection filters and partial projection filters that may be significantly affected by noise in restored images have been proposed, but should not be dealt with here.

B. Maximum Entropy Method

Suppose an image with non-negative elements is normalized. Since the value Pi of each pixel is (1) $\sum_{i=1}^{n} P_i = 1$ and (2) $\forall i: 1 > Pi > 0$, it can be regarded as a random variable. The entropy of the image can be calculated by considering the pixel value as the probability.

$$En = -\sum_{i=1}^{M \times N} fi \log fi \tag{6}$$

where En: image entropy, $M \times N$: is the number of elements in the image vector fi. However, in equation (6), the maximum value is taken when the image is flat, so a constraint condition is attached to this formula. The following restraint condition is considered,

$$F: \| Hf'\text{-}g \|^2 = \rho^2 \tag{7}$$

Then, the following Lagrange's undetermined multiplier λ is considered with MEM,

$$J = -fi \log fi + \lambda [\| Hf'\text{-}g \| - \rho^2] \tag{8}$$

$$= -fi \log fi + \lambda [(Hf'\text{-}g) * (Hf'\text{-}g) - \rho^2] \tag{9}$$

It can be said that it is equivalent to maximizing the above equation. Therefore, let f' which maximizes this energy J be the restored image in MEM. However, since this formula is a one-to-many mapping from vector to constant, and image restoration by filtering is difficult to realize, so an iterative method must be used to obtain the restored image.

That is, transcendental equations must be solved for this parameter estimation. As the parameter estimation method, methods using steady-state iterative methods such as Newton's method and quasi-Newton's method and non-steady-state iterative methods such as conjugate gradient method have already been proposed. May fall into a local solution or the solution may diverge depending on noise.

C. Parameter Estimation of Maximum Entropy Method by Annealing

The method proposed in this text is based on the image restoration based on the maximum entropy method, and in the parameter estimation, an annealing method was introduced in order to avoid the above local solution and divergence and obtain a stable minimum solution. It is a thing. Annealing is an iterative method for finding the minimum solution of an objective function. In this case, it is theoretically proved that the solution converges to a constant value regardless of the initial value [6].

The algorithm that embodies this is called simulated annealing, and is abbreviated here as annealing. The annealing algorithm is shown below.

1: Initial state *x* is set arbitrarily.
2: for *t:* = 0 to ∞ do begin
3: Perturb the state *x* and select the next state candidate *y*. Generate random numbers according to the probability distribution (..., *P* (*Y-y*), ...) ≡ (..., *P* (*x, y*), ...).
4: Set the temperature *T* (*t*): = *Δ* / ln (*t* + 2) and decide whether or not A is applied to y. The acceptance matrix is 1/{1 + exp*ΔE* / *T*}.
5: Generate random numbers according to the probability distribution (*P* (*Z* = 1), *P* (*Z* = 0)) ≡ (A (*x, y*), 1-A (*x, y*)). If Z = 1, swap *x* and *y*. If Z = 0, leave *x* unchanged.

Where *t* is the time, *Δ* is a large constant for temperature control, A is the acceptance matrix, and *P* is the perturbation matrix. Also, as shown in Fig. 1, the state after change is selected as the next state with a probability of 1/{1 + exp*ΔE* / *T*}.

The probability of transition can be reduced by gradually lowering the temperature *T* as the number of calculations increases. In other words, even if the temperature falls to a local value while the temperature is high, the transition to a higher energy is stochastically allowed, so that the local value can be escaped.

As a result, the optimal distribution of the following Boltzmann distribution defined by the objective function *F* and the positive constant *T* is determined, and the objective function reaches the minimum solution.

$$q(x) = \exp(-F(x)/T)/Z, \quad Z = \sum_{x=1}^{\infty} \exp(-F(x)/T) \qquad (10)$$

where, in the case of image restoration, the target of annealing is an image, and there is no pixel that cannot be selected when determining the next state y from a certain state *x*.
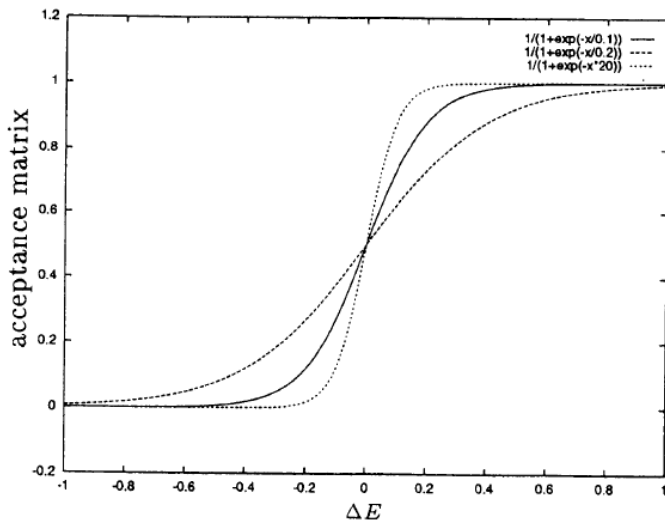
Fig 1.   A Function for Acceptance Matrix (In the Case for *ΔE/T*=0.1.
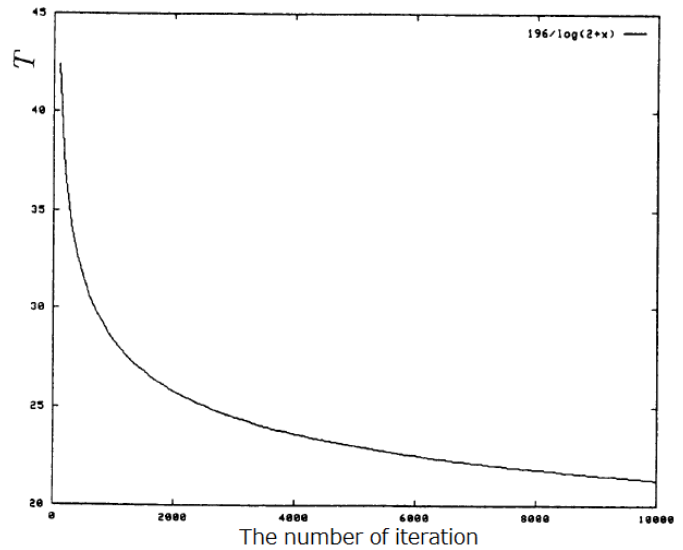
Fig 2.   Method for Decreasing of *T* in Accordance with the Number of Iteration.

TABLE I.   OPTIMUM Λ CAN BE DETERMINED FOR MINIMUM RMS ERROR

Λ	RMS ERROR
-1000	17
-2000	16
-3000	13
-4000	18
-5000	18

From this, a uniform random number is generated instead of the perturbation matrix, and the pixel to be changed next is selected according to the random number. Also, as shown in Fig. 2, instead of the acceptance matrix, the state after change is selected as the next column value with the probability of 1 / {1+ exp*ΔE* / *T*}.

The probability of transition can be reduced by gradually lowering the temperature *T* as the number of calculations increases. In other words, even if the temperature falls to a local value while the temperature is high, the transition to a higher energy is stochastically allowed, so that the local value can be escaped. Also, the value of the undetermined constant *λ* is changed and the Root Mean Square (rms) error is calculated to obtain the optimum value. For example, in the case of the image used in the experiment, the optimum λ is obtained as shown in Table I.

D. Image Restoration Method Based on the Conjugate Gradient Method

The conjugate gradient method is an effective solution for simultaneous linear equations with symmetric positive definite matrices as coefficients. If the observed image is a symmetric positive definite matrix and it is regarded as a system of linear equations represented by *g* = *Hf*, the image restoration can be performed using the conjugate gradient method.

The conjugate gradient method is performed by generating a vector sequence of iterative solutions (sequential approximations to the solution), a residual vector corresponding to the iterative solution, and a search direction

vector used for updating the iterative solution and the residual vector.

Also, for each iteration, two inner product operations are performed in order to calculate the updating scalar value that is defined so that the vector sequence satisfies the orthogonal condition. In a system of linear equations with a symmetric positive definite matrix as a coefficient, these conditions mean that the distance to the true solution is minimized under a certain norm.

The *i*-th iterative solution f^i is updated in each iteration by multiplying the search direction vector pi by a multiplier α^i.

$$f^i = f^{i-1} + \alpha^i \, p^i \tag{10}$$

The corresponding residual vector r^i is

$$r^i = r^{i-1} - \alpha \, q^i = r i^{-1} - H \, p^i \tag{11}$$

$$a = ai = r^{(i-1)\,t} \, r^{(i-1)} / p^{(i)\,t} \, H \, pi \tag{12}$$

Among all α expressed by equation (12), $r^{(i)\,t} H^{\,t} r^i$ is selected to be minimized. The search vector group is updated as follows using the residual vector.

$$p^i = r^i + \beta_{i-1} p^{i-1} \tag{13}$$

$$\beta i = r^{it} r^i / r^{(i-1)\,t} \, r^{i-1}$$

By choosing and β, p^i and $H \, p^{i-1}$, or the equivalent condition r^i and r^i are orthogonal. By selecting such βi, pi and ti become orthogonal to all previous $H \, pj$ and rj, respectively. (j = 1, 2, 3, ...) Next, the convergence condition of the conjugate gradient method will be described. In the iteration using the conjugate gradient method, noise may be increased in one iteration due to its fast convergence.

Therefore, it is necessary to introduce some noise suppression processing. Regularized Least Square Method is frequently used as this method.

$$e = \|gn - [H] \, g^2 + \gamma \|g\|^2 \tag{14}$$

is used as the error function. Where *gn* and γ are the measured data including noise and the power ratio of noise to signal. Here, assuming that noise is additive to the observed data *g*, such as *gn = g + n*, the first term on the right side of equation (14) corresponds to the noise power, and the second term is 1.

Since it is the product of / (S / N) and the power of the signal, this also corresponds to the power of noise. In the minimization of the q error function in equation (11), the first term and the second term are kept at the same magnitude, and as a result, a solution with S / N of 1 / r is obtained.

If the error function of equation (11) is partially differentiated with respect to *g* and the result is set to 0, the following is obtained.

$$[H]^{\,t} fn = ([H]^{\,t} [H] + \gamma \, [I]) \, g \tag{15}$$

Comparing equations (4) and (15), it can be seen that the matrix $[H]^{\,t} [H]$ is replaced by $[H]^{\,t} [H] + \gamma \, [I]$. In other words, by replacing $[H]^{\,t} [H]$ with $[H]^{\,t} [H] + \gamma \, [I]$ in the equation expressing the iteration of the conjugate gradient method, an image restoration algorithm based on the regularized least squares criterion is obtained. That is, it can be said that the image restoration by the conjugate gradient method is an optimization method that constrains S / N to be 1/γ.

IV. EXPERIMENT

A. Principle of Improving Resolution

By shortening the sample interval of the low-resolution sensor (this is called high-density overlap sampling), the resolution equivalent to the sample interval can be obtained. The direction *A* in the figure is the scanning direction of the satellite, and the direction *B* is the traveling direction of the satellite. For example, if the overlap sample ratio is *mo* in the satellite advancing direction and *no* in the satellite scanning direction (Fig. 3), the high-resolution image can be expressed by the following equation.

$$y_{ij} = \frac{1}{mo\,no} \sum_{k=1}^{mo} \sum_{l=1}^{no} x_{i+k,j+l} + e_{ij} \tag{16}$$

where, *mo*: satellite scan direction overlap sample ratio, *no*: satellite advancing direction overlap sample ratio, *x: i* observation target, *yij:* overlap sampled image, *l: i* miscellaneous *j* sound. This is expressed as a matrix as follows.

$$Y = AX + E \tag{17}$$
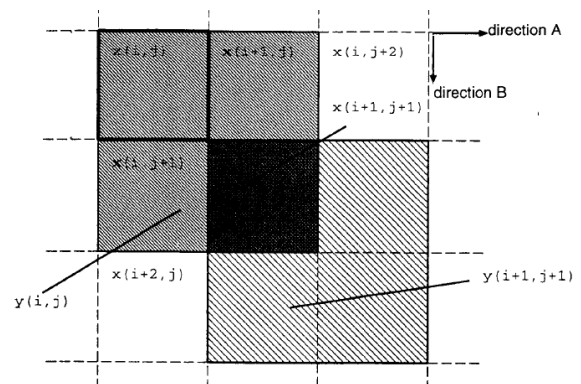
Fig 3. Spatial Resolution Enhancement with an Overlap Sampling.

The problem of restoring a high spatial resolution image from data sampled at high density can be reduced to the problem of image restoration [7]. That is, the problem of estimating high-resolution data from data with a sample interval of less than one pixel is a problem of restoring the original image *X* from the observed image *Y* in consideration of the deterioration operator *A* and noise *E*.

B. Simulation Images

In the experiment, National Oceanic and Atmospheric Administration: NOAA / Advanced Very High Resolution of Radiometer: AVHRR images (14 × 14 pixels) were used.

First, this image (original image) was multiplied by a matrix of high-density overlap samples to create an overlap sample image (observation image). Image restoration was applied to this image by the method described above to estimate the original image (estimated image or restored image), and the rms error from the original image was evaluated. Also, in that case, Table II is shown to examine the resistance to noise.

TABLE II.   GAUSSIAN NOISE USED

NOISE	MEAN	VARIANCE	S/N
A	0	30	46.7
B	0	60	33.0
C	0	150	20.8

TABLE III.   GAUSSIAN NOISE USED FOR JITTER SIMULATION

NOISE	MEAN	VARIANCE	REMARKS
a	0	1	Approx.±1% of jitter
b	0	4	Approx.±1% of jitter
c	0	9	Approx.±1% of jitter

The author prepared three types of noise, added them to the overlapped sample images, and then performed image restoration, and evaluated the noise characteristics of each restoration method. Furthermore, as the nonlinearity of the observation process, sampling jitter was simulated and added to the overlapped sample images.

This assumes that the electronic jitter of the sampling pulse is randomly mixed. The sampling interval is divided into 128 equal parts, the normal random numbers shown in Table III are generated, quantized in 256 (-127 to +128) steps, and the pixel position is shifted to be restored as an observed image.

C. Experimental Result

Fig. 4 to 7 show the original image, the overlapped sample image (overlap sample ratio is 4), and the restored images when the S / N is 33, respectively.

At this time, in the maximum entropy method with annealing (proposed method), as shown in Fig. 8, the rms error continues to decrease as the number of iterations increases. This is because it is theoretically guaranteed to approach the minimum solution infinitely. However, in the case of image restoration based on the conjugate gradient method, as shown in Fig. 9, when the noise is large (S / N = 33 or more), increasing the number of iterations may rather increase the rms error.

Fig 4.   The Original Image (Left) and the observed Image (Right).

Fig 5.   Restored Image with Noise based on MEM with the Parameter Estimation by Means of Annealing.

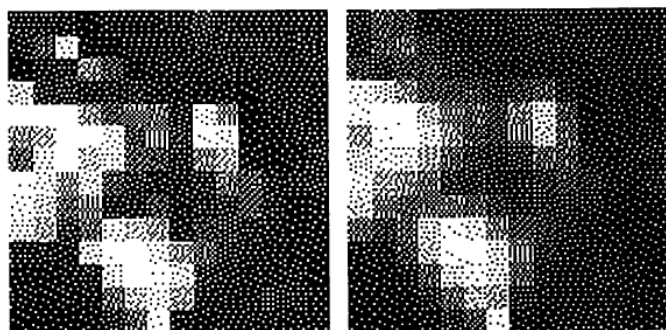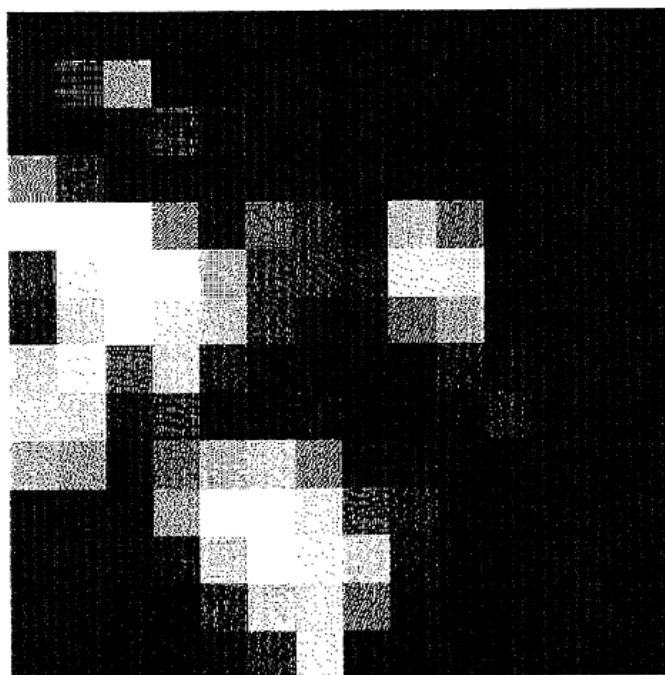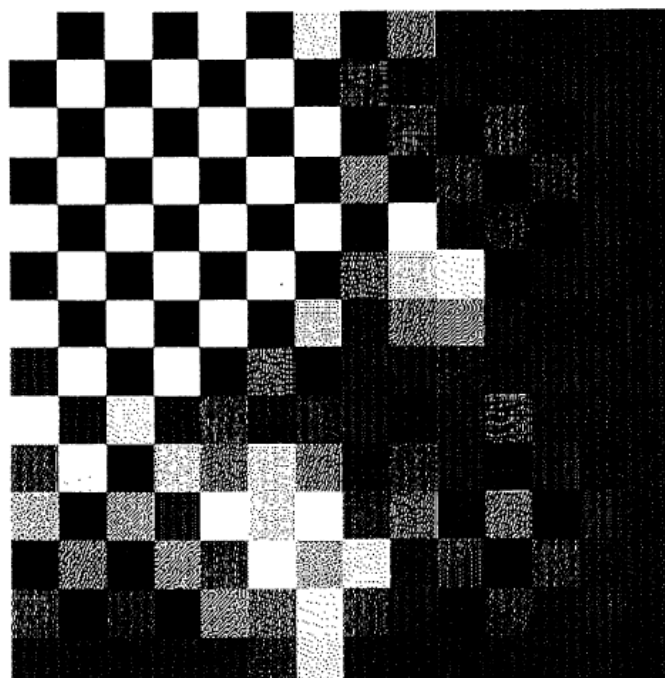Fig 6.   Restored Image with Noise based on Least Square Method: LSQ.
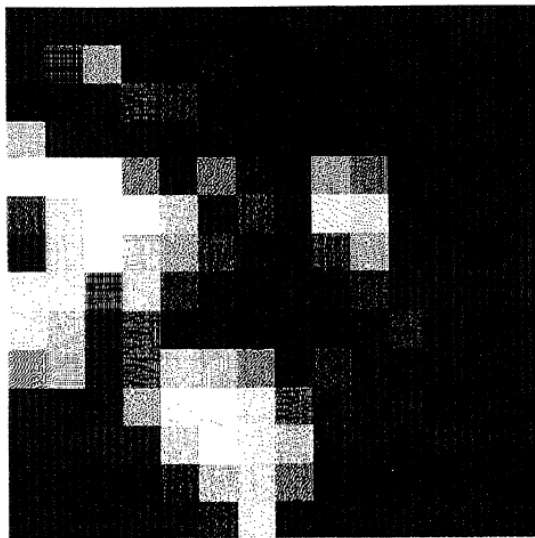
Fig 7.   Restored Image with Noise by Means of Conjugate Gradient Method.

Fig 8.   RMS Error for the Image Restoration based on MEM with the Parameter Estimation by Means of Annealing Decreases in Accordance with the Number of Iteration with the Parameter of λ.

Fig 9.   RMS Error for the Image Restoration by Means of Conjugate Gradient Method Decreases in Accordance with the Number of Iteration with the Parameter of Noise.

TABLE IV.   RMS ERROR FOR THE PROPOSED AND EXISTING METHODS (AGAINST NOISE WITHOUT JITTER)

TYPE	LEAST SQUARE	MEM WITH ANNEALING	CONJUGATE GRADIENT
0	0	10	0
A	39	10	7
B	79	11	12
C	99	23	24

TABLE V.   RMS ERROR FOR THE PROPOSED AND EXISTING METHODS (AGAINST SAMPLING JITTER WITHOUT NOISE)

TYPE	LEAST SQUARE	MEM WITH ANNEALING	CONJUGATE GRADIENT
0	0	10	0
a	69	19	22
b	88	21	28
c	98	27	33

The author was able to confirm that again. Table IV shows the results of evaluating the RMS error of each method. When the S / N is 33 (Noise Type B) or higher, the restoration effect of the proposed method (image restoration based on the maximum entropy method with annealing) is superior to that based on the conjugate gradient method in terms of rms error. The author found out that from this, it can be seen that the maximum entropy method with parameter estimation by annealing is superior in terms of resistance to noise.

It was also confirmed that noise resistance cannot be expected by the least-squares method in which the constraint conditions are not optimized.

Next, in the case of no additional noise, we tried to recover from the observed image with jitter added and evaluated the rms error with the original image. The results show that, as shown in Table V, the image restoration of the proposed method is the most tolerant to jitter. As mentioned in the foreword, this is due to the fact that the maximum entropy method is the only method that considers the nonlinearity of the observation system.

This jitter type c is equivalent to the jitter based on the normal random number of ±3% of the pixel interval considering the standard deviation. In the image restoration based on the conjugate gradient method, even if the jitter is 1%, the rms error is more than doubled compared to when it is not, and it can be said that it is sensitive, but the proposed method is relatively insensitive in that respect.

V.   CONCLUSION

It is confirmed that the image restoration based on the maximum entropy method with parameter estimation by annealing is effective in the noisy case. Furthermore, in the reconstruction problem of high spatial resolution images from high density sample images from NOAA / AVHRR data used in this experiment, if the S / N is better than 33, the image based on the conjugate gradient method is used. When restoration is sufficient, but when it is worse than that, it was found that image restoration based on the maximum entropy method with parameter estimation by annealing is superior.

Furthermore, it was confirmed that the resistance to noise cannot be expected by the least squares method in which the constraint conditions are not optimized. Comparing the

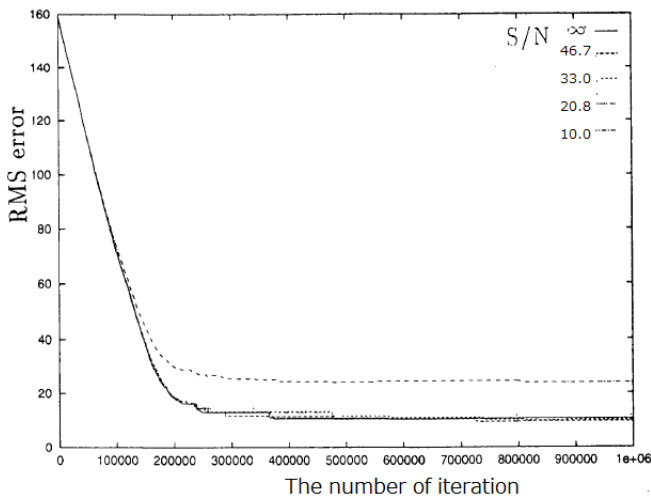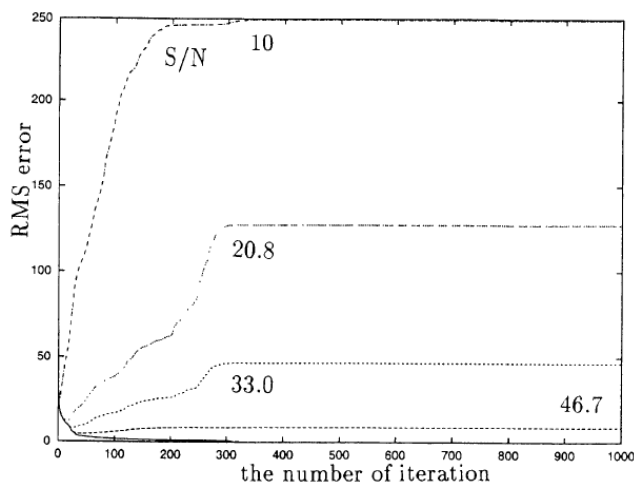resilience of the sampling pulse to electronic jitter, we find that the image restoration of the proposed method is the most tolerant to jitter. This is because the maximum entropy method can consider the nonlinearity of the observation system.

VI. Future Reaesrach Works

The proposed image restoration method is applicable to all of images. Although this time only NOAA/AVHRR remote sensing satellite images are used for the experiment, other general images have to be used for the future research works. Furthermore, it is confirmed that the proposed method takes much longer time in comparison to the other conventional image restoration methods, such as least square method. Conjugate Gradient Method: CGM. Not only restoration performance, but also required computation resources have to be evaluated in the future research works.

Acknowledgment

References

[1] Mikio Takagi, Y. Shimoda Edt, Kohei Arai. Image Analysis Handbook, The University of Tokyo Press, 1991.

[2] M.K.Ozkan, A.M.Tekalp and M.I.Sezan, "POCS-based restoration of space-varying blurred images" Vol.3, pp.450-454, Apr. 1994.

[3] Richard Barrett, Michael Berry, Tony F. Chan, James Demmle, June Donato, Jack Dongarra, Victor Eijkhout, Roldan Pozo, Charles Romine, Henk van der Vors, "Iterative Templetes", 1995.

[4] B. F. Burch, S. F. Gull and J.Skilling, Image Restoration by a Powerful Maximum Entropy Method, CVGIP, vol.23, no.2, pp.113-128, 1983.

[5] Seiichiro Yamaguchi, Sadayuki Narusawa, Tsuneo Saito, Yukio Hoshiko, Image Reproduction from Projection by Maximum Entropy Method, "IEICE Trans. (D), Vo1.J63-D, No.9, pp.1520-1527, 1982.

[6] Yoshinori Uesaka, Kazuhiko Ozeki, Algorithms for Pattern Recognition and Learning " Bunichi Sogo Shuppan, pp.163-184, 1990.

[7] Kohei Arai, Yasunori Terayama, "Improvement of spatial resolution of VTIR" Remote Sensing Society of Japan, Proceedings, pp.87-90, 1993.

[8] Shigeru Ando,"Inverse Problem in Super-Resolution and Its Solution" Mathematical Sciences, No.274, pp.56-61, 1986.

[9] S.Peleg, D. Keren and L. Schweitzer, GImproving image resolution using subpixel motionhPattern Recognition Letter, Vol.5, pp.223-226, Mar. 1987.

[10] Kohei Arai, T.Yamasaki and Y.Terayama Method for image compression with a cosmetic restoration, Proc.of the Australasian Conference on Remote Sensing, 1993.

[11] Kohei Arai, DEM Estimation with Simulated Annealing Based on Surface Reconstruction Method, Proceedings of the ISPRS Commission III, #80026, Vinne, July 1996.

[12] 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.

[13] 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.

[14] Kohei Arai, Category decomposition method for un-mixing of mixels acquired with spaceborne based visible and near infrared radiometers by means of Maximum Entropy Method with parameter estimation based on Simulated Annealing, International Journal of Advanced Research in Artificial Intelligence, 2, 4 64-69, 2013.

[15] Kohei Arai, Method for estimation of refractive index and size distribution of aerosol using direct, diffuse and aureole by means of simulated annealing, Proceddings of the COSPAR (Committee on Space Research) Congress, p.10 (Solicited Paper),2002.

[16] Kohei Arai and Jun Sakakibara, Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing, Abstracts of the 35th Congress of the Committee on Space Research of the ICSU, A1.1-0130-04, (2004)

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)

Design and Implementation of 6LoWPAN Application: A Performance Assessment Analysis

Nin Hayati Mohd Yusoff[1], Nurul Azma Zakaria[2]
Centre for Advanced Computing Technology (C-ACT)
Fakulti Teknologi Maklumat Dan Komunikasi
Universiti Teknikal Malaysia, Melaka, Malaysia

Adil Hidayat Rosli[3]
My6 Initiative Berhad 897128-T
1-21-01Sunteh @Penang Cybercity
Penang, Malaysia

Abstract—Industrial Revolution 4.0 promises an overall improvement to the communications technology by improving on the quality and flexibility of IoT application deployment. Currently, most of these applications are embedded devices from various manufacturers, networks, and technologies. As such, it would be total chaos of just getting the various and myriad devices and technologies to work together, let alone making them to work in perfect harmony. Regardless, the IoT espouses on the seamless integration and interoperability of the said devices and technologies. In realizing this goal, it would be imperative to say that the ability of the IoT system in adopting and adapting to new devices, services, and application is crucial, while at the same time it would not be in any way jeopardizing or compromising the existing system, especially the routing protocol. In view of the IP-based communication technology in WSN, the 6LoWPAN network has been chosen for the task, and the RPL protocol has been strongly considered for the 6LoWPAN solution. However, the RPL overhead tends to be spiralling upwards when additional information transmission occurs. In mitigating this anomaly, therefore, the HRPL was proposed to enhance the RPL protocol in reducing routing overhead. This study focusses on the performance analysis of RPL and HRPL based on the physical experimentation of the 6LoWPAN network in a real scenario. The results show HRPL protocol outperforms in all the performance-tested evaluations: CTO (38.7%), latency (26%), and convergence time (37%). It was also discovered that the number of DIS and DAO (RPL control message) packet is significantly reduced when the DIO message was reduced. At the same time, latency and convergence time also registered a decrease in their respective values correspondingly. Meanwhile, based on our observation, several experiments are needed to investigate how variants topology affect HRPL capabilities.

Keywords—6LoWPAN protocol; performance analysis; overhead

I. INTRODUCTION

IPv6 Low Power Area Network (6LoWPAN) is one of the IP-based communication technologies in WSN [1]. It allows a huge number of embedded devices to be connected directly to the internet (end to end communication) by using the large available IPv6 address space [2][3][4]. In these technologies, the transmission packets from a source node to destinations node over IEEE 802.15.4 based network using the multi-hop communication as we known wireless mesh network. Thus, the energy used for connectivity will save especially for data and information exchange through features of various metrics, including low bandwidth, different topologies (star, mesh, etc.), low power consumption, low-end node device cost, and scalable networks. Those, 6LoWPAN an essential part of the IoT development has been an explorer in [3] [5][6], especially for home automation systems. However, the current limitations of 6LoWPAN, such as a small frame length size could complicate to implement in real-life as has been discussed in [7]. The 6LoWPAN frame length for each data transmission limited to only 127 bytes as compared to IPv6's Maximum Transmission Unit (MTU), which is 1280 bytes. Therefore, for the successful transmission of packet frames, various techniques like header compressor, fragmentation, and reassembly technique are used to compress the User Datagram Protocol (UDP) header and IPv6 header. Unfortunately, in [8] and [9] state those techniques would affect the high packet reading of control traffic overhead. This issue contributes to the challenges in designing 6LoWPAN routing protocols that could provide efficient Quality of Services QoS [10] which ensures the network services are reliable for different users and applications [11] [12].

The IETF has proposed two working groups, Routing Over Low Power and Lossy (ROLL) networks and mobile ad hoc networks (MANET) to develop the 6LoWPAN routing algorithm requirement. Fig. 1 presents the multipath routing protocols and the taxonomy in 6LoWPAN.

According to [14] [8][15], RPL (from the ROLL working group) is effective, and simplicity has made it a strong candidate for the 6LoWPAN routing protocol. However, to suit into 6LoWPAN environments, some modifications with the effective technique are required especially in control traffic overhead. Based on a study in [12], the authors perform a detailed comparison of RPL, Ad-hoc On-demand Distance Vector (AODV), and Dynamic MANET On-demand (DYMO), exposing the network topology change, routing overhead and latency. It was found the performance of RPL is better than AODY and DYMO in terms of latency and throughputs, but obviously, RPL overhead is higher than AODY and DYMO. Besides, in 2016, [16] [17] in their study state that the RPL is effective in term of very fast network set up when implementing in 6LoWPAN but need further improved solution in overhead. Furthermore, the majority of the existing study of RPL protocol does not test in 6LoWPAN protocol stack and very limited study test real scenarios such as smart home automation. Therefore, it is necessary to study the RPL overhead to enhance the RPL protocol in order to achieve an efficient performance of QoS in real scenario 6LoWPAN network.
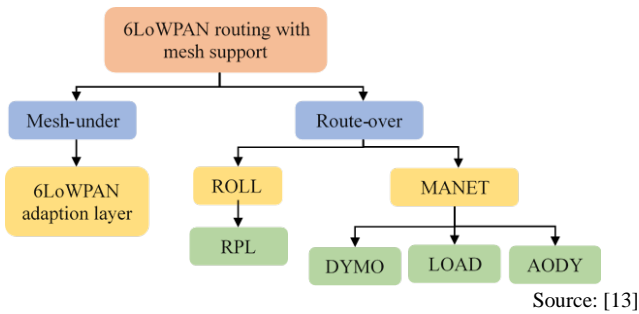
Fig. 1.   Multipath Routing Protocol Taxonomy.

Due to this matter, we conduct the intensive study via document analysis to investigate the factors that can define the causes of RPL routing overhead in 6LoWPAN [18]. We show that RPL routing overhead increased in three scenarios: i) Network topology change, ii) Additional information transmission, iii) Speed movement of node increased. As a result, a new RPL implementation for 6LoWPAN has been developed, namely HRPL, to enhance RPL that can minimize the routing overhead and provides an efficient way of communicating devices. As an extension of the work presented in [19], this paper we address the to analyze the performance of HRPL in a real application, an experiment platform based on 6LoWPAN environment has been designed and implemented. Hence, to assess the effectiveness of HRPL in the 6LoWPAN environment (grid topology) in three QoS performance metrics, there are 1) latency, 2) control traffic overhead, and 3) convergence time. Besides, in paper [20] mentioned the grid topology is considered as an advantage regarding scalability in exchange for the data between the node.

The reminders of this paper are organized as follows: Section II provided the explanations of routing protocol (RPL and HRPL), the methodology in testbed development, and the performance metric used. Result and discussion, we present in Section III, and finally, conclusion and futures work in Section IV.

II.   MATERIAL AND METHOD

A.   Routing Protocol

1) RPL: RPL protocol is a distance-vector that introduced for Low-Power and Lossy Networks (LLNs). RPL builds network topology called Destination-Oriented Directed Acyclic Graph (DODAG) using the objective function (OF) in selecting the best path and Directed Acyclic Graph (DAG). In fact, RPL has calculated the link in a network based on the least cost route between any two nodes is the route with distance and the cost of reaching a destination using Minimum Rank with Hysteresis Objective Function (MRHOF) that use minimum Expected Transmission Count (ETX) metrics or Objective Function Zero (OF0) that use minimum multi-hop metrics. To control the construction and maintenance of network topology, RPL generated three types of control

message: i) Destination-Oriented Directed Acyclic Graphs (DODAG) Information Object (DIO), ii) DODAG Information Solicitation (DIS), and iii) Destination Advertisement Object (DAO). Fig. 2 illustrates the DODAG structure, Fig. 3 shows the RPL control message sequence scenario with layer registration, Fig. 4 presents the process message flow that implements in this paper. DIS message is used for soliciting the DIO message to make a quick response to join the DODAG. First, Root Nodes (RN) broadcast the DIOs to select the root as their parent (upward route) and computes its RANK. The nodes receiving the DIOs message that contain information about RPL instance, compute the DODAGID rank, and DODAGVersionNumber allows to constructs and maintains the DODAG route. Then the RN setting as Parent Node(PN)/Intermediate Node (IN) (6LoWPAN Router (6LR) those 6LoWPAN Nodes (6LN) would re-broadcast the DIOs to the further nodes (Child Node (CN)) or Leaf Node (LN)). DAO messages are maintaining the downward routes while DAO Acknowledgement (DAO-ACK) is a unicast message in response to the DAO. These proceed repeat such as the way DODAG topology is constructed.
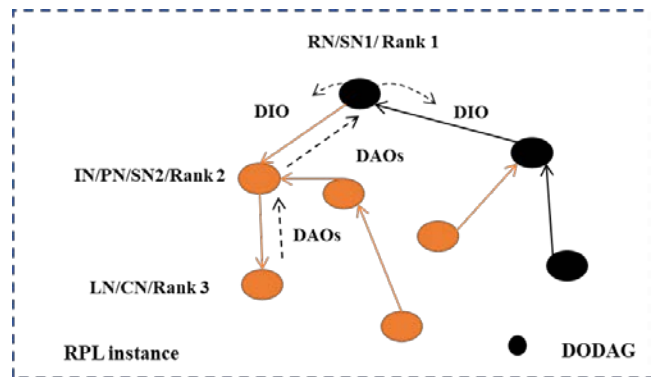
Fig. 2.   RPL Topology (DODAG Structure).
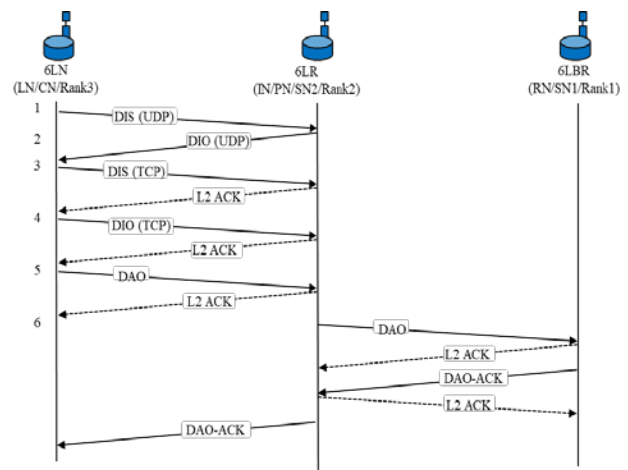
Fig. 3.   RPL Routing Layer Message Process.
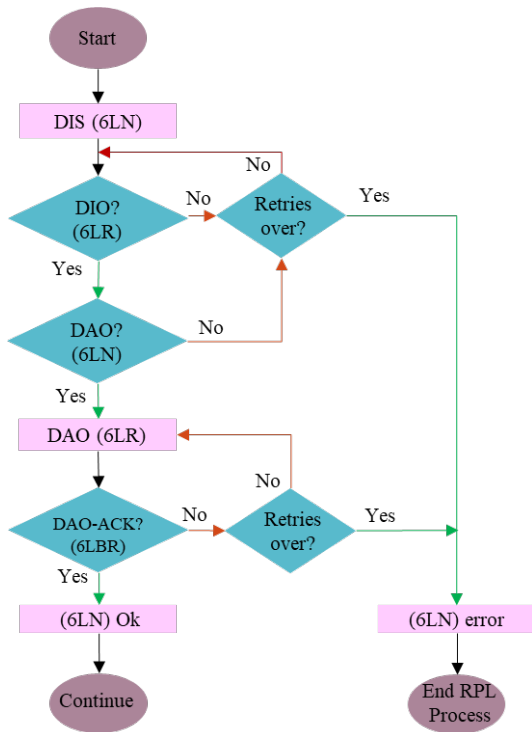
Fig. 4.   RPL Flowchart.

2) HRPL: HRPL was proposed to enhance the RPL. HRPL was introduced to reduce the Control Traffic Overhead (CTO) caused by redundant transmission. We introduced the T-Cut Off Delay to set the limit of delay and H field to respond on actions taken within the T-Cut Off Delay. As presented in Fig. 5, RN broadcast the DIO message to join the DODAG at the same time, the algorithm calculates and sets the T Cut Off Delay for the RN. If the RN receives duplicate DIO messages from other nodes before T-Cut Off Delay expires, the H field is a response based on three conditions: i) rebroadcast the DIO message) or ii) discard the DIO message or iii) to make no changes. However, when the T-Cut Off Delay expires the nodes must rebroadcast the DIOs messages. The addition of the H field is expected to reduce the control traffic overhead at the same time maximize the network lifetime, and minimize the latency.

B. Method

This research focuses on evaluating the performance of HRPL and RPL protocol in real sensor nodes (SN) based on the 6LoWPAN environment. We compare the existing RPL protocol with our proposed protocol HRPL in terms of control traffic overhead, latency, and convergence time. Fig. 6 illustrates the testing and validation design in this analysis.

C. Building the Testbed Scenario

For the validity of our approach, we use a Texas Instruments (TI) CC2538 wireless controller system-on-chip for 6LoWPAN application compliant 2.5-GHz IEEE 802.15.4. Table I shows the list of equipment required for a physical deployment scenario. All the work on the development of the 6LoWPAN testbed is presented in Fig. 7. The development

process involves the deployed hardware infrastructure, physical network, and integrated with software services for managing and controlling the hardware.
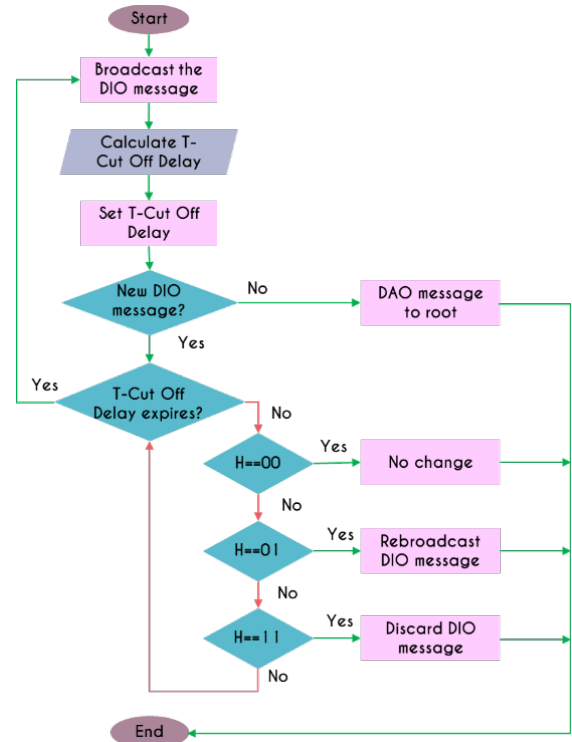
Fig. 5.   HRPL Flowchart.
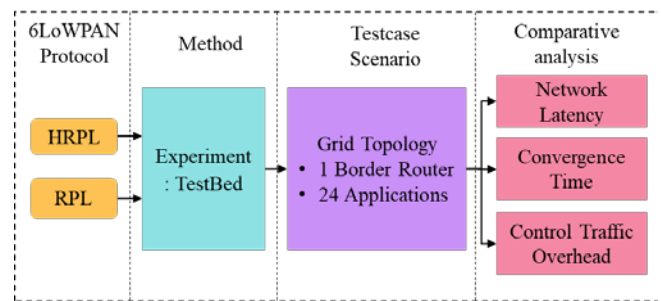
Fig. 6.   Testing and Validity Design.

TABLE I.   6LoWPAN Testbed Equipment

Hardware Equipment		
Equipment	Count	Types and Description
Sensor Node	24	• The ARM® Cortex™-M3 processor, CC2538 System-on-Chip for 6LoWPAN Application, and 2.4Ghz IEEE 802.15.4.
Border Router	1	
Sniffer	1	• CC2531 USB Dongle - Fully operational USB device that can be plugged into a PC 2 x LEDs 2 x pushbuttons USB A connector for PC 10-pin program/debug header PCB antenna (meandered inverted F) Connector pads for external sensors
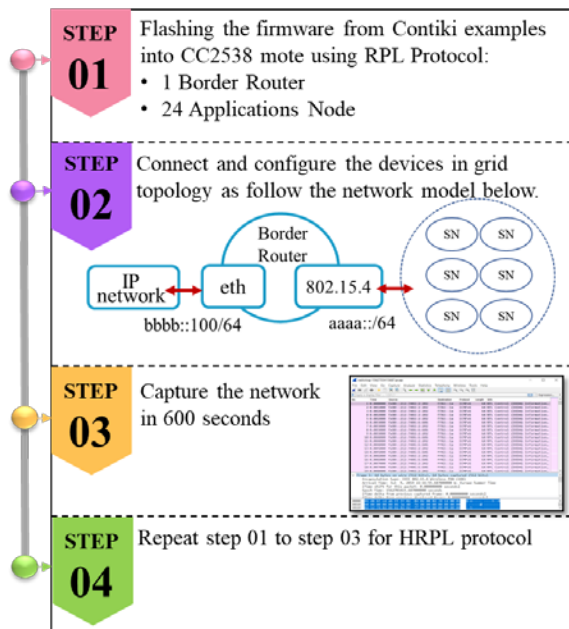
Fig. 7.   The Procedure for Development of 6LoWPAN Testbed.

TABLE II.   SYSTEM PARAMETER FOR CC2538 MOTE FIRMWARE

6LoWPAN Protocol	HRPL compared with RPL
Physical (PHY) and Data Link	IEEE 802.15.4 PHY IEEE 802.15.4 MAC and CSMA
Trickle Algorithm	Default
Nodes rank	MRHOF + ETX Metrics
DIO_Interval Min	12ms
DIO_interval doublings	8ms

TABLE III.   THE CONFIGURATION OF THE 6LBR PARAMETER

Parameter	Setting
Mode Selection	MODE=ROUTER
TAP mode - bridged or tunneled to the Ethernet network	RAW_ETH=0
Bridge Configuration	BRIDGE=1 CREATE_BRIDGE=1 DEV_ETH=eth0 DEV_BRIDGE=br0 DEV_TAP=tap0
SLIP Radio configuration	DEV_RADIO=/dev/ttyACM0 BAUDRATE=115200
Configuration Files	NVM=conf-templates/test.dat CONFIG=/etc/6lbr/nvm.conf
6LBR installation paths	BIN_6LBR=$LIB_6LBR/bin
COOJA support	SOCK_RADIO=localhost

First, a modified firmware from the Contiki example is flashed into the CC2538 mote. One of the motes is used as a Border Router (BR) (router mode) while the rest of the motes are flashed as an SN. The firmware parameter setup for HRPL and RPL are presented in Table II. For this experiment, the default trickle algorithm [21] used to control the amount of DIO message in routing traffic by setting the parameter of DIO Interval Min as 12ms and DIO interval doublings as 8ms. To achieve the best Packet Delivery Ratio (PDR), the Minimum Rank with Hysteresis Objective Function (MRHOF, RFC6719) with Expected Transmission Count (ETX, RFC 6551) metrics are used as path selection mechanism [22][23].

In the next step, the BR running the 6LBR software will act as a proxy/translator between the 802.15.4 network and the LAN network. The 6LBR can be operated as a smart bridge, router, or a transparent bridge, which in this experiment is being configured as a router. The 802.15.4 network is assigned the IPv6 subnet of aaaa::/64 while the LAN network is assigned the bbbb::/64. Fig. 8 shows the 6LoWPAN testbed using the CC2538 scenario where the nodes are deployed with grid topology in an indoor environment. Fig. 9 displays the 6LBR information and Fig. 10 exhibits the SN information in the webserver. The configuration setting of the 6LBR parameter as presented in Table III.

The RPL/HRPL control messages (DIS, DIO, DAO, DAO-ACK) are received when NDP configures the RPL/HRPL network. We observed the RPL control message to analyze 6LoWPAN packets by running a 6LoWPAN traffic sniffer and the Wireshark packet capture tool, as shown in Fig. 11. We recorded the capture after 30 minutes of the experiment running [24] and the observation time is 30 minutes. For energy monitoring, only the energy consumption of the radio transceivers is monitored based on the chip's manufacturer datasheets as shown in Table IV. We ignored other small energy consumptions such as sensors, actuators, and microcontrollers.

Fig. 8.   Development of 6LoWPAN Testbed using CC2538 Mote.

Fig. 9.   Interface of 6LBR in Web Server.

Fig. 10. Interface of Sensor Node in Web Server.
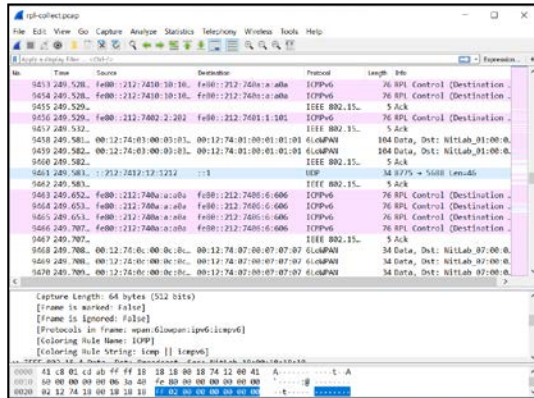
Fig. 11. Wireshark Capture Interface.

TABLE IV.   ENERGY MODEL OF CC2538

State	Current
Active-Mode Listen (RX)	20 mA
Active-Mode Transmit TX at 0 dBm	24 mA

1) Performance Metric

2) Control Traffic Overhead (CTO): The total number of DIS, DIO, DAO message transmitted by nodes for the formation of DADOG in the network. CTO is calculated using the following equation (1) [25]–[27]:

$$CTO = \sum_{k=1}^{n} DIS\,(k) + \sum_{k}^{n} DIO\,(k) + \sum_{k=1}^{n} DAO\,(k) \quad (1)$$

3) Network Latency: Network Latency is defined as the difference of the taken time between the sending and the receiving of a packet via the DADOG path [25], [28]. The total of latency as mentioned in equation (2).

$$Total\ Latency = \sum_{k=1}^{n}(Received\ Time\,(k) - Sent\ Time\,(k) \quad (2)$$

4) Convergence time (CT): CT is the total time needed for all the reachable nodes to join the DAG [29], [30] including the number of times to perform a global repair of the network after a link failure. In nutshell, CT is the process of building the DODAG from the first DIO messages sent until the last

DIO joined the DAG. CT calculated from the DIO message Equation (3) presents the formula to calculate the CT.

$$Convergence\ Time = \ Last\ DIO\ joined\ DAG - First\ DIO\ Send \quad (3)$$

III. RESULT AND DISCUSSION

A. Control Traffic Overhead

Fig. 12 to 15 shows the comparison of the control message packets (DIS, DIO, DAO, CTO) between HRPL and RPL for each node after 30 minutes of the experiment time being. In these scenarios, HRPL sends 41.6% DIS, 36.2% DIO, and 40.8% DAO lesser than RPL. The result shows that HRPL significantly decreased 38.7% CTO packets on average compared to RPL as presented in Fig. 16. In fact, our observation shows, RPL generates too many control messages to build a DODAG route, which is used to propagate the routing tables between nodes. We also monitor that the DIO message packet was dominant the rest (DIS and DAO) of the RPL/HRPL control message packet in updating and construct the destination route. By enhancing the T-Cut delay (HRPL), a new DIO message will only operate based on H field conditions as long as the T-Cut delay is not over. On the other hand, this condition has reduced the retransmission of the DIO message. Indeed, our design methods present a better CTO by choosing the route path with the lowest cost and maximum PDR. Overall results conclude that the DIO, DIS, DAO, and CTO messages have been significantly reduced with the use of HRPL.
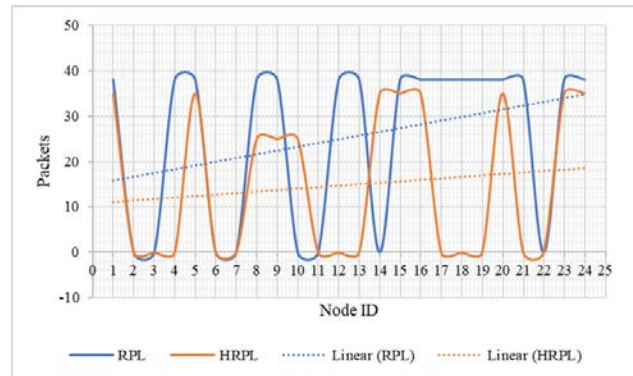
Fig. 12. DIS Messages for Each Node.
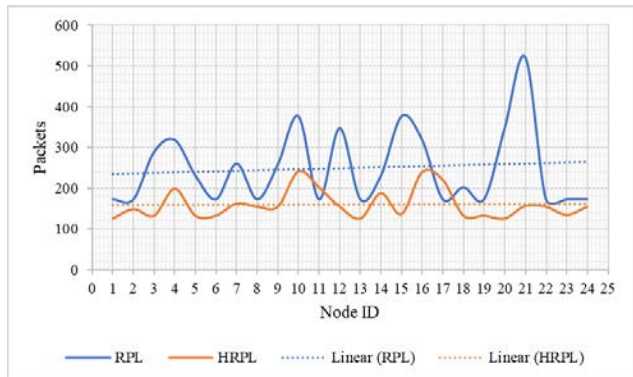
Fig. 13. DIO Messages for Each Node.
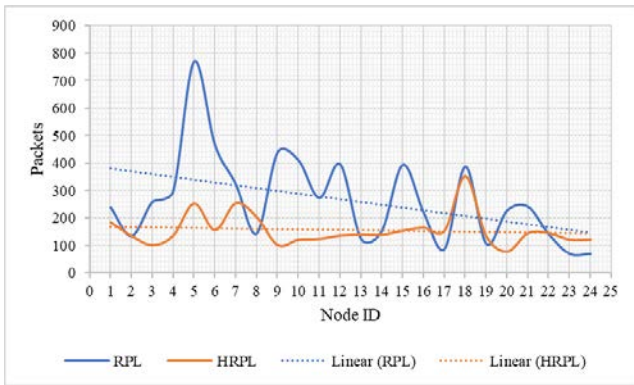
Fig. 14. DAO Messages for Each Node.
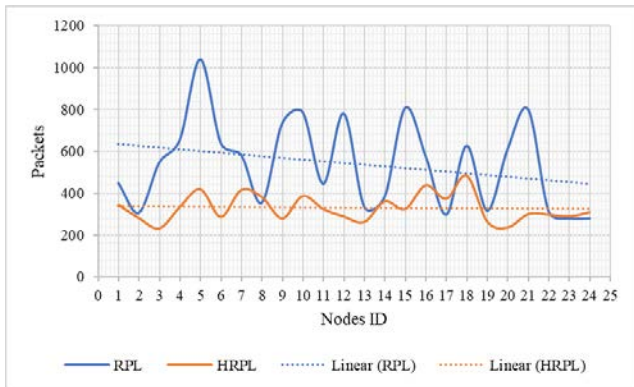
Fig. 15. CTO for Each Node.
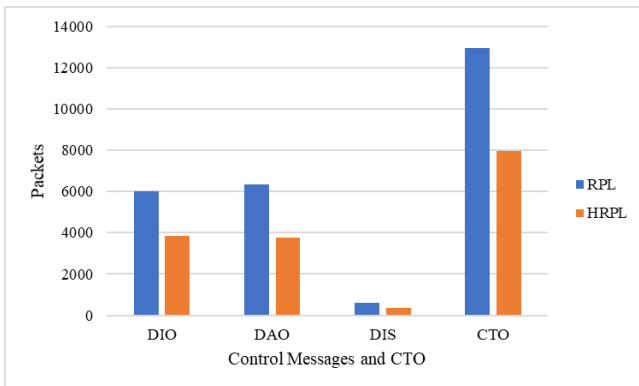
Fig. 16. Total of Control Messages and CTO.

B. Network Latency

Fig. 17 provides a comparison of the network latency between HRPL and RPL for each node. The test result shows that HRPL has lower latency (26% lower) compared to RPL. The feature of the RPL radio duty cycle is the sender nodes send the DIO Message repeatedly until it gets the DIO Acknowledgment (DIO-ACK). The receiver nodes wake up to listen on packet transmission from the neighbor and still in radio on until transmission completes. By using the cut-off delay (HRPL), the delay for waiting the receiver wakes up has been minimized. This situation could reduce network latency.

It happens because the node with less packet still in radio on until the time of cut of delay expires. Due to this matter, the extra timing for the receiver wakes up on competing for the packet transmission is reduced. Whereas in RPL the less nodes go to radio off and it takes extra timing for switching back to radio on. These results are consistent with those of [11] who found that the radio duty cycle has an impact on the latency. These differences can be explained in part by the latencies for each node as presented in Fig. 18 (RPL) and Fig. 19 (HRPL). The total of RPL latency is 0.165s and HRPL is 0.121s. We noticed that the network latency of HRPL is significantly reduced compared to RPL.

C. Convergence Time

As shown in Fig. 20, HRPL has better convergence time compared to RPL. HRPL shows 37.6% faster convergence time in this experiment. Due to this matter, the H condition in HRPL response was reducing the objective function (MRHOF-ETX) to take less time in selecting the best parent to join the DAG. At the same time, our approach successfully dropped the control message send across the network during the process of DODAG construction. In comparison with RPL, objective function took a longer time for RPL to build the DODAG, which may be caused by the complex process of objective function calculating the number of packet retransmission over the link.
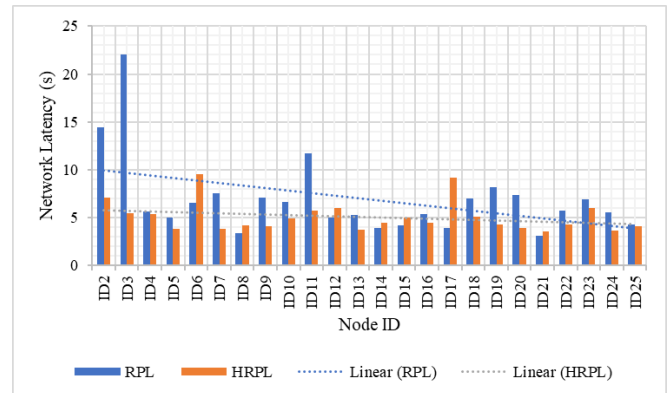
Fig. 17. Comparison of Latency between RPL and HRPL.
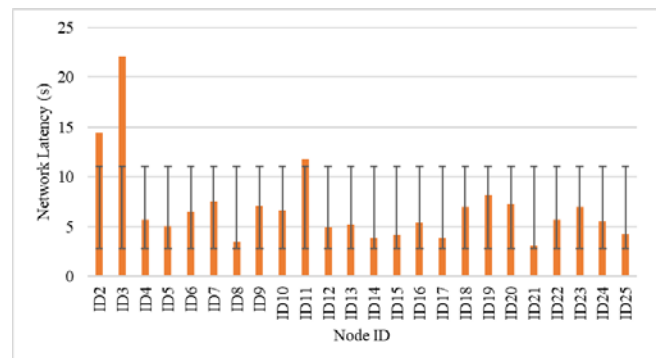
Fig. 18. The Total of RPL Latencies for Each Node.
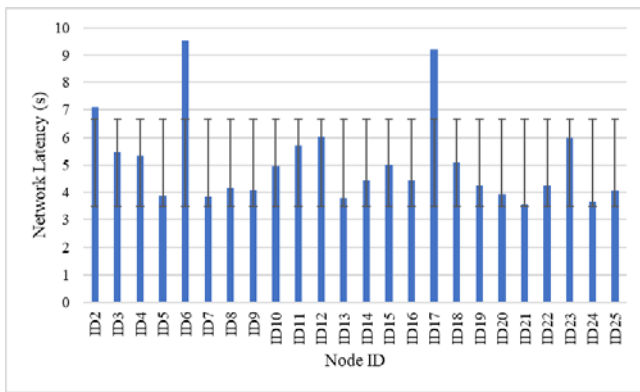
Fig. 19. The Total of HRPL Latencies for Each Node.
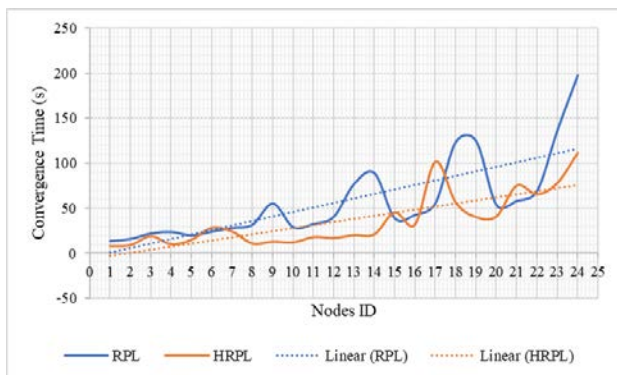
Fig. 20. RPL and HRPL Convergence Time for Each Node.

IV. CONCLUSIONS

In this paper, a testbed has been developed to support physical experimentation which addresses the performance of our HRPL protocol in the real 6LoWPAN scenario. The result shows HRPL performs much better than RPL in all performance metrics that measured (CTO, latency, and convergence time). In conclusion, the testbed developed favorable result in implementing HRPL in a real scenario. However, still many issues could be researched especially the way to reduce the latency. It ended up wasting energy, as the receiver periodically needs to wake up to listen on packet transmissions from its neighbours. This phenomenon could be explained by the fact that in the UDP message transmission via IEEE 802.15.4 frame data, a packet is sent repeatedly until it gets an acknowledgment from the receiver. Further research should be done to investigate the HRPL in large-scale network in multiple topologies, and for a different number of nodes.

REFERENCES

[1] S. Ahmad and J. Zhang, "Network-State-Aware Quality of Service Provisioning for the Internet of Things," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 6, pp. 369–376, 2016.

[2] S. El Bouanani, O. Achbarou, M. A. Kiram, and A. Outchakoucht, "Towards understanding internet of things security and its empirical vulnerabilities: A survey," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 10, pp. 337–345, 2019.

[3] B. H. Malik et al., "Investigating technologies in decision based internet of things, internet of everythings and cloud computing for smart city," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 580–587, 2019.

[4] M. Gurunathan and M. A. Mahmoud, "A review and development methodology of a lightweight security model for IoT-based smart devices," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 2, pp. 125–134, 2020.

[5] Y. Mahmood, N. Kama, A. Azmi, and S. Ya'acob, "An IoT based home automation integrated approach: Impact on society in sustainable development perspective," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, pp. 240–250, 2020.

[6] R. F. Al-Mutawa and F. A. Eassa, "A smart home system based on internet of things," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 2, pp. 260–267, 2020.

[7] Z. Shelby and C. Bormann, 6LoWPAN : The Wireless Embedded Internet. 2009.

[8] G. Keng Ee, C. Kyun Ng, N. Kamariah Noordin, and B. Mohd Ali, "A Review of 6LoWPAN Routing Protocols," Proc. Asia-Pacific Adv. Netw., vol. 30, pp. 71–81, 2010.

[9] J. J. P. C. Rodrigues and P. A. C. S. Neves, "A survey on IP-based wireless sensor network solutions," Int. J. Commun. Syst., vol. 23, no. 5, pp. 633–652, 2010.

[10] R. Garg and S. Sharma, "A study on Need of Adaptation Layer in 6LoWPAN Protocol Stack," I.J. Wirel. Microw. Technol., vol. 3, pp. 49–57, 2017.

[11] T. Zhang and X. Li, "Evaluating and analyzing the performance of RPL in contiki," in MSCC '14: Proceedings of the first international workshop on Mobile sensing, computing and communication, 2014, no. August 2014, pp. 19–24.

[12] H. Xie, G. Zhang, D. Su, P. Wang, and F. Zeng, "Performance evaluation of RPL routing protocol in 6lowpan," in Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, 2014.

[13] L. M. L. Oliveira, A. F. De Sousa, and J. J. P. C. Rodrigues, "Routing and mobility approaches in IPv6 over LoWPAN mesh networks," Int. J. Commun. Syst., vol. 24, no. 11, pp. 1445–1466, 2011.

[14] K. Naito, "A Survey on the Internet-of-Things : Standards , Challenges and Future Prospects," J. Inf. Process. Vol.25, vol. 25, pp. 23–31, 2017.

[15] A. Saaidah, O. Almomani, L. Al-Qaisi, and M. K. Madi, "An efficient design of RPL objective function for routing in internet of things using fuzzy logic," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 184–190, 2019.

[16] J. Kantert, C. Ringwald, G. Von Zengen, S. Tomforde, L. Wolf, and C. Muller-Schloer, "Enhancing RPL for robust and efficient routing in challenging environments," Proc. - 2015 IEEE 9th Int. Conf. Self-Adaptive Self-Organizing Syst. Work. SASOW 2015, pp. 7–12, 2015.

[17] K. Iwanicki, "RNFD: Routing-Layer Detection of DODAG (Root) Node Failures in Low-Power Wireless Networks," 2016 15th ACM/IEEE Int. Conf. Inf. Process. Sens. Networks, IPSN 2016 - Proc., 2016.

[18] N. H. Mohd Yusoff, N. A. Zakaria, and N. Harum, "Problem analysis of RPL overhead in 6LOWPAN using 5W1H model," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 12, pp. 5300–5305, 2019.

[19] N. H. Mohd Yusoff, N. A. Zakaria, A. Sikora, and J. S. Sebastian, "6LoWPAN Protocol in Fixed Environment: A Performance Assessment Analysis," Proc. 2019 10th IEEE Int. Conf. Intell. Data Acquis. Adv. Comput. Syst. Technol. Appl. IDAACS 2019, vol. 2, pp. 1142–1147, 2019.

[20] B. S. Hassen, S. A. S. Lafta, H. M. Noman, and A. H. Ali, "Analyzing the performances of WSNs routing protocols in grid- based clustering," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 9, no. 4, pp. 1211–1216, 2019.

[21] P. Levis, T. H. Clausen, J. W. Hui, O. Gnawali, and J. Ko, "The Trickle Algorithm," IETF RFC6206, 2011.

[22] N. Pradeska, Widyawan, W. Najib, and S. S. Kusumawardani, "Performance analysis of objective function MRHOF and OF0 in routing protocol RPL IPV6 over low power wireless personal area networks (6LoWPAN)," Proc. 2016 8th Int. Conf. Inf. Technol. Electr.

Eng. Empower. Technol. Better Futur. ICITEE 2016, no. 7863270, p. 7863270, 2017.

[23] O. Gnawali and P. Levis, "The Minimum Rank with Hysteresis Objective Function," IETF RFC6719.

[24] M. Banh, N. Nguyen, K. H. Phung, L. Nguyen, N. H. Thanh, and K. Steenhaut, "Energy balancing RPL-based routing for Internet of Things," 2016 IEEE 6th Int. Conf. Commun. Electron. IEEE ICCE 2016, pp. 125–130, 2016.

[25] H. Lamaazi and N. Benamar, "OF-EC: A novel energy consumption aware objective function for RPL based on fuzzy logic.," J. Netw. Comput. Appl., vol. 117, pp. 42–58, 2018.

[26] E. Ancillotti, R. Bruno, and M. Conti, "Reliable data delivery with the IETF routing protocol for low-power and lossy networks," IEEE Trans. Ind. Informatics, vol. 10, no. 3, pp. 1864–1877, 2014.

[27] B. Ghaleb, A. Al-dubai, E. Ekonomou, B. Paechter, and M. Qasem, "Trickle-Plus : Elastic Trickle Algorithm for Low- Power Networks and Internet of Things," in IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 2016, pp. 103–108.

[28] J. Guo, P. Orlik, K. Parsons, K. Ishibashi, and D. Takita, "Resource aware routing protocol in heterogeneous wireless machine-to-machine networks," 2015 IEEE Glob. Commun. Conf. GLOBECOM 2015, 2015.

[29] S. Biju and N. M. Shekokar, "Optimize the energy efficiency of RPL based 6L0WPAN by FL clustering," 2017 Int. Conf. Big Data, IoT Data Sci. BID 2017, vol. 2018-Janua, pp. 159–166, 2018.

[30] C.-A. La, M. Heusse, and A. Duda, "Link reversal and reactive routing in Low Power and Lossy Networks," IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC, pp. 3386–3390, 2013.

An Automated Framework for Detecting Change in the Source Code and Test Case Change Recommendation

Niladri Shekar Dey[1], Purnachand Kollapudi[2], M V Narayana[3], I Govardhana Rao[4]

Associate Professor, Department of CSE, B V Raju Institute of Technology, Narsapur, Medak Dist.Telangana State, India[1, 2]
Professor, Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad, India[3]
Assistant Professor, Department of CSE, Osmania University, Hyderabad, India[4]

Abstract—Improvements and acceleration in software development has contributed towards high quality services in all domains and all fields of industry causing increasing demands for high quality software developments. In order to match with the high-quality software development demands, the software development industry is adopting human resources with high skills, advanced methodologies and technologies for accelerating the development life cycle. In the software development life cycle, one of the biggest challenges is the change management between versions of the source codes. The versing of the source code can be caused by various reasons such as change in the requirements or adaptation of functional update or technological upgradations. The change management does not only affect the correctness of the release for the software service, rather also impact the number of test cases. It is often observed that, the development life cycle is delayed due to lack of proper version control and due to the improver version control, the repetitive testing iterations. Hence the demand for better version control driven test case reduction methods cannot be ignored. A number of version control mechanisms are proposed by the parallel research attempts. Nevertheless, most of the version controls are criticized for not contributing towards the test case generation of reduction. Henceforth, this work proposes a novel probabilistic refactoring detection and rule-based test case reduction method in order to simplify the testing and version control mechanism for the software development. The refactoring process is highly adopted by the software developers for making efficient changes such as code structure, functionality or apply change in the requirements. This work demonstrates a very high accuracy for change detection and management. This results into a higher accuracy for test case reductions. The final outcome of this work is to reduce the development time for the software for making the software development industry a better and efficient world.

Keywords—*Change detection; pre-requisite detection; feature detection; functionality detection and test case change recommendation*

I. INTRODUCTION

The improvements in the code development is a must to be performed task for all software development cycles, due to the continuous changing client requirements. The improvements or the changes in the software source code can be done in various ways such as version control or requirement tracking or using third party tools. Nonetheless, the most frequent and highly adopted method is the refactoring method as suggested by M.

Fowler et al. [1]. The effect of refactoring on the software source code is highly compatible with the change management process and further with the other phases of software development life cycle. The notable outcome by the work of E. R. Murphy-Hill et al. [2] have listed the standard phases of refactoring of source code, which deeply influences the adaptation of the process. The detailed comparative analysis of other versioning methods with refactoring is performed by N. Tsantalis et al. [3] highlighting the benefits of refactoring over other methods. The challenges of refactoring process for any source code cannot be ignored and can cause higher complexity during versioning in case of improper management as demonstrated by M. Kim et al. [4]. Another study focuses on the software development improvisation by Microsoft, suggesting similar measures as documented by Miryung Kim et al. [5]. Also, the similar study is conducted on another open source tool, GitHub, by D. Silva et al. [6] and the result is same as the previous studies recommending similar measures to be followed for safe refactoring of the source code (Fig. 1).

Therefore, understanding that the refactoring (Fig. 2) of the source code can be highly helpful for source code changing, most of the development practices uses this method.

Nevertheless, the process of refactoring the code can be helpful for making controlled changes into the code, but these changes results into further changes of testing process and test case management. Hence, the demand for change detection and test case verification without repeating the test cases for the features, which has not changed during the refactoring process, is highly prioritized by the industry practitioners and researchers. Thus, this work attempts to provide a solution to the change detection and test case reductions.

The rest of the work is furnished such as in the Section II, the outcomes from the parallel researcher are analyzed, in Section III, problem definition and the scope for improvements are listed, in Section IV, the proposed change detection algorithm is discussed, in Section V, the proposed test case detection and reduction algorithm is elaborated, in the Section VI, the proposed complete automated framework is furnished, in the Section VII, the results are discussed, in the Section VIII, the comparative analysis for understanding the improvements are discussed and in the Section IX, this work presents the final conclusion.

Fig. 1.   Source Code Change Detection.
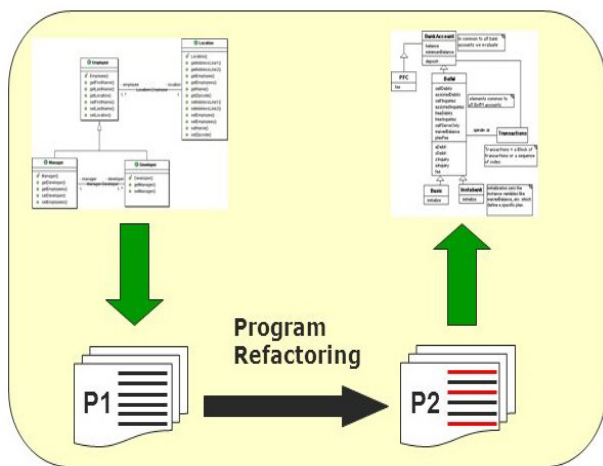
Fig. 2.   Refactoring of Source Codes.

II.   PARALLEL RESEARCH OUTCOMES

The versioning of the source code is performed in order to include changes in the source code. Often the changes are recommended by the customer or the changes are made due to the technical requirements fulfilments. Thus, refactor results into changes in pre-requisites or the feature of the source code or functionality of the source code. Hence, detecting the correct changes are the prime important task.

In order to detect the correct changes after a source code is refactored is the prime task. A number of parallel researches are taken place to accomplish this task. In this section of the work, the parallel research outcomes are analysed.

The first case study produced by E. R. Murphy-Hill et al. [2] have reported a framework that collects the historical data from the source code version control and integrates the changes into popular Eclipse IDE. The advancements of this work are done by S. Negara et al. [7], where the process of using meta data generated by version history is used. Nevertheless, this process is completely dependent on the refactoring trails or the auto-generated information during the refactoring process.

Removing the dependencies on the auto-generated information by the refactoring tools, the work of J. Ratzinger et

al. [8] proposes a framework to generate commit messages during the refactoring process. This feature enables the framework to detect all changes including the minor updates. Regardless to mention, this framework is expected to be deployed from the beginning of the code development life cycles, which makes this framework being criticized among the practitioner's community. The other popular strategies supporting this method were also made. The work of Miryung Kim et al. [5] have finetuned the framework for detecting further detection of changes.

Yet other popular methods for detecting the change are analysing the pattern and behaviours of the source code as demonstrated by G. Soares et al. [9] or analysing the software code metrics as represented by S. Demeyer et al. [10].

In the other hand, detecting refactoring using the static code analysis is also widely accepted method. The work by D. Dig et al. [11] on component-based detection of changes made the process of detection automated and specified. Also, the work by K. Prete et al. [12] have proposed an alternative method for detecting the source code changes using the templates. The major bottleneck of this process is to separate the workable templates from the templates, which does not defer any functionality. In order to improve this process, M. Kim et al. [13] proposed a logical separation of the templates using querying the construction of the code.

Furthermore, all the bottlenecks of the existing works are summarized and analysed by P. Weissgerber et al. [14]. This work takes up the recommendations and frames the generic scopes for improvements in the next section of the work.

III.   IDENTIFICATION OF SCOPE FOR IMPROVEMENTS

Furthermore, with the detailed understanding of the refactoring process outcomes by various research attempts and the strong connection with the change detection with test case management, in this section of the work, the research problems are identified.

Based on the outcomes of the parallel researches, the following short comings are identified:

- Firstly, the general-purpose regression testing is carried out on a complete set of source code which is produced and modified time to time in the software development life cycle. Most of the instances it is been observed that the pre-configured test cases are deployed in the new version of the source code. Regardless to mention that most of the test cases are configured to test the areas where no changes are made. Hence, the optimizations of the test cases are completely ignored.

- Secondly, during the manual generation of the test cases, the identification of the high priority test cases is carried out. Most of the parallel researches depends on the pre-defined functional requirements given by the customer to decide the priority of the functional requirements and based on this available information, the priority of the test cases is decided. It is natural to understand that, due to this often the hidden and critical functionalities are ignored and as well as the test cases to validate these functionalities.

- Third, automation of the test case generation is demanding area of research for regression testing. Nonetheless, the processes are far from perfection and complete acceptability.

- Finally, defining the priority test cases depends on various factors. None of the parallel researches have demonstrated all possible combinations to evolve the optimization of test cases.

This work addresses the first problem mentioned in the work.

Henceforth, in the next section of the work, the proposed change detection algorithm is discussed.

IV. PROPOSED CHANGE DETECTION

The changes made into the source code using refactoring of the codes, must be identified for reducing the test cases or generating outline of test cases.

The proposed change detection algorithm is developed in total four parts.

Algorithm - 1: Source Code Pre-Processor (SCPP)
Step - 1. Access the repository for source code files
Step - 2. Mark the previous version of the file as V(n)
Step - 3. Mark the recent version of the file as V(n+1)
Step - 4. Identify the number of lines in the V(n) and V(n+1)
Step - 5. If V(n) >= V(n+1), then mark counter = V(n)
Step - 6. Else, mark counter = V(n+1)
Step - 7. For each line in counter
a. Remove comments
b. Apply tokenizer
c. Check for variable change
d. Check for statement change
Step - 8. *Report the pre-processed V(n+1) with the changes*

The algorithm is visualized graphically here in Fig. 3.

Algorithm - 2: Prerequisite Requirement Change Detection (PRCD)
Step - 1. Load the files as V(n) and V(n+1)
Step - 2. Accept the tokenizer report
Step - 3. Build the list of "package" and "import" statements
Step - 4. For each line
a. Detect the changes in "package" and "import" statements
Step - 5. List the inclusion of Prerequisite statements
Step - 6. List the exclusion of Prerequisite statements
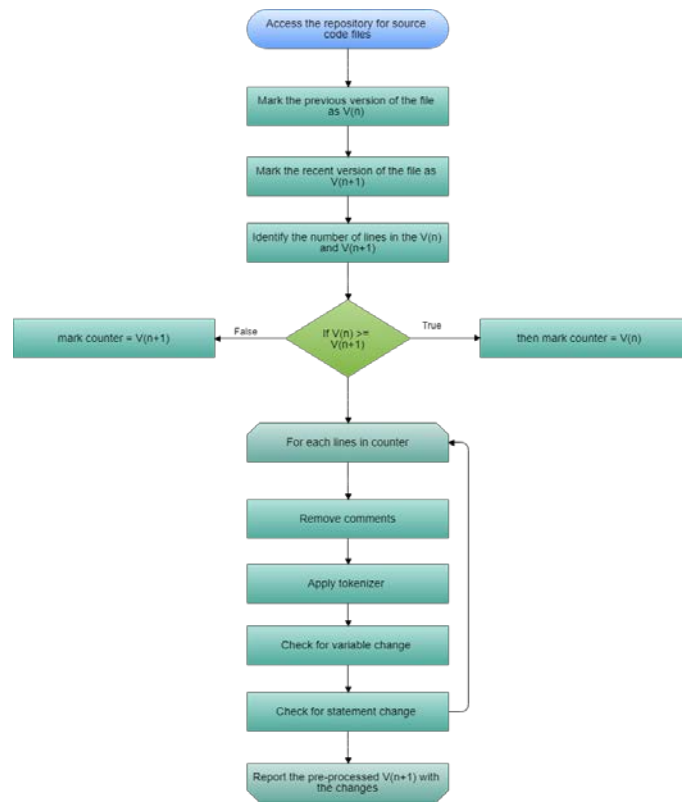
Fig. 3.   Process flow of SCPP Algorithm.

The algorithm is visualized graphically here in Fig. 4.
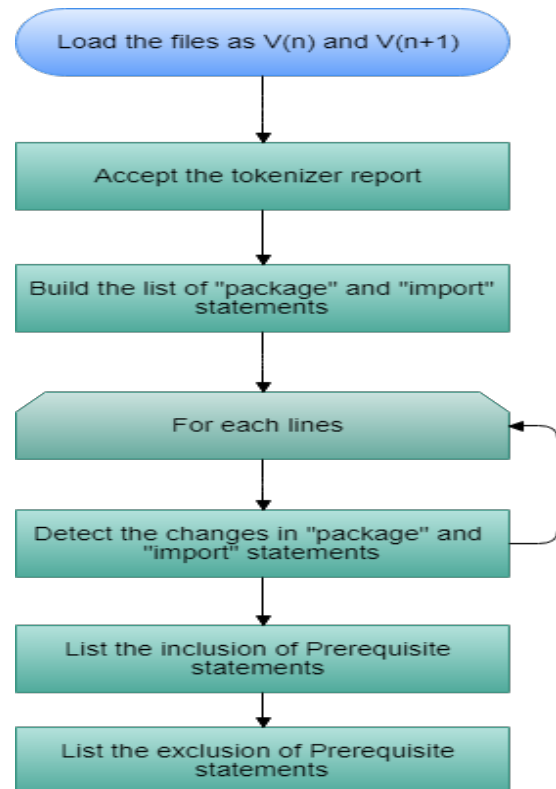
Fig. 4.   Process flow of PRCD Algorithm.

Algorithm - 3: Code Feature Change Detection (CFCD)

Step - 1. Load the files as V(n) and V(n+1)

Step - 2. Accept the tokenizer report

Step - 3. Build the list of variable identifiers

Step - 4. For each line

      a. Detect the changes in variable identifiers statements

Step - 5. List the inclusion of variable identifiers statements

Step - 6. List the exclusion of variable identifiers statements

The algorithm is visualized graphically here in Fig. 5.
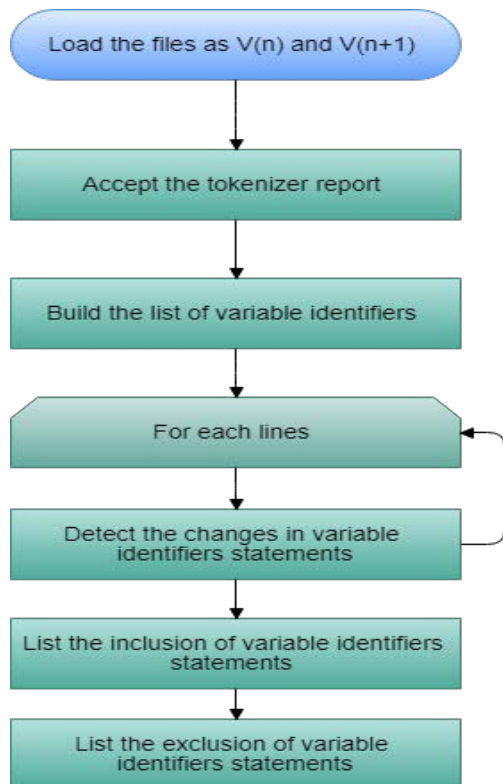
Fig. 5. Process flow of CFCD Algorithm.

Algorithm - 4: Source Functionality Change Detection (SFCD)

Step - 1. Load the files as V(n) and V(n+1)

Step - 2. Accept the tokenizer report

Step - 3. Apply programming parser on the token

Step - 4. Build the list of identified parsed token

Step - 5. For each line

      a. Detect the changes in identified parsed token statements

Step - 6. List the inclusion of identified parsed token statements

Step - 7. List the exclusion of identified parsed token statements

The algorithm is visualized graphically here in Fig. 6.
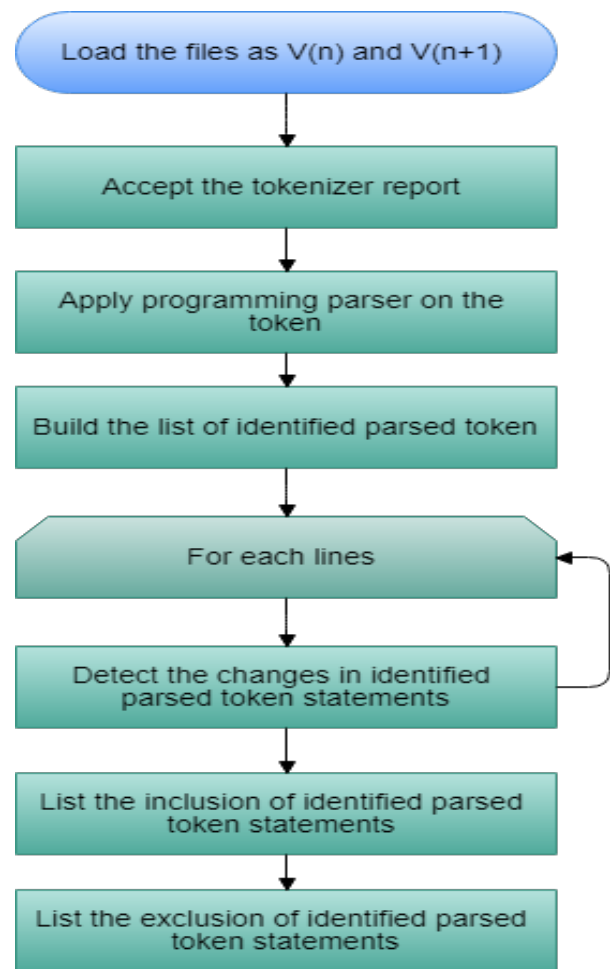
Fig. 6. Process flow of SFCD Algorithm.

Henceforth, with the detailed understanding of the proposed change recommendation algorithm, this work furnishes the test case change identification method in the next section.

V. PROPOSED TEST CASE CHANGE RECOMMENDATION

The testing is one of the most important phases in the software development life cycle. With the recent developments in software, the automation in the test cases have grown popularity. Due to the refactoring of the source codes, often the test cases are also affected. These can cause the following situations:

- Inclusion of the new test cases.

- Exclusion of the existing test cases, and.

- Removal of the duplicated test cases.

Thus, considering these factors, in this section of the work, the proposed test case change recommendation algorithm is proposed.

The algorithm is visualized graphically here in Fig. 7.

Furthermore, with the understanding of the proposed algorithms, in the next section of this work the proposed automated framework is elaborated.
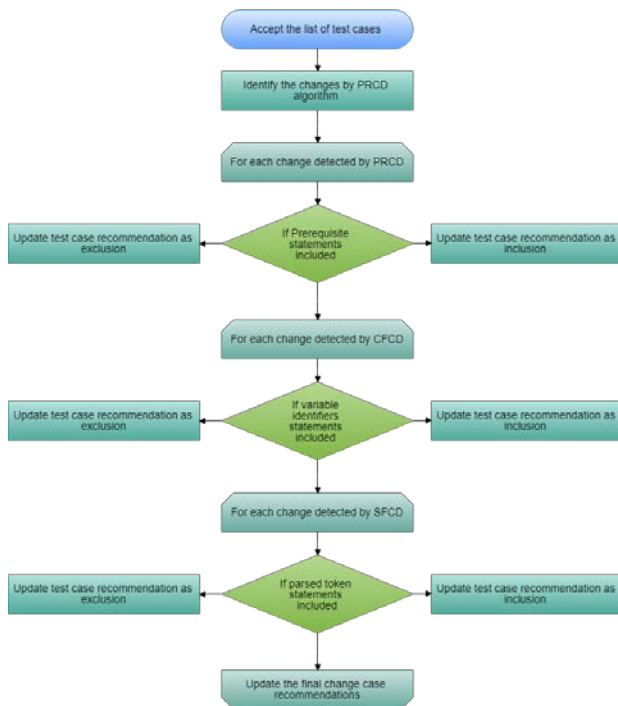
Fig. 7.   Process flow of TCCR Algorithm.

Algorithm - 5: Test Case Change Recommendation (TCCR)

Step - 1. Accept the list of test cases

Step - 2. Identify the changes by PRCD algorithm

Step - 3. For each change detected by PRCD

    a.   If Prerequisite statements included

        i.   Update test case recommendation as inclusion

    b.   Else

        i.   Update test case recommendation as exclusion

Step - 4. For each change detected by CFCD

    a.   If variable identifiers statements included

        i.   Update test case recommendation as inclusion

    b.   Else

        i.   Update test case recommendation as exclusion

Step - 5. For each change detected by SFCD

    a.   If parsed token statements included

        i.   Update test case recommendation as inclusion

    b.   Else

        i.   Update test case recommendation as exclusion

Step - 6. Update the final change case recommendations

VI. Proposed Automated Framework

In this section of the work, the proposed automated test case change recommendation framework is elaborated. The proposed framework demonstrates how different components are collaborated and coupled together for making the complete process automated (Fig. 8).
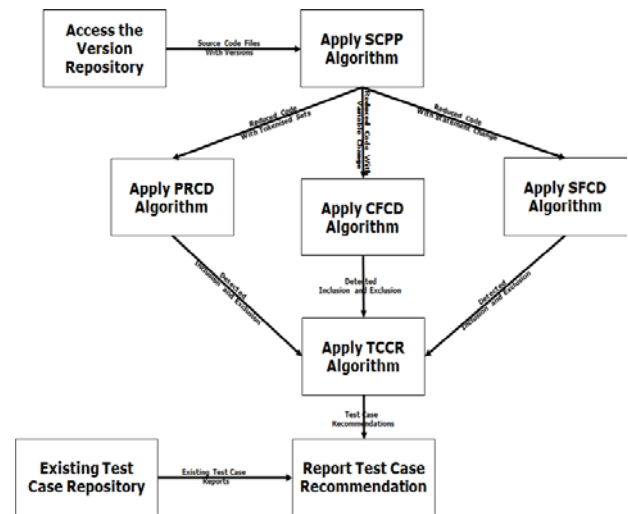
Fig. 8.   Proposed Automated Test Case Change Recommendation Framework.

The automated framework is designed to reduce the time needed for verifying and reducing or introducing test cases to the existing test case repositories.

Firstly, the source code version files are access from the location where all source codes are stored, usually called the source code repository. The source code repository is maintained by the version control tools used by any organization. This proposed framework does not apply any constraints on the version control features, rather only expects the versioning to be done only on separable source codes. After the source code files are loaded, the pre-processing algorithm is deployed on the source code to reduce the comments and to tokenize the source code files. Once the tokenization is completed, the same source code files are pushed to the proposed PRCD, proposed CFCD and proposed SFCD algorithms. The result from these algorithms are identification of pre-requisite changes, identification of feature or variable changes and identification of functionality changes, respectively. Finally, the recommendation algorithm, TCCR, generates the final recommendations based on the existing test case repository.

Further, with the detailed understanding of the complete framework work flow, in the next section of the work the results are discussed.

VII. Results and Discussion

The results obtained from the proposed automated framework is highly satisfactory and are discussed in this section of the work. Due to the highly integrated structure of the framework, the results are discussed under multiple separate factors as Experimental Setup, Pre-processor Output, Change Detection Output, Pre-Requisite Test Case Availability, Recommendation Output, Variable Test Case Recommendation Output and Functionality Test Case Recommendation Output.

A. Experimental Setup

Firstly, the experimental setup is discussed here. The primary component of the experiment relies on the Java's

"diff" utility. Diff Utilities library is an Open Source library for playing out the correlation/diff activities between writings or some sort of information: processing diffs, applying patches, creating bound together diffs or parsing them, producing diff yield for simple future showing (like one next to the other view) et cetera. The other details are discussed here in Table I.

TABLE I.   EXPERIMENTAL SETUP

Artefacts	Description
Repository Source	GitHub
Total Number of Repositories	5
Version Control Tool Used (Can be integrated with any tool)	Git
Syntax Perser	Parse Tree
Number of Iteration for Detection in each repository	10

B. Pre-Processor Output (SCPP Algorithm)

Secondly, the pre-processing outputs are listed here in Table II.

TABLE II.   SCPP ALGORITHM

Source Code Repository Name	Number of Versions Present	Number of Versions Detected	Number of Lines Present	Number of Lines Detected
Repository - 1	2	2	335	335
Repository - 2	2	2	336	336
Repository - 3	2	2	283	283
Repository - 4	2	2	332	332
Repository - 5	2	2	344	344

The result is visualized graphically here in Fig. 9.
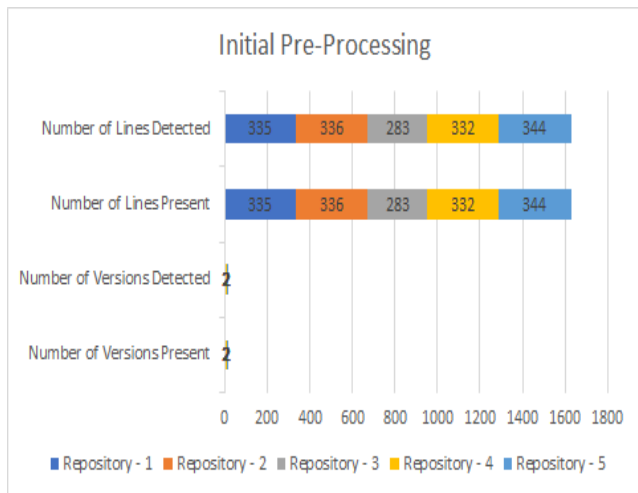
Fig. 9.   Initial Pre-Processing Phase Results.

Further, the tokenizer output is discussed in Table III.

The result is visualized graphically here in Fig. 10.

Furthermore, the comment removal phase output is discussed in Table IV.

TABLE III.   TOKENIZER OUTPUT

Source Code Repository Name	Number of Prime Tokens Present	Number of Prime Tokens Identified
Repository - 1	17	15
Repository - 2	10	8
Repository - 3	16	14
Repository - 4	15	13
Repository - 5	13	12
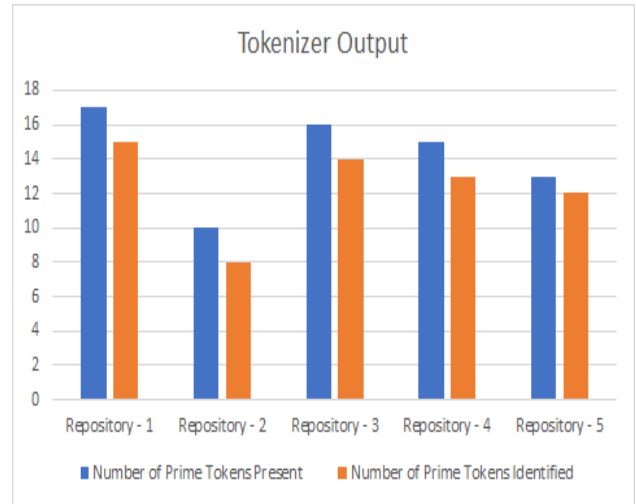
Fig. 10.   Tokenizer Phase Results.

TABLE IV.   COMMENT REMOVAL OUTPUT

Source Code Repository Name	Number of Comment Lines Present	Number of Comment Lines with Functionality	Number of Comment Lines Detected
Repository - 1	3	3	3
Repository - 2	3	2	2
Repository - 3	3	0	0
Repository - 4	11	10	10
Repository - 5	10	8	8

The result is visualized graphically here in Fig. 11.
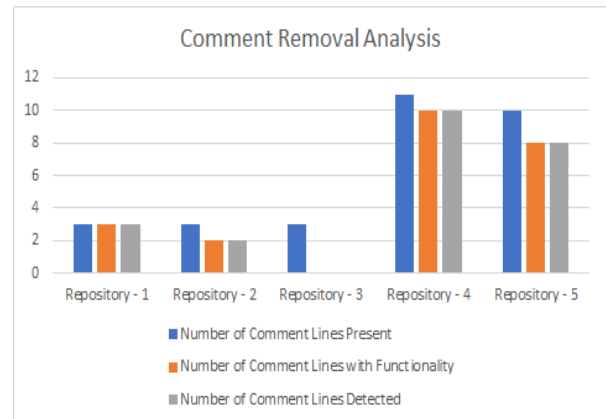
Fig. 11.   Comment Line Removal Analysis.

C. Change Detection Process Output

Thirdly, the change detection process outputs are listed here in Table V.

TABLE V. DETAILED REPORT FOR CHANGE DETECTION

Source Code Repository Name	Change Type	Change Position	Change Size
Repository - 1	Code Removed	34	0
Repository - 1	Code Removed	20	13
Repository - 1	Code Removed	5	0
Repository - 1	Code Removed	0	1
Repository - 1	Code Added	22	2
Repository - 1	Code Added	5	2
Repository - 1	Code Added	0	1
Repository - 2	Code Removed	139	0
Repository - 2	Code Removed	138	0
Repository - 2	Code Removed	134	3
Repository - 2	Code Removed	131	2
Repository - 2	Code Removed	118	12
Repository - 2	Code Removed	77	40
Repository - 2	Code Removed	76	0
Repository - 2	Code Removed	75	0
Repository - 2	Code Removed	71	3
Repository - 2	Code Removed	29	41
Repository - 2	Code Removed	26	2
Repository - 2	Code Removed	7	17
Repository - 2	Code Removed	0	6
Repository - 2	Code Added	164	1
Repository - 2	Code Added	159	4
Repository - 2	Code Added	157	1
Repository - 2	Code Added	136	20
Repository - 2	Code Added	117	18
Repository - 2	Code Added	114	2
Repository - 2	Code Added	111	2
Repository - 2	Code Added	88	22
Repository - 2	Code Added	28	59
Repository - 2	Code Added	19	8
Repository - 2	Code Added	2	15
Repository - 2	Code Added	0	1
Repository - 3	Code Removed	144	0
Repository - 3	Code Removed	143	0
Repository - 3	Code Removed	139	3
Repository - 3	Code Removed	136	2
Repository - 3	Code Removed	123	12
Repository - 3	Code Removed	82	40
Repository - 3	Code Removed	81	0
Repository - 3	Code Removed	80	0
Repository - 3	Code Removed	76	3
Repository - 3	Code Removed	34	41
Repository - 3	Code Removed	33	0
Repository - 3	Code Removed	0	32
Repository - 3	Code Added	164	1
Repository - 3	Code Added	159	4
Repository - 3	Code Added	157	1
Repository - 3	Code Added	136	20
Repository - 3	Code Added	117	18
Repository - 3	Code Added	114	2
Repository - 3	Code Added	111	2
Repository - 3	Code Added	88	22
Repository - 3	Code Added	45	42
Repository - 3	Code Added	43	1
Repository - 3	Code Added	0	42
Repository - 4	Code Removed	166	25
Repository - 4	Code Removed	164	1
Repository - 4	Code Removed	159	4
Repository - 4	Code Removed	157	1
Repository - 4	Code Removed	136	20
Repository - 4	Code Removed	117	18
Repository - 4	Code Removed	114	2
Repository - 4	Code Removed	111	2
Repository - 4	Code Removed	88	22
Repository - 4	Code Removed	45	42
Repository - 4	Code Removed	43	1
Repository - 4	Code Removed	0	42
Repository - 4	Code Added	143	0
Repository - 4	Code Added	139	3
Repository - 4	Code Added	136	2
Repository - 4	Code Added	123	12
Repository - 4	Code Added	82	40
Repository - 4	Code Added	81	0
Repository - 4	Code Added	80	0
Repository - 4	Code Added	76	3
Repository - 4	Code Added	34	41
Repository - 4	Code Added	33	0
Repository - 4	Code Added	0	32
Repository - 5	Code Removed	25	4
Repository - 5	Code Removed	22	2
Repository - 5	Code Removed	5	2
Repository - 5	Code Removed	0	1
Repository - 5	Code Added	20	13
Repository - 5	Code Added	5	0
Repository - 5	Code Added	0	1

Further, the change detection summary is presented here in Table VI.

TABLE VI. COMMENT REMOVAL OUTPUT

Source Code Repository Name	Actual Number of Changes	Number of Changes Detected	Change Detection Accuracy (%)
Repository - 1	8	7	87.50
Repository - 2	27	25	92.59
Repository - 3	23	23	100.00
Repository - 4	26	23	88.46
Repository - 5	9	7	77.78

The result is visualized graphically here in Fig. 12.
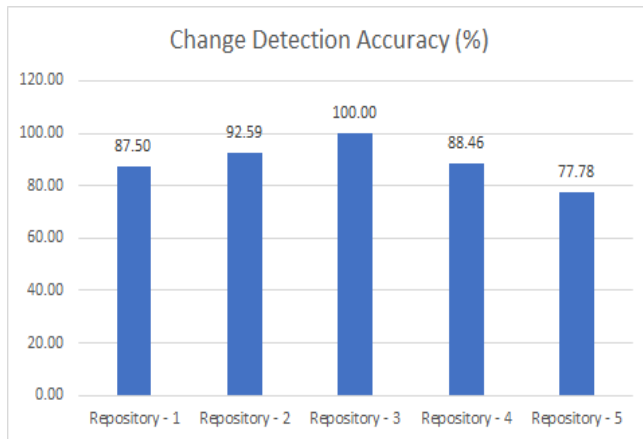
Fig. 12. Change Detection Accuracy Analysis.

D. Prerequisite Requirement Change Detection Output

Fourthly, the Prerequisite Requirement Change Detection outputs are listed here in Table VII.

TABLE VII. DETAILED REPORT FOR PREREQUISITE REQUIREMENT CHANGE DETECTION

Source Code Repository Name	Change Type	Prerequisite Details
Repository - 1	Added	import java.io.*;
Repository - 1	Added	java.util.LinkedList;
Repository - 1	Added	java.util.List;
Repository - 2	Removed	net.contentobjects.jnotify. JNotifyListener;
Repository - 2	Removed	java.io.*;
Repository - 2	Removed	java.text.SimpleDateForma;
Repository - 2	Removed	java.util.Calendar;
Repository - 2	Removed	java.util.LinkedList;
Repository - 2	Added	java.lang.reflect.Array;
Repository - 2	Removed	java.awt.Dimension;
Repository - 2	Removed	java.awt.Toolkit;
Repository - 2	Removed	javax.swing.JTextArea;
Repository - 2	Removed	javax.swing.JPanel;
Repository - 2	Removed	javax.swing.JFrame;
Repository - 2	Removed	javax.swing.JScrollPane;
Repository - 2	Added	difflib.ChangeDelta;
Repository - 2	Added	difflib.Chunk;
Repository - 2	Added	difflib.DeleteDelta;
Repository - 2	Added	difflib.Delta;
Repository - 2	Added	difflib.DiffAlgorithm;
Repository - 2	Added	difflib.InsertDelta;
Repository - 2	Added	difflib.Patch;
Repository - 3	Removed	net.contentobjects.jnotify. JNotifyListener;
Repository - 3	Removed	java.io.*;
Repository - 3	Removed	java.text.SimpleDateForma;
Repository - 3	Removed	java.util.Calendar;
Repository - 3	Removed	java.awt.Dimension;
Repository - 3	Removed	java.awt.Toolkit;
Repository - 3	Removed	javax.swing.JTextArea;
Repository - 3	Removed	javax.swing.JPanel;
Repository - 3	Removed	javax.swing.JFrame;
Repository - 3	Removed	javax.swing.JScrollPane;
Repository - 3	Added	java.lang.reflect.Array;
Repository - 3	Added	java.util.List;
Repository - 3	Added	difflib.ChangeDelta;
Repository - 3	Added	difflib.Chunk;
Repository - 3	Added	difflib.DeleteDelta;
Repository - 3	Added	difflib.Delta;
Repository - 3	Added	difflib.DiffAlgorithm;
Repository - 3	Added	difflib.InsertDelta;
Repository - 3	Added	difflib.Patch;
Repository - 4	Removed	java.util.List;
Repository - 4	Removed	difflib.ChangeDelta;
Repository - 4	Removed	difflib.Chunk;
Repository - 4	Removed	difflib.DeleteDelta;
Repository - 4	Removed	difflib.Delta;
Repository - 4	Removed	difflib.DiffAlgorithm;
Repository - 4	Removed	difflib.InsertDelta;
Repository - 4	Removed	difflib.Patch;
Repository - 4	Added	net.contentobjects.jnotify. JNotify;
Repository - 4	Added	net.contentobjects.jnotify. JNotifyListener;
Repository - 4	Added	java.io.*;
Repository - 4	Added	java.text.SimpleDateFormat;
Repository - 4	Added	java.util.Calendar;
Repository - 4	Added	java.awt.Dimension;
Repository - 4	Added	java.awt.Toolkit;
Repository - 4	Added	javax.swing.JTextArea;
Repository - 4	Added	javax.swing.JPanel;
Repository - 4	Added	javax.swing.JFrame;
Repository - 4	Added	javax.swing.JScrollPane;
Repository - 5	Added	net.contentobjects.jnotify.JNotify;
Repository - 5	Removed	java.util.LinkedList;
Repository - 5	Removed	java.util.List;

Further, the Prerequisite Requirement Change Detection summary is presented here in Table VIII.

TABLE VIII. PREREQUISITE REQUIREMENT CHANGE DETECTION SUMMERY

Source Code Repository Name	Number of Prerequisite Added	Number of Prerequisite Removed
Repository - 1	3	0
Repository - 2	8	11
Repository - 3	9	10
Repository - 4	11	8
Repository - 5	1	2

The result is visualized graphically here in Fig. 13.
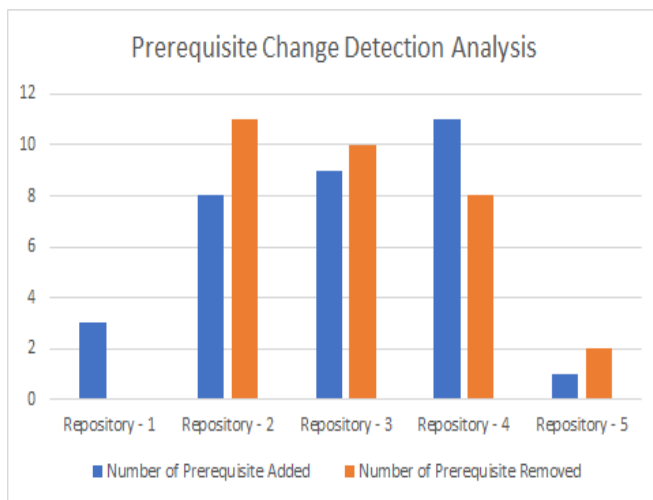
Fig. 13. Pre-Requisite Change Detection Analysis.

E. Code Feature Change Detection Output

Fifthly, the Code Feature Change Detection outputs are listed here in Table IX.

TABLE IX. DETAILED REPORT FOR CODE FEATURE CHANGE DETECTION

Source Code Repository Name	Change Type	Feature Details
Repository - 1	Remove	watchSubtree
Repository - 1	Remove	watchID
Repository - 1	Remove	res
Repository - 2	Added	N
Repository - 2	Added	M
Repository - 2	Added	MAX
Repository - 2	Added	size
Repository - 2	Added	middle
Repository - 2	Added	kmiddle
Repository - 2	Added	kplus
Repository - 2	Added	kminus
Repository - 2	Added	j
Repository - 2	Added	i
Repository - 2	Added	j
Repository - 2	Added	ianchor
Repository - 2	Added	janchor
Repository - 2	Added	static
Repository - 2	Added	newLength
Repository - 3	Remove	watchSubtree
Repository - 3	Remove	watchID
Repository - 3	Remove	res
Repository - 3	Added	N
Repository - 3	Added	M
Repository - 3	Added	MAX
Repository - 3	Added	size
Repository - 3	Added	middle
Repository - 3	Added	kmiddle
Repository - 3	Added	kplus
Repository - 3	Added	kminus
Repository - 3	Added	j
Repository - 3	Added	i
Repository - 3	Added	j
Repository - 3	Added	ianchor
Repository - 3	Added	janchor
Repository - 3	Added	static
Repository - 3	Added	newLength
Repository - 4	Added	watchSubtree
Repository - 4	Added	watchID
Repository - 4	Added	res
Repository - 4	Remove	N
Repository - 4	Remove	M
Repository - 4	Remove	MAX
Repository - 4	Remove	size
Repository - 4	Remove	middle
Repository - 4	Remove	kmiddle
Repository - 4	Remove	kplus
Repository - 4	Remove	kminus
Repository - 4	Remove	j
Repository - 4	Remove	i
Repository - 4	Remove	j
Repository - 4	Remove	ianchor
Repository - 4	Remove	janchor
Repository - 4	Remove	newLength
Repository - 5	Added	watchSubtree
Repository - 5	Added	watchID
Repository - 5	Added	res

Further, the Code Feature Change Detection summary is presented here in Table X.

TABLE X. CODE FEATURE CHANGE DETECTION SUMMERY

Source Code Repository Name	Number of Feature Added	Number of Feature Removed
Repository - 1	0	3
Repository - 2	15	0
Repository - 3	15	3
Repository - 4	3	14
Repository - 5	3	0

The result is visualized graphically here in Fig. 14.
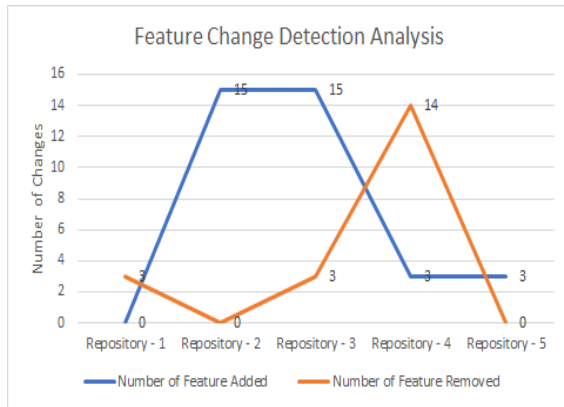
Fig. 14. Code Feature Change Detection Analysis.

F. Source Functionality Change Detection Output

Sixthly, the Source Functionality Change Detection summary is presented here in Table XI.

TABLE XI. SOURCE FUNCTIONALITY CHANGE DETECTION SUMMERY

Source Code Repository Name	Number of Functionality Added	Number of Functionality Removed
Repository - 1	7	8
Repository - 2	5	8
Repository - 3	8	8
Repository - 4	5	9
Repository - 5	5	6

The result is visualized graphically here in Fig. 15.
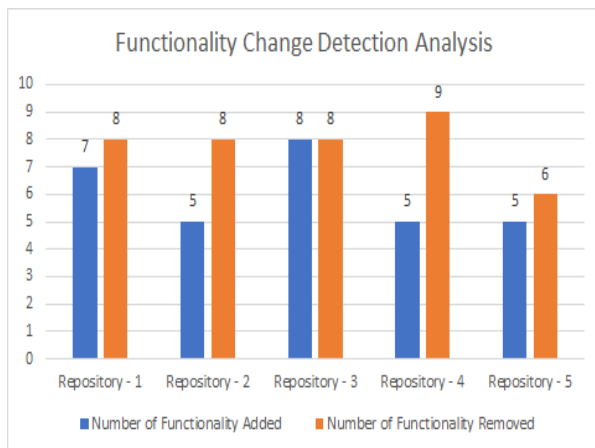
Fig. 15. Source Functionality Change Detection Analysis.

G. Test Case Change Recommendation Output

Finally, the Test Case Change Recommendation outputs are presented here in Table XII and Table XIII.

TABLE XII. SOURCE FUNCTIONALITY CHANGE DETECTION SUMMARY – INCLUSIONS

Source Code Repository Name	Prerequisite Added	Feature Added	Functionality Added	Recommendations
Repository - 1	3	1	3	Prerequisite TC Update:3 Feature TC Update:1 Functionality TC Update:3
Repository - 2	8	16	78	Prerequisite TC Update:8 Feature TC Update:16 Functionality TC Update:78
Repository - 3	9	16	78	Prerequisite TC Update:9 Feature TC Update:16 Functionality TC Update:78
Repository - 4	11	3	92	Prerequisite TC Update:11 Feature TC Update:3 Functionality TC Update:92
Repository - 5	1	3	6	Prerequisite TC Update:1 Feature TC Update:3 Functionality TC Update:6

TABLE XIII. SOURCE FUNCTIONALITY CHANGE DETECTION SUMMARY – EXCLUSIONS

Source Code Repository Name	Prerequisite Removed	Feature Removed	Functionality Removed	Recommendations
Repository - 1	0	3	4	Prerequisite TC Update:0 Feature TC Update:3 Functionality TC Update:4
Repository - 2	11	1	58	Prerequisite TC Update:11 Feature TC Update:1 Functionality TC Update:58
Repository - 3	10	3	61	Prerequisite TC Update:10 Feature TC Update:3 Functionality TC Update:61
Repository - 4	8	15	47	Prerequisite TC Update:8 Feature TC Update:15 Functionality TC Update:47
Repository - 5	2	1	1	Prerequisite TC Update:2 Feature TC Update:1 Functionality TC Update:1

Henceforth, with the complete discussions of results, in the next section, this work carries outs the comparative analysis in the next section.

VIII. Comparative Analysis

The improvements over the existing studies are the primary objective of every research and in order to justify the claim of improvements, it is must to carry out the comparative analysis. Hence in this section of the work, the comparative analysis with the popular existing works are performed on the framed metric for comparison (Table XIV).

TABLE XIV. Comparative Analysis

System Details	Change Detection Capabilities	Pre-Requisite Detection Capabilities	Feature Detection Capabilities	Functionality Detection Capabilities	Test Case Change Recommendation
M. Fowler et al. [1] 2018	Yes	No	Yes	No	No
Miryung Kim et al. [5] 2016	Yes	No	No	Yes	No
D. Silva et al. [6] 2016	Yes	No	No	Yes	No
M. Kim et al. [13] 2014	Yes	No	Yes	No	No
Proposed Automated Framework 2018	Yes	Yes	Yes	Yes	Yes

It is natural to understand that with the significant improvements and incorporation of Change Detection Capabilities, Pre-Requisite Detection Capabilities, Feature Detection Capabilities, Functionality Detection Capabilities and Test Case Change Recommendations, the proposed automated framework have outperformed the other parallel research outcomes.

IX. Conclusions

The software development industry completely relies on the accurate change management. The change driven structure or process of any organization makes it ahead of the competition among the other providers. Accommodating the client requests in terms of changes can be highly cost and time ineffective as the changes in the source code can affect the other phases of the life cycle specifically the testing. Due to any modification to the source code, the testing operations also

has to change. The challenge is to identify the current change and reduce repetition of the testing tasks. Thus, this work provides an automatic framework with Change Detection Capabilities, Pre-Requisite Detection Capabilities, Feature Detection Capabilities, Functionality Detection Capabilities and Test Case Change Recommendation for better test case managements. The major and most unique outcome of this work is to identify and recommend any changes in the test cases for making the world of software development faster and economically affordable.

References

[1] M. Fowler, Refactoring: Improving the Design of Existing Code. Addison-Wesley, 2018.

[2] E. R. Murphy-Hill, C. Parnin, and A. P. Black, "How we refactor, and how we know it," IEEE Transactions on Software Engineering, vol. 38, no. 1, pp. 5–18, 2012.

[3] N. Tsantalis, V. Guana, E. Stroulia, and A. Hindle, "A multidimensional empirical study on refactoring activity," in Conference of the Centre for Advanced Studies on Collaborative Research (CASCON), 2013, pp. 132–146.

[4] M. Kim, T. Zimmermann, and N. Nagappan, "A field study of refactoring challenges and benefits," in 20th Symposium on the Foundations of Software Engineering (FSE), 2017, pp. 50:1–50:11.

[5] Miryung Kim et al., "An empirical study of refactoring challenges and benefits at Microsoft," IEEE Transactions on Software Engineering, vol. 40, no. 7, July 2016.

[6] D. Silva, N. Tsantalis, and M. T. Valente, "Why we refactor? confessions of GitHub contributors," in 24th Symposium on the Foundations of Software Engineering (FSE), 2016, pp. 858–870.

[7] S. Negara, N. Chen, M. Vakilian, R. E. Johnson, and D. Dig, "A comparative study of manual and automated refactorings," in 27th European Conference on Object-Oriented Programming (ECOOP), 2016, pp. 552–576.

[8] J. Ratzinger, T. Sigmund, and H. C. Gall, "On the relation of refactorings and software defect prediction," in 5th Working Conference on Mining Software Repositories (MSR), 2012, pp. 35–38.

[9] G. Soares, R. Gheyi, D. Serey, and T. Massoni, "Making program refactoring safer," IEEE software, vol. 27, no. 4, pp. 52–57, 2010.

[10] S. Demeyer, S. Ducasse, and O. Nierstrasz, "Finding refactorings via change metrics," in ACM SIGPLAN Notices, vol. 35, no. 10, 2010, pp. 166–177.

[11] D. Dig, C. Comertoglu, D. Marinov, and R. Johnson, "Automated detection of refactorings in evolving components," in 20th European Conference on Object-Oriented Programming (ECOOP), 2006, pp. 404–428.

[12] K. Prete, N. Rachatasumrit, N. Sudan, and M. Kim, "Template-based reconstruction of complex refactorings," in 26th International Conference on Software Maintenance (ICSM), 2010, pp. 1–10.

[13] M. Kim, M. Gee, A. Loh, and N. Rachatasumrit, "Ref-Finder: A refactoring reconstruction tool based on logic query templates," in 8th Symposium on Foundations of Software Engineering (FSE), 2014, pp. 371–372.

[14] P. Weissgerber and S. Diehl, "Identifying refactorings from sourcecode changes," in 21st International Conference on Automated Software Engineering (ASE), 2016, pp. 231–240.

Investigating Transmission Power Control Strategy for Underwater Wireless Sensor Networks

Syed Agha Hassnain Mohsan[1]

Department of Electrical Engineering
COMSATS University Islamabad (CUI)
Islamabad, Pakistan

Hussain Amjad[2]

Department of Marine Science and Information Technology
Ocean College, Zhejiang University
Zhoushan, China

Alireza Mazinani[3]

School of Electronic and Information Engineering
Beihang University
Beijing, China

Sahibzada Adil Shahzad[4]

Department of Computer Science
National Chengchi University
Taipei, Taiwan

Mushtaq Ali Khan[5]

Marine Ecology Lab, Department of Marine Science
Ocean College, Zhejiang University
Zhoushan, China

Asad Islam[6]

School of Energy and Power Engineering
Beihang University
Beijing, China

Arfan Mahmood[7]

Complex Networks and Control Lab
Shanghai Jiao Tong University
Shanghai, China

Ahmad Soban[8]

Balochistan University of Information Technology
Engineering and Management Sciences (BUITEMS
Quetta, Balochistan

Abstract—Underwater wireless sensor network (UWSN) has proven its high stature in both civil and military operations including underwater life monitoring, communication and invasion detection. However, UWSNs are vulnerable to a wide class of power consumption issues. Underwater sensor nodes consume power provided by integrated limited batteries. It is a challenging issue to replace these batteries under harsh aquatic conditions. Thus, in an energy-constrained underwater system it is pivotal to seek strategies for improving the life expectancy of the sensors. In this paper, we propose transmission power control mechanism for underwater wireless sensor networks (UWSNs). We experimentally investigate the impact of transmission power and propose a control mechanism to enhance the performance of underwater wireless sensor network. In this proposed mechanism, source nodes will adjust its transmission power according to the location of destination node. This paper aims to provide a mechanism which is incorporated in SEEC. This study also outlines the mathematical modeling for proposed idea. Moreover, we have compared results of our scheme with previous implemented schemes.

Keywords—Component; underwater wireless sensor networks; transmission power; sensor; power consumption

I. Introduction

We have witnessed the penetration of wireless sensor networks (WSNs) into every avenue of our lives, ranging from environmental monitoring, biological medicine, national defense, underwater vehicles to marine observations [1]. A considerable research has been carried out on terrestrial sensor networks in different aspects but currently research community is attracted towards a new era of research in Underwater Wireless Sensor Networks (UWSNs). Human beings are unable to work at high pressure ocean environment for a long time. On the other hand, Terrestrial Wireless Sensor Networks (TWSNs) protocols are not feasible for UWSNs because signals transmission is elusive due to higher attenuation of aquatic environment.

Generally, UWSNs differ from TWSNs in multiple stringent design constraints such as sensors cost, dense deployment of these sensors, communication method and maximum power required for communication. The major challenges in deployments of UWSN are limited battery sources, memory and computational power. It is clear that the problem of limited battery sources is important and it is a big challenging issue for research fraternity to achieve long operation time without effecting performance of UWSN. In this new paradigm, UWSN should fulfill, despite transmission power and energy limitations, the communication requirements of several advanced technologies and manufacturing methods. Though, research community has provided several techniques to cope with these challenges due to variable characteristics of ocean but still gap lies for research.

This technology has already achieved a great stature among researchers. A growing research inclination has been noticed by industry as well as academia in UWSNs [2-3] with the

progress in underwater technologies. A key challenging issue for UWSN lifetime is the maintenance as it's difficult to replace those batteries providing power to sensor nodes. Recently, researchers are working in underwater wireless power transfer to tackle these challenges. In addition, TWSNs routing protocols cannot be implemented to UWSNs because of limited bandwidth, long propagation delay, energy constraints, high mobility and non-uniform 3D topology. Usually underwater communication is different from air due to varying conductivity, permittivity and permeability [4-5] parameters. Some researchers proved that mobile sink is an effective way to enhance throughput. In order to collect data, these mobile sinks can connect to individual sensor node or full cluster. In research study discussed in [6], authors utilized two mobile sinks to gather data from sparse regions. A SEEC protocol is proposed in [7] and we have implemented our idea on it. It is noticed when both sinks are located at adjacent positions then throughput is affected due to same sensor nodes.

Power consumption plays a vital role in efficient performance of UWSN. One major issue in power constraint of UWSN as it is unrealistic to recharge or change batteries during sensor's life span. It is not feasible to recharge these batteries frequently. Maximum power consumption occurs while data communication [8]. Power adaptation on the basis of environment conditions is a good approach towards efficient power consumption. Two major factors, wireless link quality and distance, affect the power consumption in any wireless sensor network. The link quality depends on climate conditions and physical barriers. Transmission power is controlled to obtain a good link providing successful data delivery. On the other hand, it is preferred to adopt minimum transmission power which can extend the WSN lifetime. Reliability insurance and power optimization appears as an important concern while dealing with WSN [9]. In this context, we have provided transmission power control mechanism. Reducing transmission power can put counterproductive effects to network performance. We are interested to understand better spectrum efficiency of UWSN through efficient control of transmitted power. It has the potential to minimize power consumption in UWSN. However, it still faces issues in real world deployments. Such mechanism must be robust against dynamic and complex wireless properties [10]. Through MATLAB simulation, we analyze the performance of UWSN. The obtained results of throughput, residual energy and network stability are better through power control mechanism.

In a systematic review [11], modulation techniques and channel coding is briefly discussed for discrete underwater sensor nodes. In [12], distributed algorithms are proposed to lower communication energy consumption in WSN through reducing transmission power of sensors. The appropriate selection of transmission range is discussed in [13] which greatly influences network connectivity and energy efficiency. Power consumption control problem is also briefly addressed in [14] where authors have elaborated dynamic power management.

Considering above discussion, we were inspired and implemented a power control mechanism in SEEC. The overriding objective of this study is to seek an intelligent use of transmitted power. We have implemented a previous routing scheme SEEC with incorporating a power control mechanism. We have compared our obtained results with SEEC, DBR, EEDBR and TESM. We have organized our paper as: in Section 2, we have made relevant research discussion. Section 3 will focus on transmission power control strategy. While results discussion has been carried out in Section 4. In last sections, we have presented performance tradeoff, conclusion and future work.

II. RELEVANT RESEARCH WORK

Since many years' researchers have been focusing to design several UWSNs protocols for qualitative and effective research analysis. Research community is working with the aim of paving a way for new era of underwater actuation and monitoring applications. UWSNs will help us to know our ocean with its envisioned landscape of applications. Different topologies, power control mechanisms, power management algorithms and adaptive techniques have been implemented in Wireless Sensor Networks (WSNs). Chincoli et al. [15] have presented a self-learning power control system for WSN. A topology control technique for efficient energy consumption is proposed in [16] based on power management and power control. A dynamic power management technique to enhance energy efficiency is discussed in [17]. Researchers have suggested various routing scheme to achieve better results as authors in [18] highlighted capabilities and limitations on the basis of a survey on UWSN with data aggregation. Feng Pan et al. [19] presented an energy-balanced algorithm based on energy level partition and depth threshold. In this protocol, mobile sinks can collect data at sparse region and they implied clustering on dense regions. Wan et al. [20] suggested an energy efficient scheme based on multilevel cluster selection criteria which aim to choose cluster head according to high residual energy and less radius from sink. An energy efficient routing protocol based on clustering is simulated in [21] for UWSNs. Authors proposed that clustering technique is useful to reduce energy resources consumption for a sensor network. Ma Ming et al. [22] have used a hybrid approach to maximize network lifetime and minimize the energy holes. Researchers have proposed a routing protocol AEDG which involves effective data collection through AUV [23]. In this research study, authors maximized the network lifetime in AEDG as AUV is sent to gather data from gateway nodes. The usage of multiple mobile sinks to enhance network lifetime is discussed in [24]. Authors validated that a better network lifetime can be achieved through assigning a predefined path to mobile sinks for movement.

III. TRANSMISSION POWER CONTROL STRATEGY

An efficient use of power consumption in UWSNs appears as maximum throughput and enhanced lifetime. In this study we have considered routing scheme SEEC and incorporated with transmission power control mechanism. We achieved good results which will be discussed in next sections.

A. Network Model

We consider a homogenous network model in this work. In addition, location information of all sensor nodes is given to each sensor node in this network. In our proposed idea, each sensor node forms four concentric circles around it. These

concentric circles represent different power levels. As source node maintains the location information of destination node so it will adjust the transmission power which will be in either of these concentric circles. We have discussed the concentric circles formation process in mathematical modeling section. Network model design of our proposed mechanism is shown in Fig. 1.

B. Mathematicl Modeling

In UWSN, approximately 0:68 dBm power is required for sending packets up to distance of one meter. We consider R_{sn} as the range of sensor node. The maximum transmission power for each sensor is calculated as:

$$Trans_{pwr} = I * R_{sn} \qquad (1)$$

Here we take I = 0:68, as we have divided the total transmission power into four different levels so we can calculate the radius of each concentric circle around the sensor node from following equation:

$$R_{dist} = i * (R_{sn} / 4) \qquad 1 \le i \le 4 \qquad (2)$$

We can calculate transmission power for calculated four radii as follow:

$$Trans_{pwr\,i} = I * R_{dist} \qquad (3)$$

If we consider the transmission range of the sensor node R_{sn} as 50 meters then Table I shows the calculated radii and power magnitude at each radius.

We have incorporated the proposed transmission power control mechanism in SEEC and MATLAB simulation results will be discussed in next section.
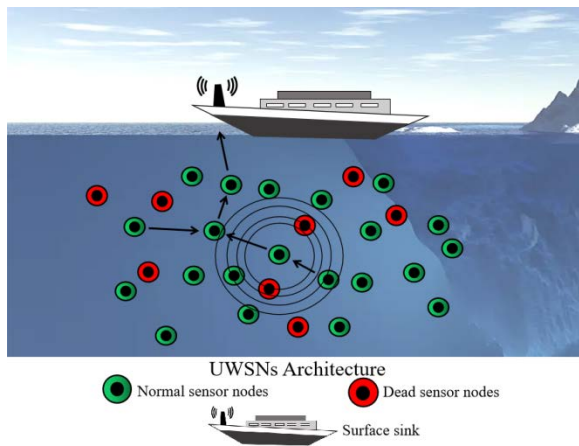
Fig. 1.   Network Model.

TABLE I.   RESEARCH KEYWORDS

R_{dist} (Meter)	Power (dBm)
12.5	8.5
25	17
37.5	25.5
50	34

IV. SIMULATION RESULTS

A. Network Lifetime

We have noticed an improvement in SEEC's network lifetime after incorporating our proposed idea of controlled transmission power mechanism. In SEEC, network was dead early as sensor nodes were sending packets with a predefined transmission power value.

While in controlled transmission power mechanism, source node adapts the transmission power on the basis of Euclidean distance from the destination node. Whenever any packet is transmitted, every time consumption power will vary. We can see a clear increase in network lifetime. In Fig. 2, nodes in SEEC were dead early while all nodes were alive till 1100 rounds after incorporating controlled mechanism.
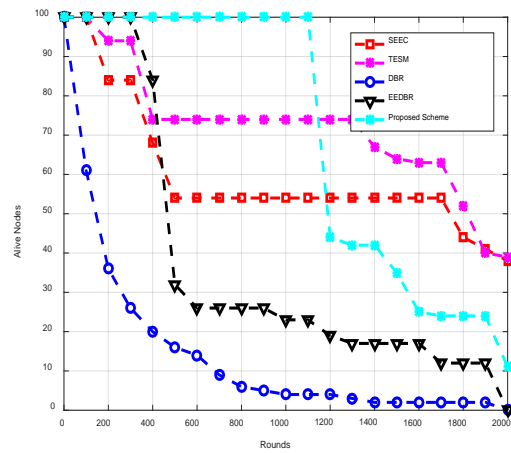
Fig. 2.   Network Lifetime.

B. Stability Period

We cannot obtain same number of sparse and dense region each time through randomly deploying sensor nodes. Network stability is compromised whenever dense regions are maximums and sensor nodes communicate with same transmission power within these dense regions. On the other hand, network stability period is improved if we adapt controlled transmission power mechanism as shown in Fig. 3. In controlled mechanism, besides controlled transmission power, stability period is better than SEEC, TESM [25], DBR and EEDBR through clustering in dense regions and mobile sinks movements in sparse regions.

C. Network Residual Energy

The transmission power factor was constant in energy consumption equation of SEEC. While sending a packet, same energy was being consumed without considering distance between two communicating nodes. In controlled transmission power mechanism, transmitted power adjustment is made on the basis of distance between source and destination node. Low transmission power in needed in case of less distance. Low requirement of transmission power results into low energy tax providing higher network residual energy. In our proposed work, energy consumption varies its return value after transmitting the packet which also gives higher network

residual energy. Fig. 4 shows the network residual energy for proposed scheme along with SEEC, DBR, EEDBR and TESM.

D. Network Throughput

In Fig. 5, the received packets per round at any sink are more than SEEC, DBR, EEDBR and TESM. We can notice an increased throughput in our proposed work with controlled transmission power mechanism. In SEEC, less packets are received as mobile sinks movement is according to sparse region and in TESM it is based on selection criteria opted for forwarding nodes.

Fig. 6 shows that the total received packets at sink are higher with controlled mechanism in SEEC. It is because of increased network lifetime by using transmission power control mechanism. EEDBR receives packets at the cost of high residual energy. In SEEC, total received packets are lower than our proposed scheme. It is because in SEEC, network is dead early and predefined transmission power also remains the same. It can be seen in Fig. 5 that received packets per round are better in our proposed controlled mechanism than SEEC, TESM, DBR and EEDBR.
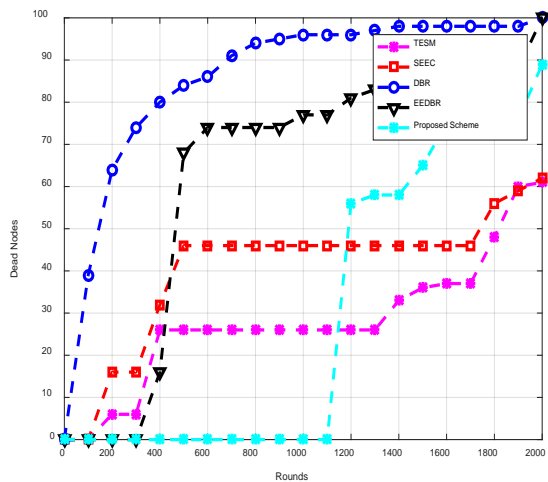
Fig. 3.   Network Stability Period.
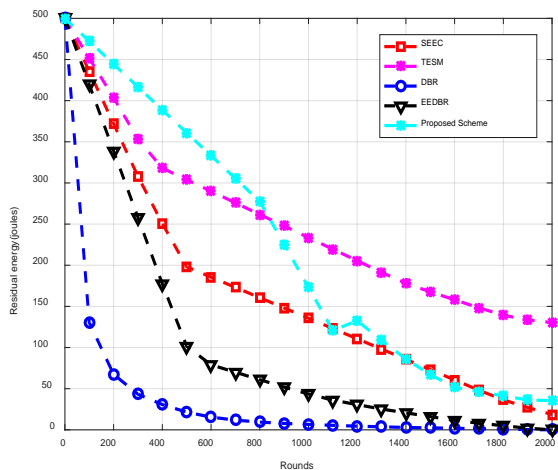
Fig. 4.   Network Residual Energy.
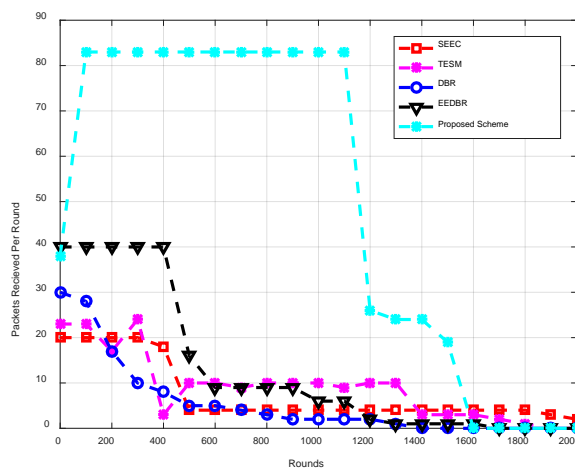
Fig. 5.   Packets Received Per round.
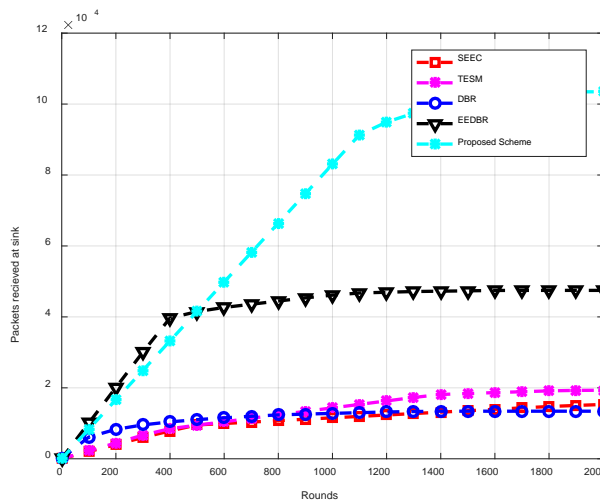
Fig. 6.   Packets Received at Sink.

V.   PERFORMANCE TRADEOFF

We have achieved better network lifetime, throughput and network residual energy. We can clearly see that more nodes are alive while dead nodes are less. Thus, we conclude a considerable progress of our work in aspects of stability period, residual energy and throughput by incorporating controlled transmission power mechanism.

VI.   CONCLUSION AND FUTURE WORK

Results presented in this work demonstrate that optimum and efficient use of transmission power in mobile sinks is crucially important in enhancing network lifetime and throughput. Our work has led us to conclude that a proper control mechanism for power transmission is pivotal. In addition, the results of any routing scheme can be improved if we monitor the proper use of resources in the protocol. In this study, we compared results of power control mechanism incorporated in SEEC with previous implemented schemes TESM, DBR and EEDBR. To sum up our work, the proposed idea is better in throughput, network lifetime and residual energy. Taking complex underwater environment and economic concerns into consideration, we conclude that

UWSN should be capable to adjust itself according to environment conditions. Thus, we can achieve this through proper protocol, power control and topology design which is self-adaptive according to the changing environment conditions. However, these design characteristics should take energy efficiency into account which is crucially essential for normal operation of UWSN. In future, we are aiming to consider sparsity control in order to improve network lifetime. Holding time phenomenon towards reducing packet drop ratio is deferred to future work.

REFERENCES

[1] Luan, Zhifu. "Calculation and Simulation of Transmission Reliability in Wireless Sensor Network Based on Network Coding." International Journal of Online and Biomedical Engineering (iJOE) 13.12 (2017): 150-161.

[2] Li, N.; Cürüklü, B.; Bastos, J.; Sucasas, V.; Fernandez, J.A.S.; Rodriguez, J. A Probabilistic and Highly Efficient Topology Control Algorithm for Underwater Cooperating AUV Networks. Sensors 2017, 17, 1022.

[3] Li, N.; Martínez, J.F.; Meneses Chaus, J.M.; Eckert, M. A Survey on Underwater Acoustic Sensor Network Routing Protocols. Sensors 2016, 16, 414.

[4] Hassnain, S. A., M. J. Mughal, and Q. A. Naqvi. "Layered Chiral Spheres with Zero Backscattering." 2019 Photonics & Electromagnetics Research Symposium-Fall (PIERS-Fall). IEEE, 2019.

[5] Hassnain, S. A., M. J. Mughal, and Q. A. Naqvi. "Analysis of Effective Medium Parameters on Polarizability of Homogeneous Chiral Sphere." 2019 Photonics & Electromagnetics Research Symposium-Fall (PIERS-Fall). IEEE, 2019.

[6] Coutinho, Rodolfo WL, et al. "Underwater wireless sensor networks: A new challenge for topology control–based systems." ACM Computing Surveys (CSUR) 51.1 (2018): 1-36.

[7] Sher, A.; Javaid, N.; Azam, I.; Ahmad, H.; Abdul, W.; Ghouzali, S.; Niaz, I.A.; Khan, F.A. Monitoring square and circular fields with sensors using energy-efficient cluster-based routing for underwater wireless sensor networks. Int. J. Distrib. Sens. Netw. 2017, 13, 1550147717717189.

[8] Khemapech, Ittipong, Alan Miller, and Ishbel Duncan. "A survey of transmission power control in wireless sensor networks." Proc. PGNet (2007): 15-20.

[9] Lavratti, Felipe, et al. "A transmission power self-optimization technique for wireless sensor networks." ISRN Communications and Networking 2012 (2012).

[10] Fu, Yong, et al. "Practical control of transmission power for wireless sensor networks." 2012 20th IEEE International Conference on Network Protocols (ICNP). IEEE, 2012.

[11] Mehedi, Syed Agha Hassnain Mohsan Md, et al. "A Systematic Review on Practical Considerations, Recent Advances and Research Challenges in Underwater Optical Wireless Communication." IJACSA 11-7 (2020).

[12] Chen, Xiao, and Neil C. Rowe. "Saving energy by adjusting transmission power in wireless sensor networks." 2011 IEEE Global Telecommunications Conference-GLOBECOM 2011. IEEE, 2011.

[13] Gao, Mingsheng, Chuan Heng Foh, and Jianfei Cai. "On the selection of transmission range in underwater acoustic sensor networks." Sensors 12.4 (2012): 4715-4729.

[14] Mokrenko, Olesia, et al. "Dynamic power management in a wireless sensor network using predictive control." IECON 2014-40th Annual Conference of the IEEE Industrial Electronics Society. IEEE, 2014.

[15] Chincoli, Michele, and Antonio Liotta. "Self-learning power control in wireless sensor networks." Sensors 18.2 (2018): 375.

[16] Nickray, Mohsen, Ali Afzali-Kusha, and Riku Jäntti. "Simultaneous power control and power management algorithm with sector-shaped topology for wireless sensor networks." EURASIP Journal on Wireless Communications and Networking 2015.1 (2015): 118.

[17] Barceló, Marc, et al. "Joint routing, channel allocation and power control for real -Transactions on sensor network Emerging Telecommunications Technologies 26.5 (2015): 945-956.

[18] Goyal, Nitin, Mayank Dave, and Anil K. Verma. "Data aggregation in underwater wireless sensor network: Recent approaches and issues." Journal of King Saud University-Computer and Information Sciences 31.3 (2019): 275-286.

[19] Feng, Pan, et al. "Improved energy-balanced algorithm for underwater wireless sensor network based on depth threshold and energy level partition." EURASIP Journal on Wireless Communications and Networking 2019.1 (2019): 1-15.

[20] Wan, Z.; Liu, S.; Ni, W.; Xu, Z. An energy-efficient multi-level adaptive clustering routing algorithm for underwater wireless sensor networks. Clust. Comput. 2018, 1–10.

[21] Huang, Xiangdang, Shijie Sun, and Qiuling Yang. "Data Uploading Strategy for Underwater Wireless Sensor Networks." Sensors 19.23 (2019): 5265.

[22] Ma, Ming, and Yuanyuan Yang. "SenCar: An energy-efficient data gathering mechanism for large-scale multihop sensor networks." IEEE Transactions on Parallel and Distributed Systems 18.10 (2007): 1476-1488.

[23] Khan, Jawaad Ullah, and Ho-Shin Cho. "A distributed data-gathering protocol using AUV in underwater sensor networks." Sensors 15.8 (2015): 19331-19350.

[24] Chen, Y.S.; Lin, Y.W.Mobicast routing protocol for underwater sensor networks. IEEE Sens. J. 2013, 13, 737–749.

[25] Yahya, Aqeb, et al. "Towards efficient sink mobility in underwater wireless sensor networks." Energies 11.6 (2018): 1471.

Attendance System using Machine Learning-based Face Detection for Meeting Room Application

Rahmat Muttaqin[1]
Telematika, Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Bandung, Indonesia

Nopendri[2]
School of Industrial and System Engineering
Universitas Telkom, Bandung, Indonesia

Syifaul Fuada[3]
Universitas Pendidikan Indonesia
Bandung, Indonesia

Eueung Mulyana[4]
School of Electrical Engineering and Informatics
Institut Teknologi Bandung, Bandung, Indonesia

Abstract—In a modern meeting room, a smart system to make attendance quickly is mandatory. Most of the existing systems perform manual attendance, such as registration and fingerprint. Despite the fingerprint method can reject the Unknown person and give the grant access to the Known person, it will take time to register first a person one-by-one. Moreover, it is possible to create long queues for fingerprint checking before entering the meeting room. Machine learning, along with the Internet of Things (IoT) technology is the best solution; it offers many advantages when applied in the meeting rooms. Generally, the method used is to create a presence by detecting faces. In this paper, we present a facial recognition authentication based on machine learning technology for connection to the meeting rooms. Furthermore, specific website to display the detection result and data storage design testing is developed. The method uses 1) the Dlib library for deep learning purposes, 2) OpenCV for video camera processing, and 3) Face Recognition for Dlib processing. The proposed system allows placing the multiple cameras in a meeting room as needed. However, in this work, we only used one camera as the main system. Tests conducted include identification of one Known person, identification of one Unknown person, identification of two people, and three people. The parameter to be focused is the required time in detecting the number of faces recorded by the camera. The results reveal that the face can be recognized or not recognized, then it will be displayed on the website.

Keywords—*Detection; face; IoT; meeting room; attendance*

I. INTRODUCTION

In this industry 4.0 era as nowadays, the development of machine learning technology is growing rapidly. Research in this area and its use-case in various fields have been demonstrated widely [1]. By using machine learning, the devices can perform separate types of learning, such as image [2-3], voice, shortest path, video pattern recognition, fault in power transmission systems [4], patterns in communication systems [5], etc. Besides, there are also developments in the Internet of Things (IoT) technology that utilize internet connections for data transfer without human intervention. This technology is usually applied to communication on microcontroller devices to send sensor data, gateways in the form of single board computers, and clouds [6]. Machine learning can be incorporated with IoT to build an intelligent system, one of them is a real-time attendance system. This technology can be a solution to streamline existing systems where most of the attendance systems still utilize the fingerprint method (including RFID-based machines). Indeed, it consumes a lot of time, especially at the personal registration and fingerprint checking (check-log) stages [7-11]. With machine learning, a person's or the people's presence in a meeting room can be detected quickly and accurately. This learning machine requires a lot of data to perform excellent detection.

Another technique like the Histogram Oriented Gradients (HOG) actually can be used as a method for face recognition application [12-15]. Facial recognition is a method for character recognition on faces that are successfully detected. To get better results compared to the method of skin color-based face detection; it can be performed by using a combination of Haar Cascade & Binary Pattern [16], or Haar-like & Adaboost features [17]. The detected objects are then classified into three types orientation based on the object's motion: horizontal, vertical, and rotational. But Machine learning method is able to provide various advantages over traditional methods when trained using large amounts of data. The more data, the better the results. For facial recognition application, we can employ two commonly used open-source libraries, i.e., Dlib and OpenCV [18]. Dlib is created using C++ language, which includes tools used for machine learning [19]. In addition, Dlib offers linear algebra where the central core uses the Basic Linear Algebra Subprogram (BLAS) [19].

Face detection is one area of research that has great potential for further development. By far, there have been many studies on Machine Learning-based face detection, the most recent research is reported by *S. Khan*, et al. [20]. Their system has successfully detected 2 to 12 people within the classroom and got 100% accurate detection results. Even so, we still want to explore this topic to get the most optimal performance facial recognition, where the system can work in real-time, which is then applied to meeting rooms. The faster the face detection, the more efficient the attendance process. Since [20] focused on the detection accuracy and the time required for detection was not explicitly presented; therefore, for this work, we will more emphasize on detection speed.

In this work, the system limitations are determined as follows: A node contains a single PC and a single camera. In the system implementation, the server and node have resided in the same PC. We limit the dataset in the form of an image under 1600 x 1500 pixels. Similar to [20], we only analyze the detection between the Unknown and Known person(s). The Unknown data used for every detected person is determined one Unknown without specific ID, it means there is the same ID for all Unknown people. The detection method employs only HOG and convolutional neural network (CNN).

Four sections compose this paper. First is Introduction that elaborates the related work and research motivation, 2) Methods discusses the intelligent attendance system for meeting room including the division of subsystems to be developed, 3) Results and Analysis, and the last section is 4) Conclusion.

II. METHODS

This system was developed under the Linux OS Ubuntu Desktop 18.04 LTS. The smart meeting room consisted of three subsystems, namely, Back-End, Front-End, and Node as illustrated in Fig. 1.

Each subsystem has a specific purpose. In this system, data is first obtained from the camera to the node. Local stream uses Real Time Streaming Protocol (RTSP). The node will process the data to be recognized by all the datasets existed on the node. After processing, the data is sent to the server (Back-End) to be saved to the database. Then Front-End will request data at specific intervals to the Back-End for real-time data up to 10 seconds before. The data will be displayed under the Back-End. To support IoT, the HyperText Transfer Protocol (HTTP) was utilized as a communication protocol between Client to Front-End.

A. Front End Design

In general, Front-End functions are for: 1) Requesting data to the Back-End by using the RESTful API; 2) Displaying data and images obtained from Back-End; and 3) giving the grant access to user/admin to the system. This section used the Vue.js framework based on the Javascript programming language. On Linux OS, the platforms that must be available are NPM and Node.js, which can be installed using the Ubuntu package manager. Some commands were carried out first until finally, it produced a folder called `dist`, which contained an `index.html` file. The file can be installed on the webserver to be accessed.

The main modules used are 1) Axios, used for RESTful API requests in the form of an HTTP POST or GET methods, 2) Vue-Router, used for routing or addressing from the display, 3) Vuetify, used for material design, and 4) Vuex as state management. The of Front-End's work-flow diagram can be seen in Fig. 2. The flowchart initially shows a list of available rooms. When a room is clicked or selected, the person/face data in the database will be displayed. A more detailed explanation will be explained in the next paragraph.

As in Fig. 2, first step is the Dashboard requests a list of rooms by doing an HTTP GET request to the Back-End part to the following link http://[IP:port]/attendance/room/. IP and port are matched with server address where Back-End is installed. The response received by the Front-End is in the form of data in JavaScript Object Notation (JSON) format, i.e., {"listdata":[{"id":1,"name":"Kelas A"},{"id":2, "name":"Ruang Rapat"}]}. Later, the data is parsed to take the ID and name parts. For example, to retrieve the ID data: ["listdata"][0]["id"]. If the room data has been obtained, the list will be displayed on the dashboard using v-for and v-bind: href tags. Later, the dashboard display is appeared; when the room name is clicked, the system will request the Back-End server, which is a list of people within the room. The data is requested using HTTP POST to the following address: http://[ip]:[port]/attendance/currentposition/.

A JSON formatted parameter that is {"location": [room_id]} must exist and be set in the body of HTTP requests. Id_ruangan is obtained on previous requests. For example, the data obtained is in the form of codes as below, where the data is a list of JSON: {"listdata:[{"person_id":1,"person__name":"syifaul","timestamp":1568689950.1496825,"profile_image_link":"/storage/image/profile/0000001_c.jpg"},{"person_id":2,"person__name": "tq","timestamp":1568699955.1496825,"profile_image_link": "/storage/image/profile/0000002_c.jpg"}]"}. If we want to take the link of profile image data, it can be achieved by the following way: data["listdata"][0]["profile_image_link"]. This data is used to display images of people present which are appeared on the dashboard.
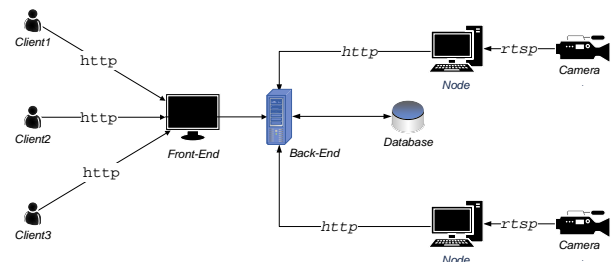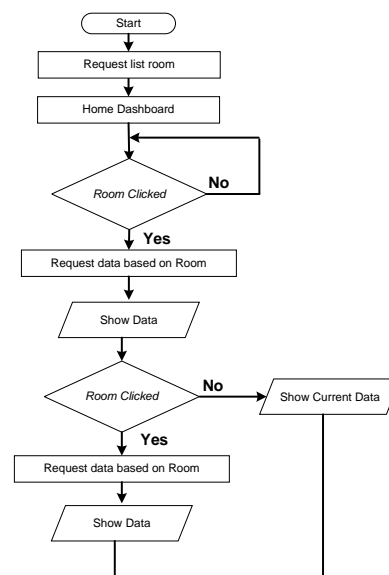

Fig. 1. Smart Meeting Room System Architecture.

Fig. 2. Flowchart of Front-End.

B. Back End Design

In general, the functions of Back-End are: 1) Providing an API to be accessed by Front-End in requesting data, 2) Providing an API used by nodes to send and store captured data, 3) Saving data to the database as seen in Table I, and 4) Registering the user names and rooms.

TABLE I. DATA BASE STRUCTURE

Name of Table	Function	Components
Person	Contains users that are recognized by the node.	ID: Integer Name: String (Length limit 40) Profile_image: Image File
Location	place / location of the camera is	ID: Integer Name: String (40)
Camera	Camera ID	ID: Integer Location: Foreignkey (Location)
Position	A collection of users detected, both registered and not	ID: Integer Person: Foreignkey (Person) Camera: Foreignkey (Camera) Location: Foreignkey (Location) Timestamp: Float Profile_image_link: String

In this work, the Back-End employs the Django v2.2 Framework, so the programming language is Python3.6. However, on Ubuntu 18.04 version, Python3.6 is already installed. Therefore, we do not need install Python again. To register, we need access to Django Admin which can be traced using a web browser to the following address: http://[ip]:[port]/admin. The first step is a request to enter the superuser's username and password that has been created. If it has been filled-in and submitted, the dashboard of Back-end admin will appear. On the dashboard, the superuser can register person, location and camera. However, registering a camera requires a room parameter. Herein, before registering it, we need to register the room first.

In the Back-End, there are several APIs that function as a communication tool with other parts, namely Login, Refresh, Position, and Current Position. Each function is described as follows:

- *Login*, which is used to get tokens in the form of JWT format, which are then used to access several available APIs. The method used is HTTP Post, the body is JSON with the following format: {"username" "admin","password" "bandung"}. Username is registered in the create superuser menu or in the Django admin window. In the response section, there are two keywords, namely refresh and access. Access is used in subsequent API requests, whereas refresh is for requesting a new token or re-requesting the token if the access has expired. The token duration before being refreshed is about 5 minutes. Meanwhile, the use of a refreshed token can be up to 1 day before the validity period expires.

- *Position*, which is used to transmit a person's data detected by the camera. A person is a list of user IDs in the dataset. Location is the location's ID where the user was detected, the camera is the ID of the camera used, and the timestamp is the time when the data was sent to the server.

- *The Unknown*, which is used to send data to the server if the camera detects an Unknown person. An Unknown person is a person who is not registered in the existing dataset. Each Unknown ID has a face photo that is sent to the server. The face photo data is converted to base64. When received by the server, the photo data is decoded into a JPG file format. The image is then saved in the storage media folder, which is useful as a link for the detected Unknown ID profile image.

- *Current Position*, which is used to get a list of users who are currently in the meeting room, even some time ago. The data taken for the past time can be adjusted according to the need, usually the time is set 2 to 300 seconds ago.

C. Node Design

In general, the functions of the node are: 1) Face coding, which is converting the image data sets in the provided data set into a 128-d vector, 2) Face capturing, which detects the location of face captured by the camera, 3) Face recognition, which is finding the owner face, and 4) Sending data to the server. Two libraries that are needed for this purpose: Python library and NVDIA CUDA.

Phyton library consists of Dlib and OpenCV: the libraries that provide tools for machine learning and image processing, respectively. Later, Face_recognition, a library that simplifies the use of Dlib. Meanwhile, NVIDIA CUDA is used as a driver to utilize the NVIDIA graphics card core in machine learning processing.

Face Encoding is the process of classifying faces into 128-d vectors (128 real numbers) that represent faces Known to the computer, as exhibited in Fig. 3. The classification is done using previously trained network (pre-trained network) using approximately three million images in the Dlib library, so we only need to use the pre-trained network. In the trainer, there is a configuration that must be set in the `config.ini` file, as follows:

```
File faceconfig.ini
[Trainer]
dataset_folder= dataset
encodings_output=
trainer_output/djangodata.pickle
detection_method=cnn
```

These parameters are filled depending on the needs. The detection method can use CNN or HOG. Dataset_folder is a folder containing the datasets in the form of images collection. The image folder can be different depending on the image identity (in the form of the name or the personal identity concerned in the image or photo). For example, in folder A contains photos of A, in folder B contains B, in folder C contains photos of C, and so on. These folders are all stored in a folder called dataset. The dataset folder is placed in the root folder so that the settings are written as follows dataset_folder = dataset. While Encodings_output is the output file encoded in this process.

In Fig. 3, initialization is processed at the beginning, such as determining the location of the dataset, encoding the output

location, the face detection method, etc. For the face detection method, CNN is used by using the Dlib library, which is set using an NVIDIA CUDA graphics card. Thus, the processing is more accurate and faster.

For each detected image, a name for the image folder will be taken and named according to the detected person's name. Then the image color changes from RGB OpenCV to RGB Dlib format. Afterward, the face detection was carried out using CNN. The detected faces are marked by making a rectangle around the face which is then converted into 128 real numbers. For each encoding of that one image, the existing encoding list is updated with matching names. The resulting encoding and name are made into a JSON model with the given format: {"encodings": knownEncodings, "names": knownNames}. KnownEncodings is a list of detected encoding (128-d vector), whereas knownNames is a detected name. Each encoding and name detected at one time occupies the same list element. The JSON result is written into a file with a pickle extension. When the format is written, all of them are converted into Hex numbers. This pickle extension file will be used to recognize faces detected by the camera.

In the Face recognition process, it requires a file with a pickle extension generated from the Face Encoding process. In this process, the `faceconfig.ini` file must be set up first, as follows:

```
Faceconfig.ini
[Detector]
camera_id= 3
location_id = 2
; time periodic for sending message to server
in seconds
periodic = 2
; time periodic for capturing/save to disk
unknown picture and send unknown data to server
in seconds
periodic_unknown = 3
; input file.pickle
encodings_input                        =
trainer_output/djangodata.pickle
; hog or cnn
detection_method=cnn
; 1 to show display or 0 to hide display
display = 1
; 1 to save video to disk, 0 not save video
save_video_to_disk = 1
;input_type can be a file,stream_local or
stream_ip
;input_type= file
input_type= stream_ip
;if input_type = file, set the video
destination file
input_video = lunch_scene.mp4
;if input_type = stream_ip, set ip
camera_address
camera_address                         =
rtsp://admin:QXJMCF@192.168.0.100
; unknown id
unknown_id = 99999999
[Server]
;username and password server
username = admin
password = bandung
server_address = 127.0.0.1
server_port = 8000
secure = 0
```
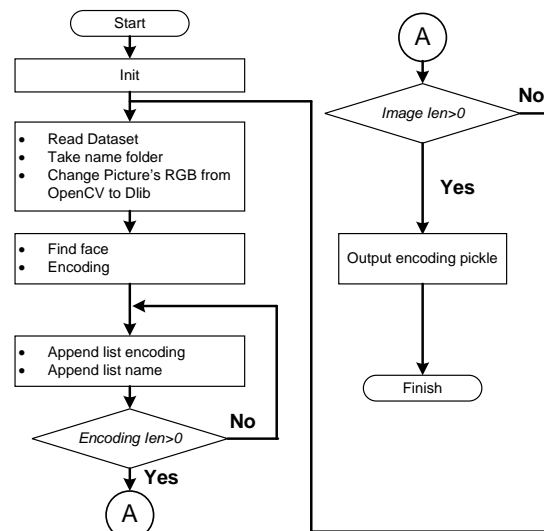

Fig. 3.   Flowchart of Face Encoding.

In the file above, the camera_id and location_id values are integers, in which the values were determined by the admin during the Back-End setup. Periodic denotes the length time required for transmission of detected data. Periodic_unknown denotes the period for sending Unknown data if the detected face is not found in the encoding file with the pickle extension. Encodings_input represents a file from the face encoding process which is used as a reference for face detection. Detection_method is used as a method for detecting faces: if CNN is selected, it can be written cnn, or hog if HOG is selected. Display is used to display video from a camera where the value 1 = yes, and 0 = no. Save_video_to_disk to save the video from the camera to a storage place, where value 1 = yes, and 0 = no. Input_type is used for input data settings. If it comes from an input file, write File, then stream_local for local stream and stream_ip for RTSP inputs. Input_video contains the address of the video input if input_type is filled with files. Camera_address represents the camera address when using stream_ip. Later, Unknown_id is the Unknown identification set on the Back-End by the admin. Username and password are created by the admin on the Back-End, where the username is used as authentication in sending files to the server. Server_address and Server_port are Back-End addresses. Secure is set to 0 if using HTTP, and 1 if HTTPS.

In Fig. 4 initialization is done at the beginning, such as specifying the location of the timeout dataset, etc.

After that in each loop, it reads the streaming video data per frame, changes the color from BGR to RGB. Accordingly, it can be read by Dlib, and reduces the reading frame size to speed up the process. The next step is face detection using CNN by utilizing an NVIDIA graphics card. Every detected face is made a kind of square or box-shaped divider to mark the face area. The marked area is encoded into a 128-d vector. The results of the encoding if there are two faces, there will be two different results. The encoding results are stored in a list. Each data in the list is compared with the encoding data that is already owned. Compare_face with the input pickle extension file can be used to perform the comparison function.
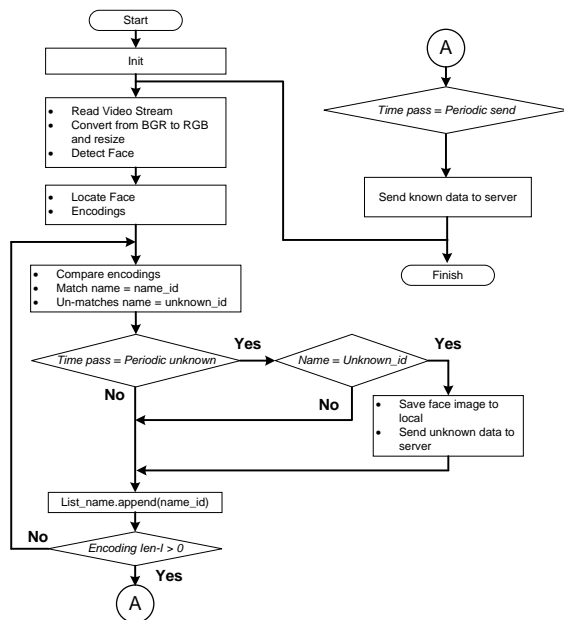
Fig. 4.   Flowchart of Face Recognition.

In the compare_face function, the tolerance parameter is set to 0.5. Every data that matches, its name or ID is stored in a name index list. In that list, a name is taken based on how often that name appears in the name index list. Only one name that comes out is put into a list again, that is the list of names containing anyone detected in a frame. If the encoding result data does not match the data in the pickle extension file, then the data is classified as Unknown. Image files (faces) retrieved from unknown data are then stored on the local drive, part of the captured data is sent to the server for display. The APIs used to send Unknown data are:

Method: POST
URL: http://ip:port/attendance/unknown/

```
Body:{"person":[99999999],"location":2,"camera":
3,"timestamp":                1556783276,'
profile_image_unknown_data':<base64encoding>}
Response: status= HTTP_201_CREATED
```

The APIs, as mentioned earlier, send Unknown data every certain period if Unknown data is detected. In this case, the Unknown data cannot be distinguished from one another. All Unknown images will be stored in the local media/Unknown folder. The admin can sort these images into additional datasets and send some of them to the server using the API above. The Unknown image sent to the server will first be converted to base64 format. The data is wrapped in a JSON, as can be seen in the previous API. If the data sent can be recognized, then the data is sent to the server within a certain period using the API. In the API, there is no image sending because the image which will be displayed taken from the profile.

Method: POST
URL: http://ip:port/attendance/position/

```
Body:{"person":[1,5,6],"location":2,"camera":3,"t
imestamp": 1556783276}
Response: status= HTTP_201_CREATED
```

III.   RESULTS AND DISCUSSION

The implementation of a real-time attendance system uses a camera with a resolution of 1080p, a laptop, a router, and connecting cables, as shown in Fig. 5.

Testing is carried out using a two-person dataset following with ID as the name of each dataset folder, where the ID used is 1 and 6. In folder 1 is inserted a user image with ID 1 and folder 6 with user ID 6. Initially, face encoding is carried out. Hence, the pickle format data is generated. Next, the first thing to do is to set the Back-End, Front-End, and Node sections, respectively. Then, Node will run face recognition.

A. Single Person Test (Known vs Unknown)

In this section, tests are performed on faces that have been registered with the system. The face has ID = 6. The face that is detected and recognized by the system will be displayed on the website, as in Fig. 6(a), where the Known's face captured by the video can be displayed smoothly on the website. The left-side image is the video display captured by the camera and processed using a particular script, whereas the right-side image is the display on the website. The data displayed is the most recent data up to 10 seconds ago. The data retrieval period can be changed as needed. The website display is updated (refreshed) every second.
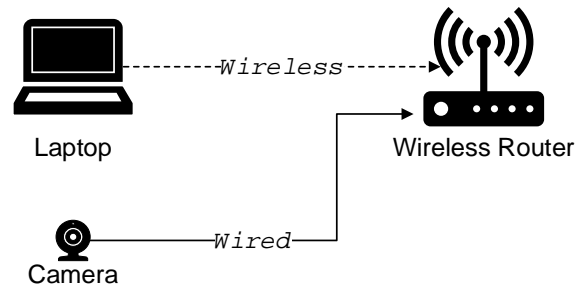
Fig. 5.   System Implementation.

(a)

(b)
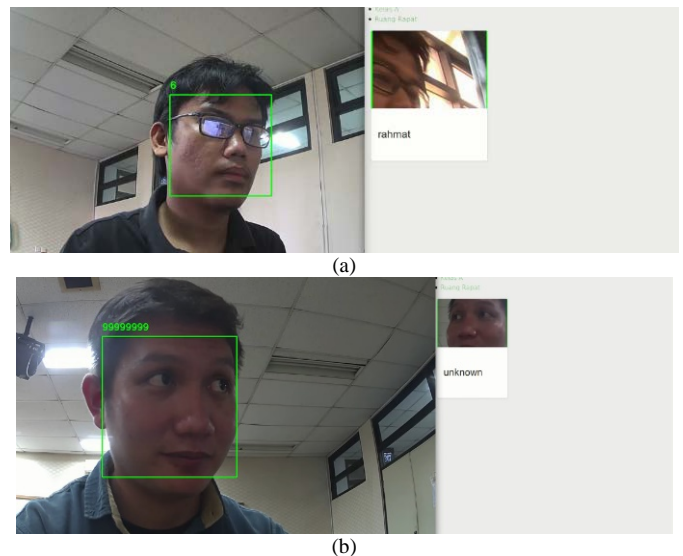
Fig. 6.   Testing for: (a) One Known Person; (b) One Unknown Person.
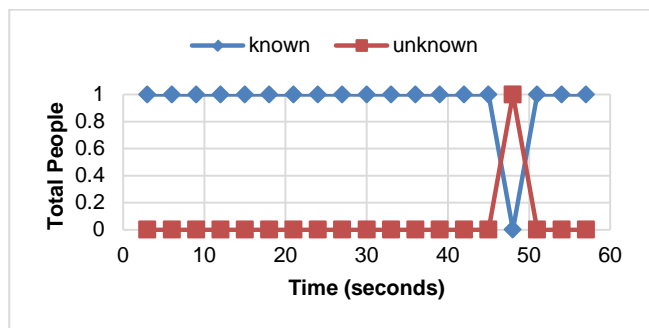
Fig. 7.   Testing for One Known Person with a Data Interval of 3 Seconds.

Fig. 7 depicts a comparison chart between registered (Known) and unregistered (Unknown) person. Data collection was done every 3 seconds with 60-second intervals. The graphics obtained are quite stable. There should be two Known straight lines on line 1 and Unknown on line 0. It showed that at 48 seconds, an error has occurred on the detected face. The Unknown test is performed on a user with an Unknown identity as visualized in Fig. 6(b). The user's face does not exist in the data set and is not even registered to the existing system. Any faces that are Unknown, will be set their ID as 99999999.

In Fig. 8, we obtain a very stable graph. There is no wrong data at all. This situation indicates that data retrieval with a more extended period can achieve more stable data than shorter periods.

Fig. 9 and Fig. 10 exhibit the Unknown graph with sample periods at intervals of 3 seconds and 5 seconds, respectively. The detection of the Unknown person tends to be unstable; this is because the Unknown faces sometimes tend to be detected as a registered user. Since the dataset has no Unknown images, the system will tend to group Unknown faces into one of the existing datasets. This tendency is caused by the tolerance factor in the library used. The tolerance value is set to 0.5; if set to be smaller, the face will be more challenging to identify. When the two images are compared, then a graph with data capture at a larger interval shows better results.

B. Two People Test

This test involves two people, namely Known and Unknown in the same camera as shown in Fig. 11. The obtained known to the unknown test chart with data interval of 3 seconds and 5 seconds are delineated in Fig. 12 and Fig. 13, respectively. A graph that should be obtained is a straight line at the number = 1. In this situation, data retrieval at intervals of 3 seconds is better than 5 seconds.

Every detected unknown face will be stored in an "Unknown" folder at a node, as shown in Fig. 14. The obtained data are set at certain intervals depending on the determined configuration. The images that have been stored in that folder are then sorted by the admin and labeled as a new dataset.

C. Three People Test

This test employs three people, composed of two known faces and one unknown face, as visualized in Fig. 15. The

result is depicted in Fig. 16. The expectation result is known as a straight line at number 2 and an unknown line at number 1. However, the results obtained are very oscillating. At that time, in 60 seconds, eight data were correctly known and five were utterly unknown. The results obtained in this study are still not stable in detecting the presence of Known and Unknown faces. This condition is probably caused by an algorithm that has not been able to handle large-scale users.
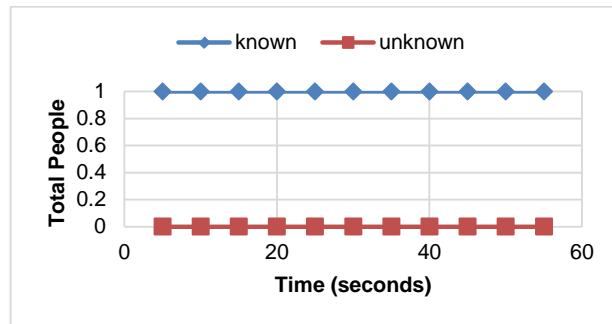
Fig. 8.   Testing for One Known Person with a Data Interval of 5 Seconds.
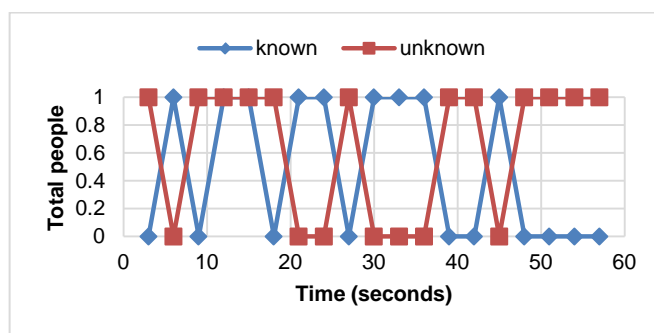
Fig. 9.   Testing for One Unknown Person with a Data Interval of 3 Seconds.
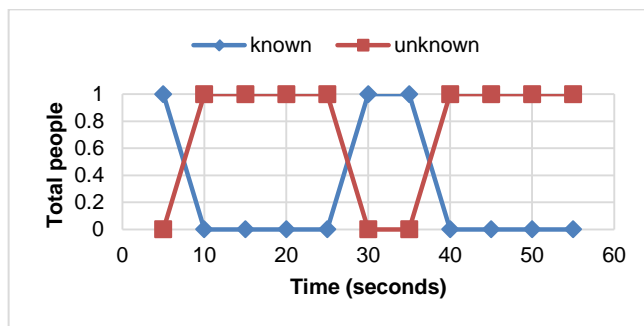
Fig. 10. Testing for One Unknown Person with a Data Interval of 5 Seconds.
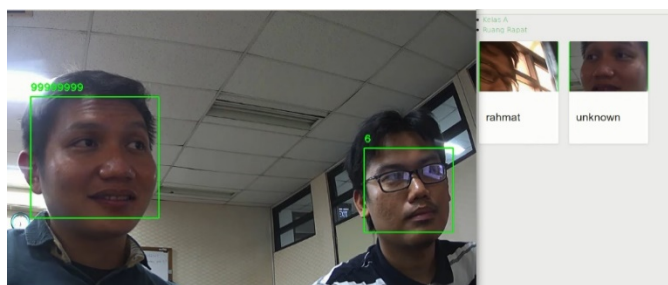
Fig. 11. Testing for Two People: Known Person vs. Unknown Person.
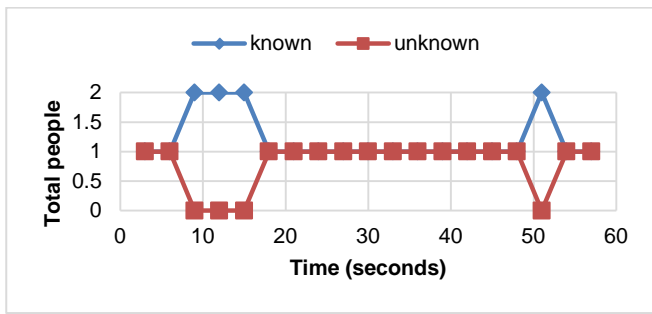
Fig. 12. Comparison Chart of Testing Two People, One Known and One Unknown, by Data Interval of 3 Seconds.
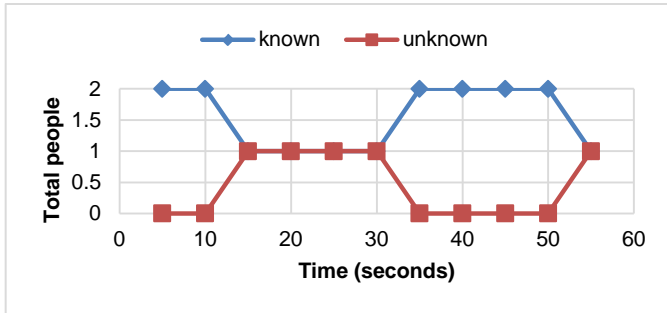
Fig. 13. Comparison Chart of Testing Two People, Known and Unknown, by Data Interval of 5 Seconds.

Fig. 14. Folder of Unknown Database in the Node.
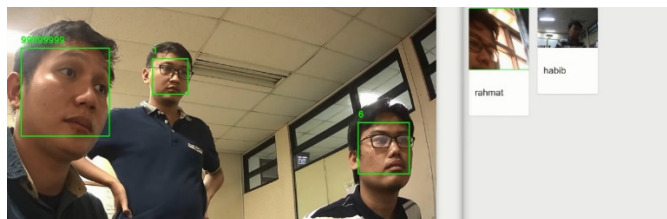
Fig. 15. Testing Three People.
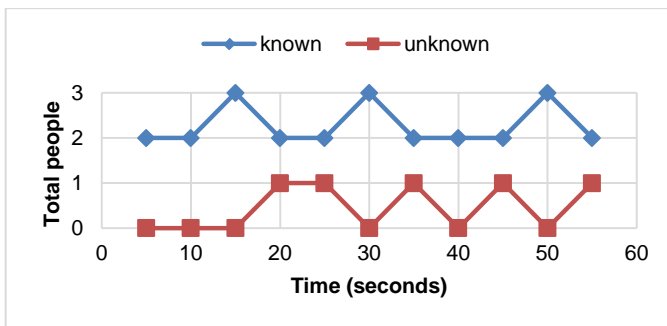
Fig. 16. Graph Testing Three People with an Interval of 3 Seconds.

IV. CONCLUSION AND FUTURE WORKS

In this work, we have developed a system that can be applied for smart meeting rooms and get results well functionally as expected. The system can detect faces, send data to the server, display the data, and save unknown faces in a folder on the node. If the detection is carried out on a registered user (known), then the system is quite stable. However, if the detection is done on an unregistered user (unknown) user is less stable. Instability result can occur when detecting the faces of three or more people in one camera. Also, there is a minimum distance to be able to detect faces accurately; which is about 2 meters.

Based on the summarized results, it is worth to improve the stability aspect in detecting more than three people for further research. Special algorithms will be applied in the system.

REFERENCES

[1] J.M. Zhang, et al., "Machine Learning Testing: Survey, Landscaped and Horizons," IEEE Transaction on Software Engineering, 2020.

[2] A. Elmahmudi and H. Ugail, "Deep face recognition using Imperfect Facial Data," Future Generation Computer Systems, Vol. 99, pp. 213-225, October 2019.

[3] S. Sharma, et al., "Face Recognition System using Machine Learning Algorithm," Proc. of the 5th Int. Conf. on Communication and Electronics Systems, 2020.

[4] S. Fuada, H.A. Shiddieqy, and T. Adiono, "A High-Accuracy of Transmission Line Faults (TLFs) Classification based on Convolutional Neural Network," Unpublished.

[5] N. Kato, et al., "Ten Challenges in Advancing Machine Learning Technologies toward 6G," IEEE Wireless Communications, Vol. 27(3), pp. 96-103, June 2020.

[6] T. Adiono, "Challenges and opportunities in designing Internet of Things," Proc. of the 1st Int. Conf. on Information Technology, Computer, and Electrical Engineering, Nov 2014.

[7] S. Fuada, et al., "Workshop Internet-of-Things untuk Guru dan Siswa Sekolah Menengah di Purwakarta, Jawa Barat, Guna Menunjang Kompetensi Era Industri 4.0," Unpublished.

[8] E. Setyowati, et al., "Mesin Absensi RFID berbasis Internet-of-Things (IoT) untuk Meningkatkan Pengetahuan Siswa di Purwakarta terhadap Teknologi," J. Pengabdian Kepada Masyarakat (DIKEMAS), Vol. 3(3), pp. 67-74, 2019.

[9] K.P. Aji, U. Darussalam, N.D. Nathasia, "Perancangan Sistem Presensi untuk Pegawai dengan RFID berbasis IoT Menggunakan NodeMCU ESP8266," J. of Information Technology and Computer Science, Vol. 5(1), pp. 25-32, 2020.

[10] Gagandeep, et al., "Biometric Fingerprint Attendance System: An Internet of Things Application," Innovations in Computer Science and Engineering, pp. 523-530, 2018.

[11] M.D. Rahmatya and M.F. Wicaksono, "Design of Student Attendance Information System with Fingerprints," IOP Conf. Series: Materials Sci. and Eng., Vol. 662(2), 2019.

[12] T. Adiono, K.S. Prakoso, C.D. Putratama, B. Yuwono, and S. Fuada, "HOG-AdaBoost Implementation for Human Detection Employing FPGA ALTERA DE2-115," Int. J. of Advanced Computer Science and Applications, Vol. 9(10), pp. 353-358, 2018.

[13] D. Monzo, et al., "A Comparative Study of Facial Landmark Localization Methods for Face Recognition Using HOG descriptors," Prof. of the 20th Int. Conf. on Pattern Recognition, pp. 1330-1333, 2010.

[14] T. Adiono, K.S. Prakoso, C.D. Putratama, B. Yuwono, and S. Fuada, "Practical Implementation of Real-time Human Detection with HOG-AdaBoost in FPGA," Prof. of IEEE Region 10 Conf. (TENCON, pp. October 2018. DOI: 10.1109/TENCON.2018.8650453.

[15] S. Fuada, et al., "A High Frame-rate Cell-based HOG Human Detector Architecture and Its FPGA Implementation," Unpublished.

[16] D. Qu, et al., "An Automatic System for Smile Recognition Based on CNN and Face Detection," Proc. of the IEEE Int. Conf. on Robotics and Biomimetics (ROBIO), pp. 243-247, 2018.

[17] B.T. Chinimili, et al., "Face recognition-based attendance system using Haar Cascade and Local Binnary Pattern Histogram Algorithm," Proc. of the 4th Int. Conf. on Trends in Electronics and Informatics, June 2020.

[18] Suwarno and Kevin, "Analysis of Face Recognition Algorithm: Dlib and OpenCV," J. of Informatics and Telecommunication Engineering (JITE), Vol. 4(1), pp. 173-184, July 2020.

[19] S. Sharma, et al., "FAREC — CNN based efficient face recognition technique using Dlib," Prof. of the Int. Conf. on Advanced Communication Control and Computing Technologies (ICACCCT), pp. 192-195, 2016.

[20] S. Khan, et al., "Real-Time Automatic Attendance System for Face Recognition using Face API and OpenCV," Wireless Personal Communication, Vol. 113, pp. 469-480, 2020.

Combined Text Mining: Fuzzy Clustering for Opinion Mining on the Traditional Culture Arts Work

Elta Sonalitha[1], Anis Zubair[2], Priyo Dari Molyo[3], Salnan Ratih Asriningtias[4]
Bambang Nurdewanto[5], Bondhan Rio Prambanan[6], Irfan Mujahidin[7]

Electrical Engineering, Merdeka University Malang[1, 7]
Information Systems Faculty of Information Technology, Merdeka University Malang[2, 5]
Information Technology, Brawijaya University[4]
Communication Engineering, Merdeka University Malang[5]
Master of management, Merdeka University Malang, Malang, Indonesia[6]

Abstract—The Indonesian government is currently intensifying work programs in the field of traditional arts and culture. In order to realize the promotion of the country's culture, the government has enacted a law on cultural promotion. One indicator of the achievement of the promotion of culture, among others, with the collection of data on traditional culture, the data mapping and data inventory can be processed into information and knowledge. In this research, indicators of performance indicators were compiled from connoisseurs of traditional works of art using data in the city of Malang, East Java, Indonesia. The results of the audience's opinion on cultural offerings can be used as a benchmark for the success of the promotion of traditional culture. When the culture is explored and tried to be displayed again, it is important to know the audience's satisfaction and understanding of the display that has just been witnessed. The results of the description of respondents in the form of opinions on the artwork will be collected as data processed using Text Mining with the Clustering of Fuzzy C-Means method to determine the audience's opinion about Feeling , which is related to feelings when viewing the beauty of the artwork, Value is related to the assessment to an art work that can be in the form of art weights contained in the work of art, and Emphasizing , which is related to empathy or respect for the art world, including professions related to the world such as dancers, musicians and others. The results achieved from this study show that has good performance on the proposed method. This can be known from the results of data testing using cluster variance V = 0.00000217901. Based on these values it can be concluded that the value of all cluster variants is good.

Keywords—Text mining; opinion mining; fuzzy clustering; arts work

I. Introduction

Cultural shifts began in the millennia era. Advances in technology led to the spread of a very diverse modern spectacle had shifted the traditions and traditional culture of the Indonesian nation. Generation who was born in the millennia era, is less familiar with the traditions and cultural heritage of the ancestors. This is very worrying and threatens the extinction of the native culture of the country [1][2]. The right step taken by the government is to pass a law on cultural promotion.

Traditional culture is the ultimate wealth of the Republic of Indonesia Unitary State, if it is not preserved properly, Indonesia's wealth will eventually disappear. For this reason, the Indonesian government promulgates Law No. 5 2017 concerning the promotion of culture. Namely the efforts that will be made to explore, record, reorganize the traditional works of the archipelago. In addition, this is reinforced by the UUD 194;32 (1); which reads culture is a future investment in the nation's civilization. There are 10 cultural fields that are defined in the presentation of the laws on cultural advancement, including: Oral Traditions, Traditional Knowledge, Manuscripts, Language, Customs, Sites, Traditional Sports, Art, Folk Games and Traditional Technology. This research raises the field of art culture.

In the world of art, there are three components that support the arts, among other artists, his appreciation for the connoisseur or awards, namely, Spectator, and Art itself as a product. This research focuses on "Audience" as a measure of the success of a show [3][4]. The opinion of the audience or viewers of a performance of art is very important for the evaluation material of the show itself. Good and bad impressions, understanding or not, enjoying or rejecting the show, whether or not satisfied will be the subject of study and input to be followed up on the next display [5].

The importance of public opinion and appreciation must receive more attention. Data collection and mapping need to be done to overcome this problem. Indicators of achieving cultural progress include the collection of data on traditional culture, data mapping, and data inventory [6][7]. The indicator of connoisseurs of art is a measure of the success of cultural promotion. When the culture is explored and tried to be displayed again, it is important to know the audience's satisfaction and understanding of the display that has just been witnessed [8][9].

Especially in the increasingly rare Traditional Art Works, requiring data and information from the public, how much attention is paid to the traditional art. The more the audience understands and understands the storyline, meaning, philosophy contained in a work of art, then the work of art can be said to be good. This understanding is a positive value from the achievement of the objectives of the cultural promotion effort. In an effort to realize the law on cultural promotion, indicators of the success of a cultural dish are mainly needed from the viewers or the public.

Based on these problems, we need a decision support system that is able to provide an overview and conclusions that can be used to benchmark policy decisions [10][11]. This research focuses on measuring the level of satisfaction of the audience of a show by using an assessment instrument that will be distributed when the audience has finished watching the show.

This study proposes the combined method of Text Mining Fuzzy Clustering as a method. Fuzzy was chosen in this study because fuzzy has the advantage of being easier to implement in various problems. Fuzzy is also often used in various kinds of problems related to forecasting, control, and clustering. F Fuzzy used in this study is Clustering of Fuzzy C-Means. Several previous studies have successfully used C-Means as a method in clustering, including Collazo-Cuevas. Text mining is used as a method of feature extraction that is filled in by the audience or the community of art lovers. While Fuzzy Clustering as a machine of machine learning to produce information [12][13]. The information generated from this decision support system is in the form of an audience's level of understanding and satisfaction with the work of art. This level of understanding and satisfaction becomes the basis of evaluation for the process of promoting culture, especially art. The achievement of the success of the appearance of a work can be measured accurately. Fuzzy Clustering can measure the level of audience satisfaction with an art performance so that it is very important to more massive promote the Indonesian culture.

The structure of this paper consists of an introduction which contains an introduction to traditional art in Indonesia as a medium to achieve the original culture and orientation of combined text mining - fuzzy clustering for this research, then the implementation process of Text Mining which is maximized by combining fuzzy c-Means to improve the results of measuring accuracy. The level of audience satisfaction with an art performance, after that result and analysis, contains an analysis of optimization analysis and improvement results from the implementation of combined text mining-fuzzy c-Means.

II. Combained Text Mining-Fuzzy C-Means

This study uses a combination of the Text Mining and Fuzzy C-Means methods as a method. Text mining is used as a method of feature extraction from the results of questionnaires or instruments that have been filled in by the audience after watching a traditional art performance. In Fig. 1, clustering data with Fuzzy C-Means or information that has gone through a weighting process [14][15].

The stages in this research starting from extracting information to producing information are shown in the figure below.
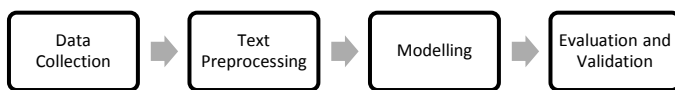
Fig. 1. Flow Diagram of the Text Mining - Fuzzy C-Means Combined Method.

A. The Data Collection

Data used in the study were 139 opinions of data from viewers of art performances in the city of Malang, East Java, Indonesia. Data in the form of text or paragraph data obtained from the questionnaire form. The data collection process is carried out by giving questionnaires or research instruments to the audience who have just seen a web-based work of art. After the data is collected, the process will then be performed Text Preprocessing [16]. The data that has been collected is shown in Table I.

B. Text Preprocessing

Text mining is a knowledge intensive process where users interact and work with a group of documents using several analysis tools. Text mining aims to get useful information from data sources in the form of documents consisting of unstructured text through identification, and exploration of a pattern. The stage in text mining is text preprocessing, namely changing unstructured data into structured data. Steps in the preprocessing text include Bag of Words and Term Weighting [17].

Bag of words consists of several processes, including folding cases to eliminate characters other than letters and change all letters to lowercase letters. Next Tokenizing which aims to break the sentence into separate words known as term or token. Furthermore filtering by removing punctuation, changing capital letters into lowercase letters and eliminating stop word that aims to delete words that are not useful or have no influence in the process [18]. Finally, stemming is to get the basic words from words that have received affixes or other information.

TABLE I. Collection of Data on Audience Opinions Performing Art

Audience ID	Opinion
1	have seen traditional art performances often watch traditional art performances ... Very important to improve or develop
2	have seen traditional art performances often watch traditional art performances in a year Awareness of the importance of traditional arts and local culture to continue to be supported and cultivated Also maintain and support the existence of traditional art ... It is important to improve or develop
3	ever see traditional art performances sometimes watch traditional art performances in a year Because love traditional art entertainment As a form of self-appreciation More have the awareness to continue to participate in preserving and preserving traditional art ... Marketing Aspects Very Important to be improved or developed
4	ever see traditional art performances rarely watch traditional art performances in a year Idly, invited or fill free time Participate in preserving and supporting the existence of traditional art More having the awareness to continue to participate in preserving and preserving traditional arts ... Important to be repaired or developed Social security of the performers Very important to be repaired or developed
...	...
139	ever see traditional art performances sometimes watch traditional art performances in a year Awareness of the importance of traditional arts and local culture to continue to be supported and cultivated as a form of appreciation Balance, ... Social security of actors is very important to be repaired or developed

Term Weighting aims to Identify the value or heft of a term based on the level of interest in the document. Period Frequency and Inverse Document Frequency (TF-IDF) is a weighting that is often used in information retrieval and text mining [19][20]. TF is the frequency of the case of a term in the document. IDF is the frequency of the case of a term in the whole document. The more frequently the term comes up in a document, the more-large the TF value and the smaller the IDF value.

$$w_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log\left(\frac{N}{df_j}\right)$$

Where w_{ij} is the the weight of the j term for document i, tf_{ij} is the number of the case of term j in document i, N is the number of literature, df_j is the number of literature containing term j.

C. Evaluation

The testing process is carried out using cluster variance. Equation for calculate the cluster variance shown in (2).

$$v_c^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} \left(d_i - \bar{d}_i\right)^2$$

Where v_c^2 is the variance in cluster c, c is the 1 ... k, k is the number of clusters, n_c is the the amount of data in cluster c, d_i is the data i on a cluster, $\bar{d}_i$ is the average of data in a cluster. This variant is used to see the results of the variance of data distribution in a cluster. The smaller the value of Vw, the better the cluster. The equation for calculating Vw shown in (3).

$$v_w = \frac{1}{N - k} \sum_{i=1}^{k} (n_i - 1)v_i^2$$

Where N is the value point of all data, k is the value point of clusters, n_i is the value point of data members in cluster i. This variant is used to see the results of variance in data distribution between clusters. The greater the value of V_b, the better the cluster[21][22]. The equation for calculating V_b is shown in (4).

$$v_b = \frac{1}{k - 1} \sum_{i=1}^{k} n_i \left(d_{ij} - \bar{d}\right)^2$$

Where k is the value point of clusters d_{ij} is the data to j in a cluster to i, $\bar{d}$ is the average of $\bar{d}_i$. To see the variance of all clusters, it can be measured by (5). If the V value is getting smaller, the cluster value is getting better.

III. RESULT AND ANALYSIS

In this section, the results of the tests that were carried out in the previous chapter are explained, starting from Text Preprocessing, Modeling, and Evaluation. After the data is collected, the process carried out is Text Preprocessing, the results of which are shown in Table II. In this process, punctuation has been successfully removed.

Furthermore, after the punctuation in the previous process has been lost, the data is processed using Tokenizing, which breaks the sentence into separate words. The results of this test are shown in Table III.

TABLE II. THE RESULT OF CASE FOLDING PROCESS IN TEXT PREPROCESSING

Audience ID	Opinion
1	ever see traditional art performances often watch traditional art performances in a year of awareness of the importance of traditional arts and local culture to continue to be supported and cultivated to take care ... it is very important to improve or develop
2	... adequate audio system sound stage and accompaniment that you have ever enjoyed has adequate quality accompaniment equipment from the show that you have ever seen conditions are very adequate equipment from the show includes costume property ... developed social security for artists is very important to be repaired or developed
3	... the educational values of messages or social criticism that you can catch enough to understand traditional art very well understand the meaning and meaning of the verbal language used to understand the meaning of conversations in the accompaniment narrative dialogue used during the performance can capture the message conveyed through the stage layout includes lighting and the use of property in staging messages of value education and social criticism delivered can be understood and understood messages delivered beautifully and deserve to strongly approve or agree on the message and values ...
4	... about traditional art, lack of understanding of the meaning and meaning of the verbal language that is lacking in understanding the meaning of conversational songs, narrative accompaniment used during ... developed price standards are very important to be improved or developed social security of the actors is very important to be repaired or developed
...	…
139	... through word of mouth rarely get information about the show schedule through the leaflet flyer leaflet brochure etc. sometimes get information about the show schedule through announcements in a line conversation groups etc ...

TABLE III. The Result of Tokenizing and Filtering Process in Text Preprocessing

Audience ID	Opinion
1	ever traditional art performances often watch traditional art performances a year awareness of traditional arts supported local culture cultivated keeping supporting the existence of traditional art having awareness of preserving wayang people condition ...
2	... adequate stage lighting quality ever visited adequate audio stage sound system accompaniment enjoy adequate quality accompaniment equipment performances have seen conditions are very adequate performance equipment including costume property ...
3	... supporting equipment has been watched very adequate location of the show quite easy to reach publication information about the event is available often information schedule social media shows often information schedule shows word of mouth sometimes schedule information ...
4	... marketing is very important to be improved developed production management is very important to be improved developed price standards are very important to be improved developed social security is very important actors to improve developed ...
...	…
139	... the ability of presenters is quite important to be improved developed facilities are very important to be improved developed equipment is very important to be improved developed quality of cultivation is quite important to be improved developed creative process of exploration is quite important to be improved developed use ...

The last Text Preprocessing is Steptemming. This process aims to get the basic words from words that have received affixes or other information. The results of this process are shown in Table IV.

After obtaining the results in Table IV, the weighting process is then performed using the TF-IDF Term Weighting technique. Weighting is done based on the level of importance in the document. The number of weights for each audience is 34. The figure is obtained based on the results of the formation of opinion variables. In Table V, these opinion variables are formulated into Keywords. So, each audience has 34 keywords.

Based on Table VI, the technique for determining data clusters that are Fuzzy clustering in which the existence of each data point in a cluster is determined by the degree of membership. The determination of data clusters is based on the Euclidean Distance form to measure the proximity between data. The concept of fuzzy clustering is to indicate the center of the cluster in advance which is each cluster center on average

location. In the cluster center of the initial condition, the value level is still not precision, repetitive improvements are made to the membership degree and the center of the cluster in each point of data so that the center of cluster moves to the right position. At this stage, cluster testing has been carried out 5 times. Tests carried out using an initial cluster of 2 clusters with an error e <0.001. The results of this test are shown in Table VII.

To determine the accuracy of the cluster formed, the cluster variance testing technique Vw, Vb, and V. Based on the results shown in Table VII, it can be seen that the value of Vw = 0.284667123 the smaller, the better the distribution of data in the cluster. Meanwhile, to find out the accuracy of the variants of all clusters by using the calculation of the value of V. It can be seen in Table VII that the value of V is getting smaller so it can be concluded that the better cluster value V = 0.00000217.

TABLE IV.   The Result of Stemming Process in Text Preprocessing

Audience ID	Opinion
1	ever traditional art performances often watch traditional art performances in the year of awareness of traditional arts local culture support cultivated keep supporting the existence of traditional arts awareness of preserving sustainable wayang people the condition of the building ever visited ...
2	... stage lighting has never been enough adequate stage sound system ever enjoyed the adequate quality of accompaniment equipment shows have seen very adequate conditions of equipment performances costumes property equipment support ever ...
3	the show is quite easy to reach the publication of event information is quite available often information schedule social media shows often information schedule shows word of mouth sometimes information schedule shows announcements by message group conversations etc. quality
…	…
139	... developed information dissemination reviewing news narratives is quite important developed marketing aspects are quite important developed production management is quite important developed price standards are quite important developed social guarantees art behavior is quite important to be developed

TABLE V.   Keywords that have been Formed based on Audience Opinion

No	Keyword	No	Keyword
1	never	18	very bad
2	ever	19	bad
3	often	20	very understanding
4	rarely	21	quite understand
5	sometimes	22	not understand
6	awareness	23	do not understand
7	cheer up	24	understand
8	prankster	25	really understand
9	appreciation	26	quite understand
10	keep	27	do not completely understand
11	support	28	do not understand
12	lesson	29	understand
13	introspection	30	very catch
14	sustainable	31	just catch it
15	very good	32	less catch
16	quite good	33	not catch
17	very nice	34	catch it

TABLE VI. TF-IDF Calculation Results

No	Keyword	DF	IDF
1	appreciation	26	1.676
2	very nice	103	0.3
3	bad	0	0
4	quite good	55	0.927
5	quite understand	53	0.964
6	quite understand	76	0.604
7	just catch it	51	1.003
8	support	137	0.014
9	cheer up	65	0.76
10	introspection	12	2.45
…	…	…	…
34	not catch	2	4.241

TABLE VII. Results of the Audience Opinion Clustering Test

No.	Experiment …	Vc	Vw	Vb	V
1	Experiment 1	Cluster_1 = 11.833549726654	0.284671533	130582.4281	2.18001E-06
		Cluster_2 = 10.111108133333			
2	Experiment 2	Cluster_1 = 10.111108133333	0.715328467	130582.4281	5.47798E-06
		Cluster_2 = 11.833549726654			
3	Experiment 3	Cluster_1 = 11.833549726654	0.284671533	130582.4281	2.18001E-06
		Cluster_2 = 10.111108133333			
4	Experiment 4	Cluster_1 = 11.833549726654	0.284671533	130582.4281	2.18001E-06
		Cluster_2 = 10.111108133333			
5	Experiment 5	Cluster_1 = 10.111108133333	0.284667123	130582.4281	2.17901E-06
		Cluster_2 = 11.833549726654			
Average			0.456934307	130582.4281	3.4992E-06

IV. CONCLUSION

Combined Text Mining Fuzzy C-Means Clustering method that is proposed in this research. This method can be implemented in the clustering opinion of an audience of art performances to assess the level of audience satisfaction with an art show. This step is carried out as an effort to promote Indonesian culture. The novelty of this research that can measure the level of audience satisfaction with an art performance so that it is very important to more massive promote the Indonesian culture. The results of testing the value of V = 0.00000217 show that the combined method of Text Mining and Fuzzy C-Means Clustering has good performance. This is indicated by the decreasing value of V in each test. This can indicate that all cluster variants are getting better.

REFERENCES

[1] S. C. Satapathy, V. Bhateja, and A. Joshi, "Proceedings of the International Conference on Data Engineering and Communication Technology : ICDECT 2016. Volume 2," Int. J. Eng. Res. Technol., 2020.

[2] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," IEEE Access, 2019.

[3] A. Abdul Aziz and A. Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches," IEEE Access, 2020.

[4] I. Mujahidin, D. A. Prasetya, Nachrowie, S. A. Sena, and P. S. Arinda, "Performance tuning of spade card antenna using mean average loss of backpropagation neural network," Int. J. Adv. Comput. Sci. Appl., 2020.

[5] T. Traylor, J. Straub, Gurmeet, and N. Snell, "Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator," in Proceedings - 13th IEEE International Conference on Semantic Computing, ICSC 2019, 2019.

[6] X. Peng, Y. Bao, and Z. Huang, "Perceiving Beijing's 'city Image' across different groups based on geotagged social media data," IEEE Access, 2020.

[7] I. Mujahidin, D. A. Prasetya, A. B. Setywan, and P. S. Arinda, "Circular Polarization 5.5 GHz Double Square Margin Antenna in the Metal Framed Smartphone for SIL Wireless Sensor," in 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2019, pp. 1–6.

[8] X. Chen, Y. Zhao, Z. Cui, G. Meng, Y. Liu, and Z. Wang, "Large-Scale Empirical Studies on Effort-Aware Security Vulnerability Prediction Methods," IEEE Trans. Reliab., 2020.

[9] R. Yuwono and I. Mujahidin, "Rectifier using UWB microstrip antenna as electromagnetic energy harvester for GSM, CCTV and Wi-Fi transmitter," J. Commun., 2019.

[10] Z. Li, C. Zhang, S. Jia, and J. Zhang, "Galex: Exploring the evolution and intersection of disciplines," IEEE Trans. Vis. Comput. Graph., 2020.

[11] I. Mujahidin, S. H. Pramono, and A. Muslim, "5.5 Ghz Directional Antenna with 90 Degree Phase Difference Output," 2018.

[12] D. Buenano-Fernandez, M. Gonzalez, D. Gil, and S. Lujan-Mora, "Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach," IEEE Access, 2020.

[13] A. Karami, M. Lundy, F. Webb, and Y. K. Dwivedi, "Twitter and Research: A Systematic Literature Review through Text Mining," IEEE Access, 2020.

[14] H. Gao, Y. Xu, Y. Yin, W. Zhang, R. Li, and X. Wang, "Context-Aware QoS Prediction with Neural Collaborative Filtering for Internet-of-Things Services," IEEE Internet Things J., 2020.

[15] R. Yuwono, I. Mujahidin, A. Mustofa, and Aisah, "Rectifier using UFO microstrip antenna as electromagnetic energy harvester," Adv. Sci. Lett., 2015.

[16] I. Hafeez, M. Antikainen, A. Y. Ding, and S. Tarkoma, "IoT-KEEPER: Detecting Malicious IoT Network Activity Using Online Traffic Analysis at the Edge," IEEE Trans. Netw. Serv. Manag., 2020.

[17] A. Gavioli, E. G. de Souza, C. L. Bazzi, K. Schenatto, and N. M. Betzek, "Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods," Biosyst. Eng., 2019.

[18] A. Seyedzadeh, S. Maroufpoor, E. Maroufpoor, J. Shiri, O. Bozorg-Haddad, and F. Gavazi, "Artificial intelligence approach to estimate discharge of drip tape irrigation based on temperature and pressure," Agric. Water Manag., 2020.

[19] E. H. Kim, S. K. Oh, W. Pedrycz, and Z. Fu, "Reinforced fuzzy clustering-based ensemble neural networks," IEEE Trans. Fuzzy Syst., 2020.

[20] H. Gan, "Safe Semi-Supervised Fuzzy C -Means Clustering," IEEE Access, 2019.

[21] M. B. Ferraro, P. Giordani, and A. Serafini, "Fclust: An R package for fuzzy clustering," R J., 2019.

[22] A. Alsarhan, Y. Kilani, A. Al-Dubai, A. Y. Zomaya, and A. Hussain, "Novel Fuzzy and Game Theory Based Clustering and Decision Making for VANETs," IEEE Trans. Veh. Technol., 2020.

Increasing User Satisfaction of Mobile Commerce using Usability

Ninyikiriza Deborah Lynn[1]
Magister Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia 55281
Department of Computer Science and
Information Technology
Kabale University, Kabale, Uganda

Arefin Islam Sourav[2]
Magister Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia 55281
Department of Computer Science and
Engineering, Daffodil International
University, Dhaka, Bangladesh

Djoko Budiyanto Setyohadi[3]
Magister Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia 55281

Abstract—**Online shopping continues to simplify people's way of life in the present world. People no longer need to physically visit stores to buy items for home use or other purposes. This can be done online using mobile applications to order for preferred items, which can be delivered in return. However, the increase in the features of mobile applications and mobile devices makes usability testing a necessary aspect in today's advancement of technology. This paper uses an experiment-based usability testing evaluation on Lazada Indonesia mobile application to understand four usability factors, namely, ease of use, efficiency, functionality, and satisfaction. The test was performed using 40 participants and all were students from Universitas Atma Jaya Yogyakarta, Universitas Gadjah Mada, and Universitas Sanata Dharma. These performed 6 tasks on the mobile application and later answered a questionnaire to capture their views concerning the usability of the mobile application. The results were analyzed using SPSS software and descriptive statistics were displayed using standard deviation and arithmetic means. The results from the evaluation showed that the mobile application is easy to use, efficient, and has good functionality. However, some issues were mentioned by participants which indicated that users were not fully satisfied with the mobile application. Therefore, this calls for designers to consider these usability issues to increase user satisfaction.**

Keywords—*Usability; usability testing; mobile commerce; mobile application*

I. INTRODUCTION

Currently, mobile commerce has changed the nature of business worldwide. Mobile commerce can be defined as the act of conducting transactions of goods and services of any monetary value [1], through mobile wireless devices such as smartphones and tablets. With mobile commerce, products can be searched and purchased online by consumers [2], anywhere through internet-enabled wireless devices, without any worry or need to use other devices like laptops. This activity has stimulated the growth rate of the production of mobile applications in app stores. These applications play a big role in the lives of users and have led to the rapid development of the world thus making the world a better place every day. They can be used for several purposes such as online shopping [3],[4], marketing and product distribution [5],[6], delivering and easing Health services [7],[8],[9], monitoring and compiling records, playing games, exercising and home

therapy [10], connecting with friends [11], photography, medical support and interactions [12], disaster management [13], accessing educational information by students [14], and so many others. However, because several programmers are coming up with new applications every day, this calls for extra attention to the users, as well as to the developers to ensure that these applications truly perform the roles they are meant for or else it will be a waste of time. Therefore, developers must develop applications of quality and go ahead to prove that these applications play their roles to catch up with the rising market, and this can only be done through usability testing. One of the important factors that give a mobile application quality is its usability. If the usability of the app is complex, then this brings down its quality. Developers should always consider several factors before developing a mobile application and ensure that the application suits the purpose it is meant for. Design constraints such as personalization, data entry models, freedom to navigate should all be considered to give the users satisfaction. In addition to that, applications that deal well with information overload should be considered as it is an important factor that greatly troubles online shoppers nowadays [15]. Usability according to users simply deals with ease of use of the application, the efficiency of the application, effectiveness, and satisfaction of users [16].

Some applications make users frustrated while using them, they are not enjoyable, others do not give users a chance to navigate freely because they have a lot of restrictions, others are very difficult to use, and some also require a lot of strength or thinking to be used by the user. This leaves the users disappointed and unsatisfied with some of the mobile applications, yet they could be important to them. This makes usability testing a vulnerable aspect of mobile application development. Therefore, developers need to carry out usability tests on the users, to understand the effectiveness, ease of use, the efficiency of the applications, and find out if the users are satisfied with the mobile application and their features. This helps developers to improve their designs and thus achieving higher market and competency. In e-commerce business activities, consumers must get and understand information on mobile apps first, before they carry out online purchases. This, therefore, requires application designs that are satisfactory to the users to help them find convenient information faster. The kind of process needs the ease of use of mobile apps in e-

commerce activities [17]. Users will always stop using any mobile application if it doesn't show its major business intentions and doesn't pay attention to consumer needs. This is because of the absence of all this makes online shopping difficult and uncomfortable for consumers. In the end, consumers will always be attracted by other mobile applications whose human-computer interaction is friendly and attractive. This, therefore [17], makes evaluating usability especially the "ease of use" point very necessary.

Online shopping today continues to ease ways of doing business to the business sector worldwide. With improved technology today, e-commerce has truly shown a big impact and rise through the improvement of online trade since the 1990s. Moreover the trade has now spread to all consumers and sellers both locally and internationally around the globe. People can sell their commodities while in rural areas to people in urban areas, and vice versa just easily through the internet.

Indonesia currently greatly supports online platforms to buy and sell commodities. There are several platforms used in Indonesia such as Lazada Indonesia, Shopee, Tokopedia, Amazon and so many others. The use of mobile smartphones for mobile shopping in Indonesia is increasing due to the rapid growth of the use and access to the internet. In 2017 [18], the number of internet users in Indonesia was reported to rise to 143 million of the total population. This means that the highest number of Indonesians (more than 50%) possess smartphones [19]. Lazada is an e-commerce website that was launched in March 2012 in Indonesia, Malaysia, the Philippines, Thailand, and Vietnam, all in south-east Asia.

Lazada Indonesia is a very popular e-commerce site in Indonesia that began in 2012. It is one of the fastest-growing online shopping platforms in Indonesia. People use the application to buy and sell several types of products, ranging from home decor, beauty products, elegant products, to health products. This is simply done by accessing the site and application of Lazada. The application is supported by various payment modes including "cash-on-delivery" the most preferred payment mode which makes it simpler for all Indonesian consumers to get all items they order. In the year 2014, Lazada was ranked as the highest visited site in Indonesia. This study concentrated on the Lazada Indonesia application for mobile phones. The objective of this study is to evaluate the usability of the Lazada Indonesia mobile application.

In mobile commerce, users usually experience problems with using online shopping mobile applications. While these problems may result from a lack of experience in using the mobile application, other issues may result due to the poor performance of the mobile application. This study, therefore, aims to understand the performance and issues faced by Lazada Indonesia mobile application users. This will help to fulfill the gaps of usability issues facing users by providing designers with the required information concerning usability challenges from users. This will be done by carrying out a usability test on the mobile application to understand the performance of its features and options used by its users to attain services offered by the mobile application. Four usability aspects i.e., efficiency, functionality, effectiveness, and satisfaction are tested. Understanding the existing issues experienced by users will help Lazada mobile application designers to maintain their strengths and improve the weaknesses of the mobile application following the results from the usability test performed in this study. Filling this gap of usability issues faced by Lazada Indonesia mobile application users will promote its good usability and thus increase user satisfaction of the mobile application.

II. Literature Review

A. Smartphones and Mobile Shopping

Presently, the use of mobile smartphones and internet penetration has increased compared to ancient times. In many countries, more than 50% of people who own phones possess smartphones and this has made mobile phones the quickest channel for mobile commerce around the world. Consumers use mobile commerce for specific transactions such as ordering for mobile fast foods [20], mobile transport [21], mobile banking [22], mobile payment [23], mobile learning [24], and mobile health [25]. Some researchers claim that mobile commerce and its applications [26] are a rising area as researchers continue to show their interest in this field for research. With the evolution of smartphones, the traditional web-based mobile commerce continues to be replaced by mobile shopping applications [27]. Since carrying out mobile commerce transactions requires the practitioner to own a smartphone [28], mobile online stores theoretically support multiple shopping activities to take place such as searching for information and buying products [29]. However, Sometimes mobile users encounter challenges compared to users shopping using personal computers [30]. The two searches are different in behavior. In some instances, the users may encounter increased costs due to smaller screens of mobile devices and encounter unclear interfaces. However, mobile phones continue to be preferred since they are portable, that is; sellers and buyers can perform business transactions and carry out advertisements at anytime and anywhere [30], and users can have a variety of mobile applications [19]. In addition to that, other users prefer smartphones to view images of commodities displayed in online shops, perform shopping via social media platforms such as Facebook and Instagram [31], as well as listen to music and watch videos [32]. Thus, the popularity of mobile commerce continues to bring a positive impact on society as well as to the business world.

Usually, mobile applications are so significant to users as they are designed by companies to widen their businesses using mobile platforms [33], therefore, most mobile applications are usually mobile versions of online websites. The existence of mobile smartphones provides an alternative channel of doing shopping activities, following offline and online channels. Engaging mobile applications to do mobile shopping is time convenient and since users are usually directly in control, it reduces stress and saves time that would be used moving to shops physically and reduces costs. In addition to that, some retailers can also provide their customers with access to virtual stores [34], which can be useful to consumers to locate and purchase commodities.

However, the increase in the behavior of mobile shopping greatly depends on the experience of the user. As the users'

experience to use mobile applications increases, the more the online shoppers install more applications. The good usability of any mobile application makes it enjoyable and completely brings reduced irritation while using the mobile application [35]. This is the reason as to why when smartphone owners get more experience of using smartphones [33], they usually tend to possess a higher number of mobile shopping apps. One of the significant quality attributes of a mobile application is the ease of use and freedom of navigation which helps and encourages electronic commerce retailers to carry out their operations effectively [12].

B. Usability and Usability Testing

Usability can be defined as the ability of the user to be able to interact with the system and perform certain tasks. Usability appears among the key concepts of human-computer interaction which are all covered under the concept of usability engineering. Usability engineering deals with aspects of user interactive experience [17], such as game usability, web usability, and mobile usability. Currently, several researchers have carried out researches about the ease of use of mobile applications. However, the activity of online shopping using mobile applications can only be successful if customers achieve adequate satisfaction. Therefore, usability evaluation in e-commerce becomes a very necessary process to ensure acceptability, ease of use, effective and efficient mobile applications and acts as a significant reference for developers seeking advice for design and feature improvements [36]. In application development, usability evaluation helps in raising the profits, improving designs, and reaching the exact users' expectations. Developers need to aim at producing the best and suitable designs on the market [37], and this can help them to achieve the goal of usability evaluation.

Testing the usability of applications is a crucial activity in the application development life cycle. It confirms the quality as well as the usefulness of the effort applied in developing the mobile application [37]. Several usability evaluations usually require to be carried out on all products, be it commodities, before they are released to exist on the market. This helps to curb down the disappointment and loss of customers in case of any usability issues application, which would exist in the real application designed. The increased use of the internet and the activity of usability testing gives more opportunities to designers to identify issues via usability testing [38], and thus makes the use of mobile applications and mobile devices/gadgets more inherent in our daily lives. Usability Evaluation is based on various activities such as planning the tasks, identifying the evaluation methods, choosing participants, and carrying out the task, analyzing, and recommending the results. Generally, usability testing of mobile applications continues to show its significance at a high rate today. According to some studies, [39][40], some designers have designed and tested the usability of different mobile applications.

III. Methodology

In this study, the researchers followed a five-step procedure to carry out usability testing, namely; planning, choosing participants, performing the usability test, analyzing results from the test, and finally documenting the results. This study

invited 40 (100%) participants who participated willingly and helped researchers to evaluate the usability of the Lazada Indonesia mobile application. Of the 40 participants, 30 were females (75%), and 10 (25%) were males. The participants were all students from universities of Atma Jaya, Gadjah Mada, and Sanata Dharma. All the 40 participants were between the ages of 20 to 35. However, 12 (30%) of the participants were in the age category of 20-25 while all the rest 28 (70%) participants were between the ages of 25-35.

Among the participants, 30 (75%) were master's degree students, 6 (15%) were bachelor's degree students and 4 (10%) were Ph.D. students. All the participants in the study had a clear knowledge of computer operation and android smartphones. However, not all of them were familiar with the operation of the Lazada mobile application. The participants were graded into three main categories of users, i.e. novice/ beginner users, intermediate users, and expert/ experienced users of the application. Out of the total number of participants in the evaluation test, 4 (10%) were novices/ beginner users, 8 (20%) were intermediate users and 28 (70%) were experts/experienced users.

The novices/ beginner users owned smartphones in their daily lives, but they had never installed Lazada Indonesia mobile application. Therefore, these users had never interacted with the mobile application before. The intermediate users had had the Lazada application in their smartphones for not more than 3 months before the test was carried out but had not interacted enough with the mobile application. Each of the intermediate users had at least ever ordered one item from Lazada online shopping, but not more than two items. Expert/ experienced users confessed they had good knowledge about the mobile application and had fully interacted with it for at least 2 or more years. These users knew about online shopping and each of the 28 expert users had ever bought more than 5 items using the Lazada Indonesia mobile application.

The test was carried out at Mrican Residence, in one big free room which was close to the hostels in which all the participants lived. The residence had a very clear and quiet environment and that is why it was chosen to make the participants comfortable first. All participants were asked to carry out the task while seated so that the researchers would learn clearly what they were doing. A questionnaire was distributed to each participant after the test to elicit the participants' perceptions on the usability of Lazada's mobile application's interface. Finally, after conducting the study, SPSS software was used by the researchers for data analysis. The data was illustrated in the form of a mean and standard deviation per item.

A. Tools used for the Experiment

To carry out this experiment/test, there are several necessary and useful tools that the researchers used in the usability evaluation process. Since the participants were 40 in number, the test was carried out using android smartphones that were owned by each of the participants. The phone brands included Samsung, iPhone, Oppo, and Vivo. Also, A digital video camera (16MP High definition Digital video Camcorder 1080p) was used to record the participants' activities and take their pictures. Another iPhone camera on the researcher's

mobile phone was used to move around the participants to watch their attitude and reactions while they carried out the given task on the Lazada mobile application. To limit the time for every single task, a stopwatch digital professional LCD was employed. All the 40 participants were provided with a set of test task/scenario that was to be performed (40 sets of test tasks). A set of questionnaires (after-task questionnaire) for each participant was also prepared to be answered at the end of the practical session. The Questionnaire contained demographic part and usability questions. Each usability aspect had four usability questions required to be answered. These questions would help the researchers to evaluate the four intended usability aspects in the study. The statements about the 'ease of use' aspect required the users to answer whether it was easy for them to search items, interact with the interface, whether the mobile application was confusing to them, and whether it was easy for them to locate the search box. On the hand of the 'efficiency aspect', the researchers needed to know whether it did not require much thought, and time for the users to attempt the task, whether they did not get confused about the task, and also to know if they needed extra time to accomplish the task. To consider the 'functionality' of the mobile application, the researchers sought to know from the users whether it was easy for them to navigate through the mobile application as they carried out the task, whether the features of the application were clear and understandable, whether the steps of making a purchase were easy and fast and lastly, whether the application features were responding fast and clearly. Moreso, to prove 'user satisfaction' of the mobile application, the researchers through the questionnaire asked the participants whether carrying out the transaction was satisfying and clear with feedback, whether the users were satisfied with the features and looks of the mobile application, whether they felt comfortable while they used the mobile application, and finally whether they had fun and freedom with the mobile application as they carried out the task. All questions on each aspect were to be answered based on the 5 Likert Scale. i.e., Strong disagree, disagree, Less Agree, Agree and Strong Agree.

The whole test session started with a briefing and a thorough checking by researchers to ensure all participants had a clear and functioning version of Lazada mobile application. For the novice/beginner users who did not have the application, the Lazada application was installed and tested by the researchers on their android phones before the test session started. To ensure a smooth activity, the researchers decided to first carry out a quick pilot study before the participants began attempting the task. This was to ensure that the time given to the participants would be enough for them to carry out the scenarios provided for usability testing, as well as answering the post-task questionnaire. After carrying out the pilot study, the participants were then briefed about the time limit and set to begin performing the task. Two moderators conducted the study to ensure the smooth running of the study and do a proper recording with a camera to capture participants' attitudes.

B. Test Description/ Task Scenarios

In this experiment, the researchers prepared the task scenarios that were to be carried out by the users/ participants.

The participants were asked to carry out the following six task scenarios on the Lazada mobile application: (1) Search for an iPhone (iPhone 11 pro max) using the search box and add it to the trolley. (2) Return to the home page. Go to the trolley and find the product that was added and buy it using the "pay on delivery" payment option. (3) Search any product of choice and order it using the bank transfer (BCA) payment method, then after completing the process, cancel the order. (4) Search for any 2 household items such as (lunch boxes) of different colors and add each of them to the trolley. (5) Go to the trolley and order for the two lunch boxes at once. (6) Go to the trolley and delete all the items that were added in that task session. The mentioned were the six task scenarios that were to be attempted by the participants.

The test was conducted under specific procedures. Since the participants were many, to effectively carry out the usability evaluation test, the researchers divided them into four groups (A, B, C, D). Each group contained 10 participants. These groups were to participate in the experiment, one group after the other. Choosing which participants to join groups was done randomly by the test moderators. There were no special criteria followed in forming the four groups. After informing each participant to which group they belonged, the place was arranged and participants were tested beginning with group A. Ensuring that the participants were comfortable and ready for the test, each group at a time was briefed and let to begin the test while they were being recorded.

Following the timer after carrying out the task scenarios, participants were asked to answer the questionnaire to help the researchers evaluate the usability of the mobile application. The recording was then stopped, questionnaires were collected from participants and moderators thanked the participants, sent them off, and called for the next group to repeat the exercise until Group D was done, and the exercise finished. The results were later analyzed using SPSS and descriptive statistics such as mean and frequency distribution were used to display and present results of the study.

IV. RESULT AND DISCUSSION

The usability evaluation test of this study was purposed to understand the performance and know the challenges faced by Lazada Indonesia mobile application users, as well as to finally know whether the users were satisfied with the mobile application. The test was carried out on four desired usability factors about Lazada mobile application i.e. (1) ease of use of the mobile application; (2) efficiency of the mobile application; (3) functionality of the mobile application and finally, (4) satisfaction attained by users of Lazada mobile application. This section defines the results of the study that were attained from the experiment/ test. The rating scale of the tested factors is also included. Apart from the observation that was done by the moderators, the data analysis greatly considered the answers of the participants since they were the people that were directly involved in the interaction with the mobile application. Therefore, it was considered important to consider their feelings about the usability nature of the mobile application. This is therefore the reason why the researchers mainly relied on the questionnaire answers and comments to make the result of this study.

Tables I to IV were computed from compiling the responses from the questions that were asked about the four aspects of the mobile application, compiled in the post-task questionnaire. Table V displays the rating scale. The questionnaire contained four questions for each of the researched usability factors of the mobile application which were to be answered by each of the 40 participants. The participants' answers were given based on the 5 Likert Scale. These were represented by 1,2,3,4 and 5 in SPSS to obtain the mean and standard deviation. Next is a summary and explanation of the results, ranging from ease of use, efficiency, functionality, and satisfaction of users with Lazada Indonesia mobile application.

TABLE I. EASE OF USE (EU)

Item	N	Min	Max	Mean	St. Deviation	Likert Scale				
						Strongly Disagree	*Disagree*	*Less Agree*	*Agree*	*Strongly Agree*
EU1	40	1	5	4.33	0.944	0	2	1	15	22
EU2	40	1	5	4.43	0.931	1	1	3	10	25
EU3	40	1	5	4.55	0.904	0	1	2	7	30
EU4	40	1	5	4.68	0.859	0	1	1	4	34

TABLE II. EFFICIENCY (EF)

Item	N	Min	Max	Mean	St. Deviation	Likert Scale				
						Strongly Disagree	*Disagree*	*Less Agree*	*Agree*	*Strongly Agree*
Ef1	40	1	5	4.52	0.960	0	2	1	7	30
Ef2	40	1	5	4.38	1.055	1	1	2	10	26
Ef3	40	1	5	4.48	0.987	0	2	2	7	29
Ef4	40	1	5	4.60	0.871	0	1	1	7	31

TABLE III. FUNCTIONALITY (F)

Item	N	Min	Max	Mean	St. Deviation	Likert Scale				
						Strongly Disagree	*Disagree*	*Less Agree*	*Agree*	*Strongly Agree*
F1	40	3	5	4.78	0.480	0	0	1	7	32
F2	40	2	5	4.40	0.744	0	1	3	15	21
F3	40	2	5	4.52	0.679	0	1	1	14	24
F4	40	1	5	4.28	0.847	0	1	1	20	18

TABLE IV. SATISFACTION (S)

Item	N	Min	Max	Mean	St. Deviation	Likert Scale				
						Strongly Disagree	*Disagree*	*Less Agree*	*Agree*	*Strongly Agree*
S1	40	1	5	4.60	0.900	0	0	2	6	32
S2	40	4	5	4.58	0.501	0	0	0	17	23
S3	40	3	5	4.48	0.554	0	0	1	19	20
S4	40	3	5	4.73	0.506	0	0	1	9	30

TABLE V. RATING OF LAZADA MOBILE APPLICATION

Scale	Weighted Mean Range	Grading of Usability aspects
5	4.51 - 5.00	Very High
4	3.51 - 4.50	High
3	2.51 - 3.50	Medium
2	1.51 - 2.50	Less
1	1.00 - 1.50	Least

A. *Ease of use of Lazada Mobile Application*

Table I contains the descriptive statistics for ease of use of Lazada Indonesia mobile application and each item is indicated by the mean and standard deviation. The results obtained from the post-test questionnaire answered by the participants after they completed the exercise indicated that most of the participants strongly agreed that the application was very easy for them to use during the evaluation test that they were given by the moderators. Ease of use is very important in a way that if it deteriorates, the usefulness of the application also deteriorates and vice versa [41]. The responses from the "ease of use" questions were graded with a mean of 4.68 out of 5.00, and a standard deviation of 0.859. This is graded as "Very High" according to the rating scale in Table V.

The highest percentage of participants confessed that they did not find complications while they searched for the items, they easily interacted with the interface without any confusion. Only a few participants less agreed and this was because they were novice users, so they had never interacted with the Lazada mobile application before. The results portray that Lazada mobile application has a good ease of use design to the users. Usability is greatly associated with the ease of use of a mobile application, therefore, mobile application designers have to focus on the usability aspect as this can even affect the market of mobile devices [42]. If users cannot interact with any mobile application, the application's value degrades and will have fewer users. Basing on the results of this aspect, Mobile users are recommended to buy devices and install applications that are simple for them to use [43]. Mobile devices like smartphones and tablets contain applications that are intended for end-users as their main feature [44] thus, developers must take care of Human-Computer Interaction (HCI) aspects.

B. *Efficiency of Lazada Mobile Application*

The descriptive statistics for the efficiency of the use of Lazada mobile applications are displayed in table two. The efficiency of Lazada mobile application was graded with a mean of 4.60 out of 5.00, and a standard deviation of 0.871. According to Table V, this is also perceived as "Very High", indicating that the mobile application is very efficient in use. Similarly, most participants in the study agreed that Lazada mobile application is efficient to use. The highest number strongly agreed that the application did not require much thought from them to complete the task, they did not spend a lot of time on performing the task, and they were not confused on how to carry out the task. Most of them were able to complete the task successfully and confessed in the post-test completed questionnaire that they did not require extra time. Only a few novice participants found some issues because they were new users, however, they were also quick at understanding when they sought help. They were able to complete the task although with just a little help by explanation.

The answers from the questionnaire and result compilation show that the application is efficient while the users are using it. Without a doubt, one of the most important factors to be considered by designers is the efficiency of the mobile application, as it also allows users to utilize the energy of their mobile devices [45]. Therefore, designers should ensure that users take less time to complete their goals while using an application. The efficiency of the application means that it is clearly functioning [46], to carry out its purpose using minimal resources and thus finally bring satisfaction to users. This becomes the only reason as to why the rating of any product greatly depends on efficiency [47], because high efficiency attracts high ratings as it enhances satisfaction and vice versa.

C. *Functionality of Lazada Mobile Application*

Table III illustrates the functionality of Lazada mobile application according to the test that was carried out in this study. The functionality aspect attained a mean of 4.78 out of 5.00 and a standard deviation of 0.480. This is also graded as "Very High" according to the grading scale in Table V. The result, therefore, indicates that most of the participants agreed that Lazada mobile application has good functionality. In Table III, the mean and standard deviation of each item are shown and the result implies that the functionality of the interface was pleasing to the participants.

Most of them strongly agreed to statements in the questionnaire such as "the application is user friendly and allows free navigation of the user, "the features on the site are clear and understandable", "The application features were fast to respond", and the steps for making the purchase were not complex to perform. Users were seen to prefer the search button because their feedbacks were returning fast. The more the features respond fast the more the users enjoy the functionality of the application and thus attracting satisfaction in the long run [48]. Following the results, the functionality of a mobile application is very important as it attracts users to continue using the application [49]. If the functionality of the system is not nice, such as doesn't give free navigation to the users, users will leave frustrated and will not use the application again as it won't let them achieve their goals [50]. Poor functionality leaves users dissatisfied, and this may finally lead to the disloyalty of the application.

D. *User Satisfaction with Lazada Mobile Application*

In the last part of the activity, the researchers needed to know how much the users were satisfied with the mobile application. Table IV shows the descriptive statistics of user satisfaction of participants from using the Lazada Indonesia mobile application. The mean and standard deviation for each item are shown in the table.

Although most participants including new users agreed that Lazada mobile application and its features were an enjoyable application to use even for the novice users, they claimed they were not fully satisfied with the mobile application. According to the results, Lazada mobile application gives comfort to its users but not complete satisfaction [51]. Accumulated satisfaction can be attained by building the customers' trust in the mobile application [52], as this further greatly influences the intended purchase of users. Based on the results, it can be argued that satisfaction is a very important aspect of usability evaluation as it brings users positive thoughts and attraction to the mobile application. Satisfaction does not only depend on the mobile application itself but also depends on the mobile application store [53], i.e. service providers. Satisfaction gives users a feeling of fun and comfort while using the mobile application. Perceived satisfaction is the only key factor to

entice mobile users to engage in using online shopping mobile applications [54]. This is because satisfaction makes users develop a positive attitude and trust towards products sold and positive word-of-mouth that increases product loyalty and creates uniqueness in the mobile application.

In this study as we tested user satisfaction, participants pointed out some issues, which indicated that users were not fully satisfied with the operation of the mobile application. Novice users as well found some usability complications in carrying out the task such as locating the trolley while they were on the main page. In addition to that, the intermediate users had difficulties in making transactions using the bank transfer method, as it involved extensive steps. The other challenging issue was a deep understanding of the language by foreign participants since Lazada Indonesia uses Bahasa Indonesia. In addition to the mentioned issues, some experienced participants complained that the android application was slow in loading and displaying the trolley, and even when displayed, it required much time lag to perform any action where almost 90% of the time error message prompts displayed.

Other users complained that during their search of items as they performed the task, a lot of irrelevant items appeared in their search which was unpleasing. "It required a lot of time to locate the wanted item", the users confessed. And lastly, the participants complained about the second to last task they were asked to perform. They complained that it was not possible to check out the items in the trolley using the "pay on delivery" if items that could pay on delivery were checked out with items that were not allowed to pay on delivery. Users suggested that it would be better if items allowed to pay on delivery would be differentiated from items that are not allowed to pay on delivery if the user tried to check out, rather than denying the whole payment option of "pay on delivery". All these complaints confirmed to the researchers that users were not fully satisfied with the mobile application, thus designers need to attend to the issues first to increase user satisfaction.

V. CONCLUSION

This study was intended to understand the performance and issues facing Lazada Indonesia mobile application users through carrying out a usability test on the mobile application by further considering four usability aspects i.e. ease of use, efficiency, functionality, and satisfaction attained by users from using the mobile application. Overall, from the results obtained from data analysis, the performance of the mobile application proved good to most of the participants of the study. The mobile application showed good ease of use, adequate efficiency, and effectiveness, but did not provide complete satisfaction to the users.

Based on the results, users were not fully satisfied with the mobile application due to the following usability issues that were mentioned by some participants, i.e. (1) language barrier experienced by foreign users, (2) Slowness of the mobile application in loading the trolley, (3) Appearing of irrelevant items during product search by users, and (4) Failure to separate items that allow COD from those which don't allow COD during the checkout process. In conclusion, therefore, the designers of Lazada Indonesia mobile application need to

improve their design by solving the usability challenges experienced by some users, which in the actual sense could make a large number if considered out of the whole total population of Lazada Indonesia mobile application users. Solving users' complaints can make the application more marketable and trustworthy, and overall increase user satisfaction.

VI. LIMITATIONS AND FUTURE WORK

Like any other researches, this study encountered some limitations during the experiment. The study used only 40 participants for the experiment since it was not easy to gather many people as this study was carried out during the COVID-19 pandemic. Therefore, this limited the expected sample size. The researchers hope that the results would have been much better with bigger sample size. Since this study only concentrated on Lazada Indonesia mobile shopping application, future studies should consider testing the users of the Lazada Indonesia desktop application because the two are different in their nature of search. Furthermore, this test was carried out by mixing the users of different experiences of using the mobile application. Future research can be carried out by grouping and testing the users differently according to their levels of experience to understand the issues experienced by users at different levels of experience. This will help the designers to understand which usability problems to solve first and at what stage of using the application these usability problems are encountered, thus providing quick solutions to challenges and increase the satisfaction of Lazada Indonesia mobile application users.

REFERENCES

[1] N. Shaw and K. Sergueeva, "The non-monetary benefits of mobile commerce: Extending UTAUT2 with perceived value," Int. J. Inf. Manage., vol. 45, no. October 2018, pp. 44–55, 2019, doi: 10.1016/j.ijinfomgt.2018.10.024.

[2] E. Marinao-Artigas and K. Barajas-Portas, "Precedents of the satisfaction of mobile shoppers. A cross-country analysis," Electron. Commer. Res. Appl., vol. 39, p. 100919, 2020, doi: 10.1016/j.elerap.2019.100919.

[3] T. Natarajan, S. A. Balasubramanian, and D. L. Kasilingam, "Understanding the intention to use mobile shopping applications and its in fl uence on price sensitivity," J. Retail. Consum. Serv., vol. 37, no. March, pp. 8–22, 2017, doi: 10.1016/j.jretconser.2017.02.010.

[4] G. McLean and A. Wilson, "Shopping in the digital world: Examining customer engagement through augmented reality mobile applications," Comput. Human Behav., vol. 101, no. July, pp. 210–224, 2019, doi: 10.1016/j.chb.2019.07.002.

[5] S. Nuanmeesri, "Mobile application for the purpose of marketing, product distribution and location-based logistics for elderly farmers," Appl. Comput. Informatics, 2019, doi: 10.1016/j.aci.2019.11.001.

[6] R. Pitakaso, K. Sethanan, and T. Jamrus, "Hybrid PSO and ALNS algorithm for software and mobile application for transportation in ice manufacturing industry 3.5," Comput. Ind. Eng., vol. 144, p. 106461, 2020, doi: 10.1016/j.cie.2020.106461.

[7] S. Yoo et al., "Developing a mobile epilepsy management application integrated with an electronic health record for effective seizure management," Int. J. Med. Inform., vol. 134, p. 104051, 2020, doi: 10.1016/j.ijmedinf.2019.104051.

[8] R. Paramastri et al., "The use of mobile applicatiosn to improve nutrition behaviour: A systematic review," Comput. Methods Programs Biomed., vol. 192, p. 105459, 2020, doi: 10.1016/j.cmpb.2020.105459.

[9] P. Piran et al., "Medical Mobile Applications for Stroke Survivors and Caregivers," J. Stroke Cerebrovasc. Dis., vol. 28, no. 11, p. 104318, 2019, doi: 10.1016/j.jstrokecerebrovasdis.2019.104318.

[10] K. Valdes, E. Gendernalik, J. Hauser, and M. Tipton, "Use of mobile applications in hand therapy," J. Hand Ther., 2020, doi: 10.1016/j.jht.2019.10.003.

[11] M. Anshari and Y. Alas, "Smartphones habits, necessities, and big data challenges," J. High Technol. Manag. Res., vol. 26, no. 2, pp. 177–185, 2015, doi: 10.1016/j.hitech.2015.09.005.

[12] C. Campbell-Grossman, D. B. Hudson, K. M. Hanna, B. Ramamurthy, and V. Sivadasan, "Ease of Use and Acceptability of a Smartphone App for Young, Low-Income Mothers," J. Technol. Behav. Sci., vol. 3, no. 1, pp. 5–11, 2018, doi: 10.1007/s41347-017-0031-5.

[13] V. Sukhwani and R. Shaw, "Operationalizing crowdsourcing through mobile applications for disaster management in India," Prog. Disaster Sci., vol. 5, p. 100052, 2020, doi: 10.1016/j.pdisas.2019.100052.

[14] A. Razzaq, Y. T. Samiha, and M. Anshari, "Smartphone habits and behaviors in supporting students self-efficacy," Int. J. Emerg. Technol. Learn., vol. 13, no. 2, pp. 94–109, 2018, doi: 10.3991/ijet.v13i02.7685.

[15] J. V. Chen, A. Tran, and T. Nguyen, "Computers in Human Behavior Understanding the discontinuance behavior of mobile shoppers as a consequence of technostress : An application of the stress-coping theory," Comput. Human Behav., vol. 95, no. January, pp. 83–93, 2019, doi: 10.1016/j.chb.2019.01.022.

[16] S. N. Upadhe, D. D. Shinde, and U. S. Mugale, "Ease of Use Experimentation of Isometric Template," Procedia Manuf., vol. 20, pp. 296–299, 2018, doi: 10.1016/j.promfg.2018.02.044.

[17] X. Li, X. Zhao, W. (Ato) Xu, and W. Pu, "Measuring ease of use of mobile applications in e-commerce retailing from the perspective of consumer online shopping behaviour patterns," J. Retail. Consum. Serv., vol. 55, no. March, p. 102093, 2020, doi: 10.1016/j.jretconser.2020.102093.

[18] F. V. Haucke, "Smartphone-enabled social change: Evidence from the Fairphone case?," J. Clean. Prod., vol. 197, pp. 1719–1730, 2018, doi: 10.1016/j.jclepro.2017.07.014.

[19] A. A. Pinem, A. Yeskafauzan, P. W. Handayani, F. Azzahro, A. N. Hidayanto, and D. Ayuningtyas, "Designing a health referral mobile application for high-mobility end users in Indonesia," Heliyon, vol. 6, no. 1, p. e03174, 2020, doi: 10.1016/j.heliyon.2020.e03174.

[20] U. Akram, A. R. Ansari, G. Fu, and M. Junaid, "Feeling hungry? let's order through mobile! examining the fast food mobile commerce in China," J. Retail. Consum. Serv., vol. 56, no. April, p. 102142, 2020, doi: 10.1016/j.jretconser.2020.102142.

[21] R. Septiani, P. W. Handayani, and F. Azzahro, "Factors that Affecting Behavioral Intention in Online Transportation Service: Case study of GO-JEK," Procedia Comput. Sci., vol. 124, pp. 504–512, 2017, doi: 10.1016/j.procs.2017.12.183.

[22] A. Shankar and B. Rishi, "Convenience matter in mobile banking adoption intention?," Australas. Mark. J., no. xxxx, 2020, doi: 10.1016/j.ausmj.2020.06.008.

[23] S. H. Liao and L. L. Yang, "Mobile payment and online to offline retail business models," J. Retail. Consum. Serv., vol. 57, no. 151, p. 102230, 2020, doi: 10.1016/j.jretconser.2020.102230.

[24] D. B. Setyohadi, sri Kusrohmaniah, C. Efrans, T. D. Luciana, and P. S. Bening, "M-Learnıng Interface Desıgn Based On Emotıonal Aspect Analysıs," Conf. Pap. Lect. Notes Comput. Sci., vol. 1, no. January, pp. 15–26, 2017, doi: 10.1007/978-3-319-52503-7.

[25] T. Chen, "Assessing factors critical to smart technology applications to mobile health care − the fgm-fahp approach," Heal. Policy Technol., vol. 9, no. 2, pp. 194–203, 2020, doi: 10.1016/j.hlpt.2020.02.005.

[26] J. J. Hew, "Hall of fame for mobile commerce and its applications: A bibliometric evaluation of a decade and a half (2000-2015)," Telemat. Informatics, vol. 34, no. 1, pp. 43–66, 2017, doi: 10.1016/j.tele.2016.04.003.

[27] D. L. Kasilingam, "Understanding the attitude and intention to use smartphone chatbots for shopping," Technol. Soc., p. 101280, 2020, doi: 10.1016/j.techsoc.2020.101280.

[28] A. K. Y. Tang, "A systematic literature review and analysis on mobile apps in m-commerce: Implications for future research," Electron. Commer. Res. Appl., vol. 37, no. February, p. 100885, 2019, doi: 10.1016/j.elerap.2019.100885.

[29] S. Sohn, "A contextual perspective on consumers' perceived usefulness: The case of mobile online shopping," J. Retail. Consum. Serv., vol. 38, no. April, pp. 22–33, 2017, doi: 10.1016/j.jretconser.2017.05.002.

[30] K. Y. Goh, J. Chu, and J. Wu, "Mobile advertising: An empirical study of temporal and spatial differences in search behavior and advertising response," J. Interact. Mark., vol. 30, pp. 34–45, 2015, doi: 10.1016/j.intmar.2014.12.002.

[31] J. Vithayathil, M. Dadgar, and J. K. Osiri, "International Journal of Information Management Social media use and consumer shopping preferences," Int. J. Inf. Manage., vol. 53, no. October 2019, p. 102117, 2020, doi: 10.1016/j.ijinfomgt.2020.102117.

[32] A. Y. L. Chong, "Mobile commerce usage activities: The roles of demographic and motivation variables," Technol. Forecast. Soc. Change, vol. 80, no. 7, pp. 1350–1359, 2013, doi: 10.1016/j.techfore.2012.12.011.

[33] M. Kim, J. Kim, J. Choi, and M. Trivedi, "Mobile Shopping Through Applications: Understanding Application Possession and Mobile Purchase," J. Interact. Mark., vol. 39, pp. 55–68, 2017, doi: 10.1016/j.intmar.2017.02.001.

[34] E. Pantano, "Innovation drivers in retail industry," Int. J. Inf. Manage., vol. 34, no. 3, pp. 344–350, 2014, doi: 10.1016/j.ijinfomgt.2014.03.002.

[35] Y. Lee and H. Y. Kim, "Consumer need for mobile app atmospherics and its relationships to shopper responses," J. Retail. Consum. Serv., vol. 51, no. June 2017, pp. 437–442, 2019, doi: 10.1016/j.jretconser.2017.10.016.

[36] F. J. Martínez-López, Y. Li, H. Liu, and C. Feng, "Do safe buy buttons and integrated path-to-purchase on social platforms improve users' shopping-related responses?," Electron. Commer. Res. Appl., vol. 39, no. October 2019, p. 100913, 2020, doi: 10.1016/j.elerap.2019.100913.

[37] A. Kaur and K. Kaur, "A COSMIC function points based test effort estimation model for mobile applications," J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2019, doi: 10.1016/j.jksuci.2019.03.001.

[38] N. Bento, "Calling for change? Innovation, diffusion, and the energy impacts of global mobile telephony," Energy Res. Soc. Sci., vol. 21, pp. 84–100, 2016, doi: 10.1016/j.erss.2016.06.016.

[39] S. Meedya et al., "Developing and testing a mobile application for breastfeeding support : The Milky Way application," Women and Birth, no. 2019, 2020, doi: 10.1016/j.wombi.2020.02.006.

[40] S. Clavijo-Buendía et al., "Construct validity and test-retest reliability of a free mobile application for spatio-temporal gait analysis in Parkinson's disease patients," Gait Posture, vol. 79, pp. 86–91, 2020, doi: 10.1016/j.gaitpost.2020.04.004.

[41] A. M. Lund, "Measuring usability with the USE questionnaire," Usability interface, vol. 8, no. 2, pp. 3–6, 2001, doi: 10.1177/1078087402250360.

[42] M. A. Uddin, H. Xu, and M. T. Azim, "Factors affecting mobile handset (MH) buying decision: An empirical study," Int. J. Manag. Bus. Res., vol. 5, no. 3, pp. 225–236, 2015.

[43] M. Dunlop and S. Brewster, "The challenge of mobile devices for human computer interaction," Pers. Ubiquitous Comput., vol. 6, no. 4, pp. 235–236, 2002, doi: 10.1007/s007790200022.

[44] A. Meneses-Viveros, E. Hernández-Rubio, S. Mendoza, J. Rodríguez, and A. B. Márquez Quintos, "Energy saving strategies in the design of mobile device applications," Sustain. Comput. Informatics Syst., vol. 19, pp. 86–95, 2018, doi: 10.1016/j.suscom.2018.07.011.

[45] K. Yan, J. Tan, and X. Fu, "Improving energy efficiency of mobile devices by characterizing and exploring user behaviors," J. Syst. Archit., vol. 98, no. November 2018, pp. 126–134, 2019, doi: 10.1016/j.sysarc.2019.07.004.

[46] A. Hussain, E. O. C. Mkpojiogu, N. H. Jamaludin, and S. T. L. Moh, "A usability evaluation of Lazada mobile application," AIP Conf. Proc., vol. 1891, no. October, 2017, doi: 10.1063/1.5005392.

[47] A. Hussain, E. O. C. Mkpojiogu, J. Musa, and S. Mortada, "A user experience evaluation of Amazon Kindle mobile application," AIP Conf. Proc., vol. 1891, 2017, doi: 10.1063/1.5005393.

[48] Z. Mushtaq and A. Wahid, "Inclusion of Functional and Non-Functional Parameters for the Prediction of Overall Efforts of Mobile Applications.," Comput. Stand. Interfaces, vol. 71, p. 103404, 2020, doi: 10.1016/j.csi.2019.103404.

[49] A. J. G. Silvius and C. M. Silvius, "Exploring Functionality of Mobile Applications for Project Management," Procedia Comput. Sci., vol. 64, pp. 343–351, 2015, doi: 10.1016/j.procs.2015.08.498.

[50] M. Rezae, N. Chen, D. McMeekin, T. Tan, A. Krishna, and H. Lee, "The evaluation of a mobile user interface for people on the autism spectrum: An eye movement study," Int. J. Hum. Comput. Stud., vol. 142, no. May, p. 102462, 2020, doi: 10.1016/j.ijhcs.2020.102462.

[51] A. Hussain, E. O. C. Mkpojiogu, H. Abubakar, and H. M. Hassan, "The usability evaluation of Mudah.my on mobile device," AIP Conf. Proc., vol. 1891, 2017, doi: 10.1063/1.5005391.

[52] W. T. Wang, W. M. Ou, and W. Y. Chen, "The impact of inertia and user satisfaction on the continuance intentions to use mobile communication applications: A mobile service quality perspective," Int. J. Inf. Manage., vol. 44, no. May 2018, pp. 178–193, 2019, doi: 10.1016/j.ijinfomgt.2018.10.011.

[53] J. Song, J. Kim, D. R. Jones, J. Baker, and W. W. Chin, "Application discoverability and user satisfaction in mobile application stores: An environmental psychology perspective," Decis. Support Syst., vol. 59, no. 1, pp. 37–51, 2014, doi: 10.1016/j.dss.2013.10.004.

[54] F. Luna-Perejon et al., "Evaluation of user satisfaction and usability of a mobile app for smoking cessation," Comput. Methods Programs Biomed., vol. 182, p. 105042, 2019, doi: 10.1016/j.cmpb.2019.105042.

Recognition of Local Birds of Bangladesh using MobileNet and Inception-v3

Md. Mahbubur Rahman[1], Al Amin Biswas[2], Aditya Rajbongshi[3], Anup Majumder[4]

Crowd Realty, Tokyo, Japan[1]

Dept. of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh[2]

Dept. of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh[3, 4]

Abstract—**Recognition of bird species can be a challenging task due to various complex factors. The purpose of this work is to distinguish various local bird species of Bangladesh from the image data. The MobileNet and Inception-v3 model which is mainly an image classification model used here to accomplish this work. Here, we have used a total of four approaches namely Inception-v3 without transfer learning, Inception-v3 with transfer learning, MobileNet without transfer learning, and MobileNet with transfer learning to accomplish the task. To evaluate our experimental results, we have calculated F1 Score besides the model's accuracy and also presented the ROC curve to evaluate the model's output quality. Then we have done a comparison among the applied four approaches. The experimental result has proved the working capability of the applied four approaches. Among these four approaches, MobileNet with transfer learning outperforms the others and obtained a test accuracy of 91.00%. For each of the classes, MobileNet with transfer learning obtained the highest F1 Score than other approaches.**

Keywords—*Recognition; MobileNet; Inception-v3; transfer learning; computer vision; Bangladeshi bird*

I. Introduction

The bird is one of the amazing creations of God. Besides adding charm and beauty to nature, they also help to sustain a balance of the fresh environment in the world. Bangladesh has a friendly environment to live for the birds. It is very difficult to recognize the local birds because of the bird's nature, color, structure, and voice.

Images can be classified in several ways. Image classification aims to category a distinct picture according to a set of probable categories. From a deep learning aspect [1], the image classification problem can be solved by transfer learning. This learning technique is becoming more popular day by day in computer vision because of its ability to construct accurate models within a very short time [2]. The main advantages of transfer learning techniques [3], there is no need for a huge training dataset and also no need for much more computational power. In computer vision, Transfer learning is normally exposed through the practice of pre-trained models. A pre-trained model [1] is a technique that was trained on a huge dataset to resolve a problem similar to the one problem that individual desires to solve.

Here, we have mainly performed the task of recognition of the seven birds. Seven local birds of Bangladesh namely Bulbuli(বুলবুলি), Chorui(চড়ুই), Kaak(কাক), Doel(দোয়েল), Machranga(মাছরাঙ্গা), Shalik(শালিক), and Kokil(কোকিল) including the national birds which are shown in the Fig. 1. In this paper, we have implemented two models (two different approaches for each of the models) to accomplish this work of local bird recognition in Bangladesh. Firstly, we have performed Inception-v3 and MobileNet without using transfer technology and then we have again performed Inception-v3 and MobileNet using transfer technology. Then we have compared our observations on the basis of performance evaluation metrics namely accuracy and F1 Score.

This paper is divided into five sections and presented as follows: Literature Review is presented in Section II. Section III describes the Methodology of this work. Experimental Result and Analysis are presented in Section IV. Conclusion is presented in Section V. Lastly; Future Work is mentioned in Section VI, respectively.

(a) বুলবুলি   (b) চড়ুই   (c) কাক   (d) দোয়েল

(e) মাছরাঙ্গা   (f) শালিক   (g) কোকিল

Fig. 1. An Introduction of Seven Local Birds.

II. Literature Review

Much research has been worked on transfer learning by many researchers but few have been applied this learning technique on local birds of Bangladesh. Many researchers have worked on bird recognition with many traditional diverse solutions.

A method of bird recognition based on Inception-v3 is presented by J. Bai et al. [4]. To improve the robustness and generalization of the model, they have applied some of the data augmentation techniques. Lastly, they evaluated their system in BirdCLEF test data and obtained 0.055 of classification mean average precision.

A deep learning neural network technique is presented to identify the bird species by S.K. Pillai et al. [5]. To implement this work, they used the Tensor Flow framework and twelve bird species were taken from the Caltech dataset to test their system. From their result, the true prediction (training dataset) as well as true prediction (unknown image) is always more than 0.5 and in some cases, both the values are more than 0.9 which proves the accuracy of the algorithm. For this dataset, train accuracy starts from 29% (at step 0) and goes till 100% (at step 1500). For their work, it is observable that the relationship between validation accuracy and the cross-entropy is inversely proportional.

M. Lasseck [6] presented deep learning methods for the detection of the acoustic bird. Deep Convolutional Neural Networks originally outlined for the classification of the image, are adjusted and fine-tuned to recognize the bird's presence in audio recordings. To improve model performance, various data augmentation techniques were implemented here.

N. R. Gavai et al. [7] have presented the experimental performance of the MobileNets model to retrain the flower kind datasets, which can exceedingly reduce the time and space for flower classification compromising the accuracy insignificantly.

J. Bankar and N. R. Gavai [8] used a machine learning technique to classify the animal into particularized classes. They used the transfer learning technology to retrain the animal kind datasets based on the Inception-v3 model in the TensorFlow platform, which can exceedingly enhance the accuracy of classification in this circumstance.

Detection of pest birds in the field of agriculture is addressed by S. Lee et al. [9]. They proposed adapting deep learning techniques to certain cropped small areas of a frame where there is a huge chance of bird presence, based on the image processing's result. By conducting the background subtraction based on Gaussian Mixture Model, the moving objects are extracted. After that, color extraction and median filter eliminate the undesirable things in the agricultural environment. Lastly, the neural network object classifier was used to classify and minimize moving objects. The test results showed that employing neural networks to image's precise areas causes greater accuracy than the entire original frame.

S. S. Londhe and S. S. Kanade [10] studied the way of automatic bird's species identification by their vocalization. Bird sounds are classed by their function into songs and calls which are moreover subdivided into hierarchical levels of phrases, syllables, and elements. It is pointed out here that the syllable is a suitable unit to recognize bird species. Variety within diverse types of syllables birds are capable of producing is large. To identify species, the Support Vector Machine is chosen here.

A real-time detection system of birds in flight based on background subtraction and tracking through point correspondence is described by Moein Shakeri and Hong Zhang [11]. They have used a single fixed camera for their bird detection system and used Zivkovic's background subtraction technique. They appended a correspondence component based on point-tracking to the background subtraction technique to reach reliable bird detection. Investigations were carried to analyze the detection performance accepting objects of varied sizes, colors, and velocity. The obtained results showed the accuracy and efficiency of their system.

M. R. Islam et al. [12] developed a MobileNet model for local bird classification using CNN. They have used a small dataset. They utilized 500 images in 5 classes of local birds. Their obtained accuracy is about 100%.

X. Xia et al. [13] utilized the Inception-v3 model for the classification of flowers. For retraining the datasets, transfer learning technology was practiced. Two types of datasets namely Oxford-17 and Oxford -102 flowers were employed and the classification process was conducted using four steps. 95% percent accuracy is found for the dataset Oxford-17 and 94% for the dataset Oxford-102.

Howard et al. [14] presented a model which is the application of a mobile vision named ModelNet. The MobileNet architecture is based on streamlined. For choosing the right sized model based on the problems, they mentioned two hyperparameters which are global. MobileNet effectiveness is also demonstrated by focusing on various classification problems.

III. Methodology

This section represents the way we have used to implement this work. This section basically consists of five subsections namely Dataset, Data Preprocessing, Model Description, Training, and lastly Testing. The implementation procedure of our work is presented in Fig. 2.

The detailed description of the abovementioned five subsections is presented below.

A. Dataset

In this work, we have collected seven local birds' images. Details of the seven local birds with mentioning the local name, English name, and also scientific name is presented in Table I. Primarily, we have collected 100 images for each bird species. As we are working with seven types of local birds. So, our total collected data is 700 image data.
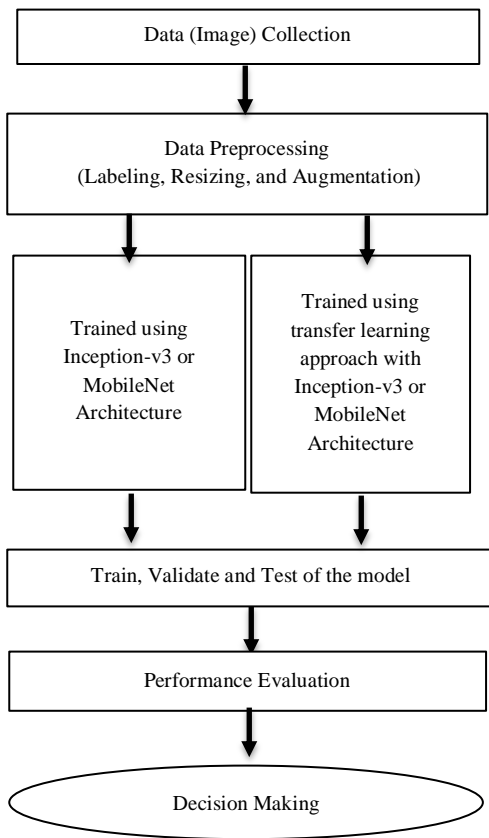
Fig. 2. The Implementation Procedure of Recognition of Local Birds of Bangladesh using MobileNet and Inception-v3.

TABLE I. REPRESENTATION OF THE SEVEN LOCAL BIRDS

Bird's Local Bangla Name	Bird's English Name	Bird's Scientific Name
বুলবুলি পাখি (Bulbuli)	Red-vented bulbul	*Pycnonotus cafer*
চড়ুই পাখি (Chorui)	House Sparrow	*Passer domesticus*
কাক পাখি (Kaak)	Crow	*Corvus corax*
দোয়েল পাখি (Doel)	Magpie	*Copsychus saularis*
মাছরাঙ্গা পাখি (Machranga)	Common kingfisher	*Alcedo atthis*
শালিক পাখি (Shalik)	Myna	*Acridotheres tristis*
কোকিল পাখি (Kokil)	Cuckoo	*Cuculus canorus*

B. Data Preprocessing

The data preprocessing step is included with data resizing, labeling, and augmentation. Since, in deep learning, we need a sufficient amount of data but in some cases, it is not feasible to collect. So, in this type of case, data augmentation techniques can help individuals to rescue. So, to increase the

size of the dataset, we have performed the data augmentation technique on the original dataset. Data augmentation can be done by several operations. The most commonly used operations are rotation, shearing, zooming, cropping, flipping, and contrast adjustment using histogram equalization, adaptive histogram equalization technique. We got 500 images for each class after augmentation. So, after the augmentation of data, we have a total of 3500 image data. We have resized this augmented label dataset for training and testing purposes. Then, we divided the data randomly into two sets (train-tests) with an 80% to 20% ratio.

C. Model Description

Here, we worked on Inception-v3 and MobileNet architecture in two different ways. One, we have used Inception-v3 and MobileNet architectures without transfer learning. And another way is using the transfer learning approach and for this, we used the pre-trained weight of these architectures on the imagenet dataset of 1000 classes. The details of MobileNet and Inception-v3 architectures for the imagenet dataset are presented in Table II. For our training purpose, we discarded the output layer of 1000 neurons and added an output layer of 7 neurons as in our dataset we have only worked on the 7 classes. During transfer learning with Inception-v3, we make the top 2 inception blocks freeze and the rest of the layer trainable. We made all the layers trainable when we did transfer learning with MobileNet architecture. As we were getting high overfitting with transfer learning, we used a dropout layer with a rate of 0.5 after the final fully connected layer to reduce the overfitting.

D. Training

To train the models, we used an Nvidia Geforce 2080 GPU with 8GB Memory. We trained the data in 100 epochs with a batch size of 8. For transfer learning, we used the pre-trained weight on the imagenet dataset. During training, we used the test set as the validation set. The accuracy with respect to epochs for all the models is presented in the Fig. 3. Blueline is used here to present the training accuracy and the other one is used here to present the validation accuracy.

E. Testing

After training, we tested the newly trained models with the test data. We got the following 7*7 confusion matrix for the class set {'Bulbuli', 'Chorui', 'Kaak', 'Doel', 'Machranga', 'Shalik', 'Kokil'} which is shown in Table III.

The binary representation of the 7*7 confusion matrix shown in Table III is presented in Table IV. A Binary matrix is easy to interpret and understand. As 700 images are used to testing so, the summation of TP, FP, FN, and TN is equal to 700 for each of the bird classes.

TABLE II.   MOBILENET AND INCEPTION-V3 ARCHITECTURE

MobileNet			Inception-v3		
Type/Stride	Filter Shape	Input Size	Type/ Stride	Patch size	Input size
Conv / s2	3×3×3×32	224×224×3	Conv /s2	3×3	299×299×3
Conv dw / s1	3×3×32 dw	112×112×32	Conv /s1	3×3	149×149×32
Conv / s1	1×1×32×64	112×112×32	Conv padded /s1	3×3	147×147×32
Conv dw / s2	3×3×64 dw	112×112×64	Pool /s2	3×3	147×147×64
Conv / s1	1×1×64×128	56×56×64	Conv / s1	3×3	73×73×64
Conv dw / s1	3×3×128 dw	56×56×128	Conv / s2	3×3	71×71×80
Conv / s1	1×1×128×128	56×56×128	Conv /s1	3×3	35×35×192
Conv dw /s2	3×3×128 dw	56×56×128	3×Inception	Inception Module	35×35×288
Conv / s1	1×1×128×256	28×28×128	5×Inception	Inception Module	17×17×768
Conv dw /s1	3×3×256 dw	28×28×256	2×Inception	Inception Module	8×8×1280
Conv / s1	1×1×256×256	28×28×256	Avg Pool	Pool 8×8	8×8×2048
Conv dw / s2	3×3×256 dw	28×28×256	FC	2048 × 1000	1×1×2048
Conv / s1	1×1×256×512	14×14×256	Softmax	classifier	1×1×1000
Conv dw/s1 5× Conv / s1.	3×3×512 dw 1×1×512×512	14×14×512 14×14×512			
Conv dw /s2	3×3×512 dw	14×14×512			
Conv / s1	1×1×512×1024	7×7×512			
Conv dw /s2	3×3×1024 dw	7×7×1024			
Conv / s1	1×1×1024×1024	7×7×1024			
Avg Pool	Pool 7×7	7×7×1024			
FC	1024×1000	1×1×1024			
Softmax	Classifier	1×1×1000			
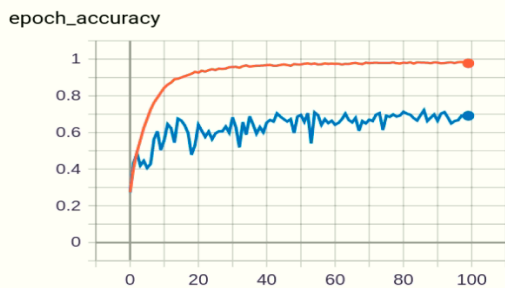
(a) Inception-v3 with Transfer Learning.
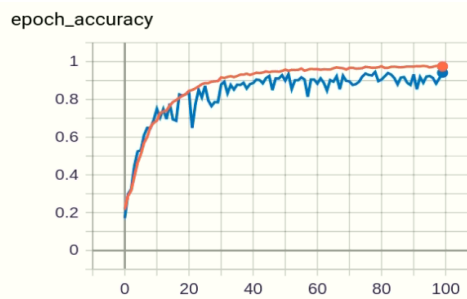
(b). Inception-v3 without Transfer Learning.
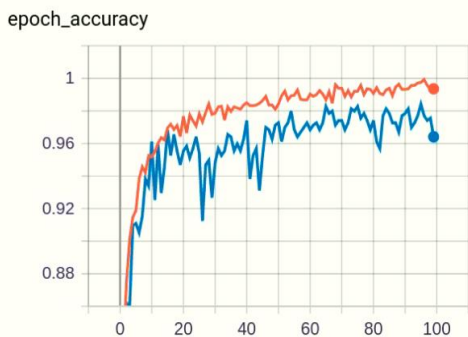
(c). MobileNet with Transfer Learning.
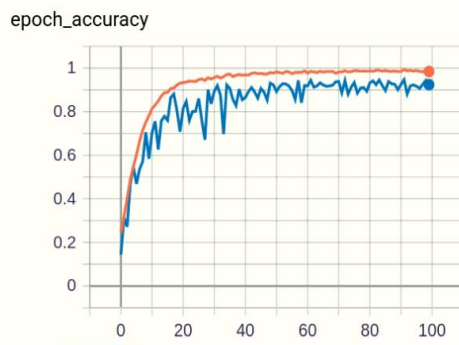
(d). MobileNet without Transfer Learning.

Fig. 3.   Accuracy over the Epochs of different Models.

TABLE III.   CONFUSION MATRIX BASED ON THE TEST SET

Method: Inception-v3 without transfer learning	Predicted						
	Bulbuli	**Chorui**	**Kaak**	**Doel**	**Machranga**	**Shalik**	**Kokil**
Actual **Bulbuli**	87	8	3	0	0	1	1
Chorui	1	82	10	0	0	3	4
Kaak	0	0	87	6	0	1	6
Doel	0	0	0	91	8	0	1
Machranga	0	0	3	0	89	5	3
Shalik	2	1	1	1	2	82	11
Kokil	8	2	7	2	0	0	81
Method: MobileNet without transfer learning	**Predicted**						
	Bulbuli	**Chorui**	**Kaak**	**Doel**	**Machranga**	**Shalik**	**Kokil**
Actual **Bulbuli**	86	8	2	1	0	0	3
Chorui	1	88	9	1	0	0	1
Kaak	0	0	84	9	0	0	7
Doel	0	0	2	90	8	0	0
Machranga	1	0	2	2	88	5	2
Shalik	1	1	6	1	0	81	10
Kokil	8	0	8	0	0	0	84
Method: Inception-v3 with transfer learning	**Predicted**						
	Bulbuli	**Chorui**	**Kaak**	**Doel**	**Machranga**	**Shalik**	**Kokil**
Actual **Bulbuli**	53	8	8	4	5	11	11
Chorui	1	64	15	4	3	12	1
Kaak	6	0	76	5	0	6	7
Doel	1	2	3	71	14	8	1
Machranga	6	10	1	4	64	12	3
Shalik	4	5	7	3	6	64	11
Kokil	9	1	7	2	4	7	70
Method: MobileNet with transfer learning	**Predicted**						
	Bulbuli	**Chorui**	**Kaak**	**Doel**	**Machranga**	**Shalik**	**Kokil**
Actual **Bulbuli**	90	7	2	0	0	0	1
Chorui	0	91	8	0	0	0	1
Kaak	0	0	92	7	0	0	1
Doel	0	0	1	90	8	0	1
Machranga	0	0	1	0	92	7	0
Shalik	0	0	1	0	1	90	8
Kokil	8	0	0	0	0	0	92

TABLE IV.   CONFUSION MATRIX AS A BINARY FORMAT

Class Name	Method	True Positive	False Positive	False Negative	True Negative
বুলবুলি (Bulbuli)	Inception-v3 without transfer learning	87	11	13	589
চড়ুই (Chorui)		82	11	18	589
কাক (Kaak)		87	24	13	576
দোয়েল (Doel)		91	9	9	591
মাছরাঙ্গা (Machranga)		89	10	11	590
শালিক (Shalik)		82	10	18	590
কোকিল (Kokil)		81	26	19	574
বুলবুলি (Bulbuli)	MobileNet without transfer learning	86	11	14	589
চড়ুই (Chorui)		88	9	12	591
কাক (Kaak)		84	29	16	571
দোয়েল (Doel)		90	14	10	586
মাছরাঙ্গা (Machranga)		88	8	12	592
শালিক (Shalik)		81	5	19	595
কোকিল (Kokil)		84	23	16	577
বুলবুলি (Bulbuli)	Inception-v3 with transfer learning	53	27	47	573
চড়ুই (Chorui)		64	26	36	574
কাক (Kaak)		76	41	24	559
দোয়েল (Doel)		71	22	29	578
মাছরাঙ্গা (Machranga)		64	32	36	568
শালিক (Shalik)		64	56	36	544
কোকিল (Kokil)		70	34	30	566
বুলবুলি (Bulbuli)	MobileNet with transfer learning	90	8	10	592
চড়ুই (Chorui)		91	7	9	593
কাক (Kaak)		92	13	8	587
দোয়েল (Doel)		90	7	10	593
মাছরাঙ্গা (Machranga)		92	9	8	591
শালিক (Shalik)		90	7	10	593
কোকিল (Kokil)		92	12	8	588

IV. EXPERIMENTAL RESULT AND ANALYSIS

The following formula ((1) to (4)) is used to calculate the performance evaluation metrics.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \qquad (1)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \qquad (2)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \qquad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \qquad (4)$$

In Table V, we can see that MobileNet with Transfer Learning among the four approaches achieved the highest accuracy of 91.00%. We have also calculated another performance evaluation metrics F1 Score of the applied four approaches for each of the seven classes. From Table V, we found that F1 Score of MobileNet with the transfer learning approach for the class বুলবুলি পাখি (Bulbuli), চড়ুই পাখি (Chorui), কাক পাখি (Kaak), দোয়েল পাখি (Doel), মাছরাঙ্গা পাখি (Machranga), শালিক পাখি (Shalik), and কোকিল পাখি (Kokil) is 90.91%, 91.92%, 89.76%, 91.37%, 91.54%, 91.37%, and 90.20%, respectively. Here, MobileNet with the transfer learning approach has outperformed than other approaches for each of the classes.

Receiver Operating Characteristic (ROC) metric is usually used to evaluate the quality of classifier's output. In the ROC curve, true positive rate places on the Y-axis and on the other hand false positive rate places on the X-axis. This denotes that the top left corner of the plot is the ideal point because here FPR is lowest, and TPR is highest. This is not very practical, but it does indicate that a larger AUC is usually better. From Fig. 4, we can see that the ROC of MobileNet with transfer learning approaches has larger AUC than others which indicates that this approach is working better here.

TABLE V.   RESULT ANALYSIS BASED ON PERFORMANCE EVALUATION METRICS

	Without Transfer Learning		With Transfer Learning	
	Inception-v3	MobileNet	Inception-v3	MobileNet
Accuracy (Test)	85.57%	85.86%	66.00%	91.00%
F1 Score [বুলবুলি (Bulbuli)]	87.88%	87.31%	58.89%	90.91%
F1 Score [চড়ুই (Chorui)]	84.97%	89.34%	67.37%	91.92%
F1 Score [কাক (Kak)]	82.46%	78.87%	70.05%	89.76%
F1 Score [দোয়েল (Doel)]	91.00%	88.24%	73.58%	91.37%
F1 Score [মাছরাঙ্গা (Machranga)]	89.45%	89.80%	65.31%	91.54%
F1 Score [শালিক(Shalik)]	85.42%	87.10%	58.18%	91.37%
F1 Score [কুকিল (Kokil)]	78.26%	81.16%	68.63%	90.20%
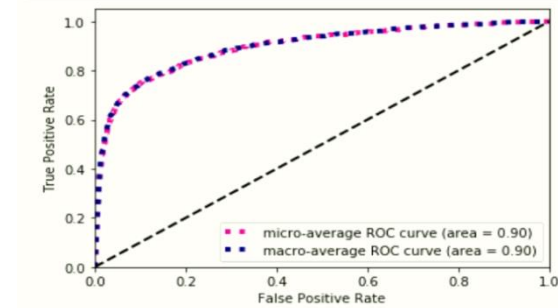
(a) Roc of Inception-v3 with Transfer Learning.
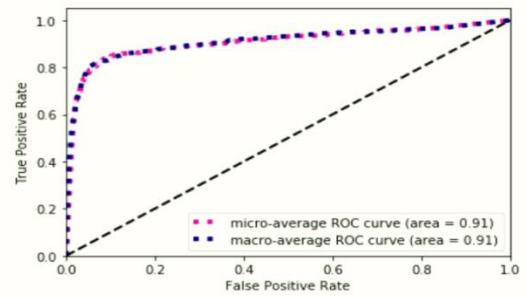
(b) Roc of Inception-v3 without Transfer Learning.

(c) ROC of MobileNet with Transfer Learning.

(d). ROC of MobileNet without Transfer Learning.

Fig. 4.   ROC Curve of Four different Approaches.

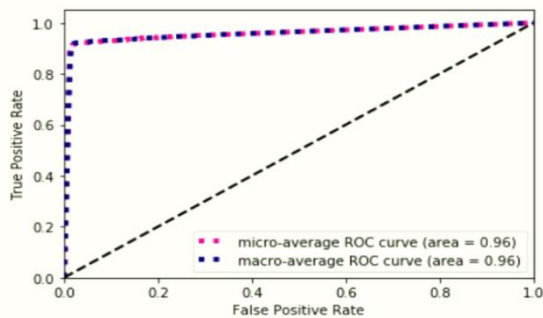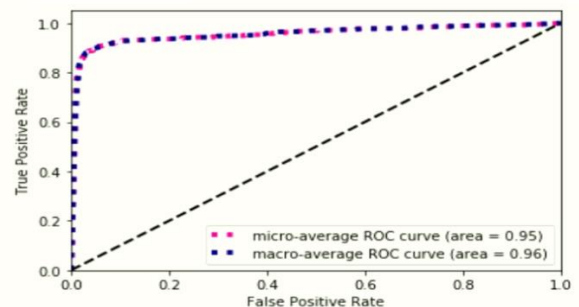ROC curves are usually used in binary classification to analyze the output of a classifier. In order to extend the ROC curve and ROC area to multi-label classification, it is essential to binarize the output. The two evaluation measures for multi-label classification are micro-averaging and macro-averaging. This is possible to draw a ROC curve per label, but it is also possible to draw a ROC curve by considering each element of the label indicator matrix as a binary prediction (micro-averaging). Macro-averaging, basically known as another evaluation measure for multi-label classification which gives equal weight to the classification of each label.

V.   CONCLUSION

This research work is mainly done on the seven local bird's species of Bangladesh, namely, বুলবুলি পাখি (Bulbuli), চড়ুই পাখি (Chorui), কাক পাখি (Kaak), দোয়েল পাখি (Doel), মাছরাঙ্গা পাখি (Machranga), শালিক পাখি (Shalik), and কোকিল পাখি (Kokil). Here, we have applied four approaches. To accomplish the work, we have considered 2800 images data as a training set and 700 images data as a test set. Among the four approaches, MobileNet with transfer learning performed better here in terms of performance evaluation metrics which is shown in Table V.

VI.   FUTURE WORK

There are so many local bird species in Bangladesh. So, in the future, we will work with more local bird species with many other transfer learning techniques. Also, we will work on the recognition of the local bird species from the video data in the future.

REFERENCES

[1]   "Transfer learning from pre-trained models," Available Online: https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751 [Last Access: 26 December 2019].

[2] "Transfer Learning for Image Classification in Keras ," Available Online: https://towardsdatascience.com/transfer-learning-for-image-classification-in-keras-5585d3ddf54e [Last Access: 26 December 2019].

[3] "Deep Learning For Beginners Using Transfer Learning In Keras," Available Online: https://towardsdatascience.com/keras-transfer-learning-for-beginners-6c9b8b7143e [Last Access: 26 December 2019].

[4] J. Bai, B. Wang, C. Chen, Z. H. Fu, and J. Chen, "Inception-v3 Based Method of LifeCLEF 2019 Bird Recognition," CLEF working notes, 2019.

[5] S. K. Pillai, M. M. Raghuwanshi, and U. Shrawankar, "Deep Learning Neural Network for Identification of Bird Species," In Computing and Network Sustainability, vol. 75, pp. 291-298, 2019.

[6] M. Lasseck, "Acoustic Bird Detection with Deep Convolutional Neural Networks, "in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), pp. 143-147, 2018.

[7] N. R. Gavai, Y. A. Jakhade, S. A. Tribhuvan, and R. Bhattad, "MobileNets for Flower Classification Using TensorFlow," International Conference on Big Data, IoT and Data Science (BID), pp. 154-158, 2017.

[8] J. Bankar, and N. R. Gavai, "Convolutional Neural Network-based Inception v3 Model for Animal Classification," International Journal of Advanced Research in Computer and Communication Engineering, vol. 7, no. 5, pp. 142-146, 2018.

[9] S. Lee, M. Lee, H. Jeon, and A. Smith, "Bird Detection in Agriculture Environment using Image Processing and Neural Network," In 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), pp. 1658-1663, 2019.

[10] S. S. Londhe, and S. S. Kanade, "Bird Species Identification Using Support Vector Machine," International Journal of Computer Science and Mobile Computing, vol. 4, no. 9, pp. 125-135, 2015.

[11] M. Shakeri, and H. Zhang, "Real-Time Bird Detection Based on Background Subtraction," In Proceedings of the 10th World Congress on Intelligent Control and Automation, pp. 4507-4510, 2012.

[12] M. R. Islam, N. Tasnim, and S. B. Shuvo, "MobileNet Model for Classifying Local Birds of Bangladesh from Image Content Using Convolutional Neural Network," In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-4. IEEE, 2019.

[13] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," In 2017 2nd International Conference on Image, Vision and Computing (ICIVC), pp. 783-787. IEEE, 2017.

[14] A. G. Howard., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

EDES-ACM: Enigma Diagonal Encryption Standard Access Control Model for Data Security in Cloud Environment

Sameer[1]

Ph.D Scholar

Department of Computer Science and Applications

Chaudhary Devi Lal University, Sirsa-125055(India)

Dr. Harish Rohil[2]

Associate Professor

Department of Computer Science and Applications

Chaudhary Devi Lal University, Sirsa-125055(India)

Abstract—The data management across the different domains is the foremost requirement for many organizations universally. The organization establishes the cloud computing paradigm to handle the data effectively due to its robust scaling at a low cost. In recent times the usage of cloud and its data is increasing with the multiuser environments. This resulted in the issue of ensuring the security of data in the cloud environment uploaded by the owners. The cloud service providers and researchers implemented several schemes to ensure data security. However, the task of providing security with multiuser remains tedious with data leakage. A novel Enigmatic Diagonal Encryption Standard (EDES) algorithm to provide access control over the cloud is proposed. The framework with the proposed algorithm is named as EDES-ACM. The Inverse Decisional Diffie Hellman (IDDH) technique is used for generating the signature of the group. The data is encrypted with the EDES algorithm by the data owner. The encrypted data is provided to the user and is accessed by the EDES based private key. The group manager monitors the cloud and provides the activity report to owners based on which the revocation is performed. The framework is validated for its performance on security parameters and compared with the existing models on computation cost. The EDES-ACM framework is effective with low computation cost. The future notion for the proposed framework is to include the block chain technology that may improve the security and better accumulation of data.

Keywords—*Cloud; security; multiuser; EDES-ACM; computation cost*

I. INTRODUCTION

Cloud computing (CC) has emerged in the various platforms to provide services at flexible time and place with internet connection [1]. Many organizations and enterprises establish cloud as the effective solution with low cost, rapid scaling. The only concern for the data owners is that the cloud service providers have to guarantee the privacy and security of data to remain secure and confidential among the dynamic cloud environment [2]. Most of the CSP's rely on the third-party operation in storing the data that are case sensitive. Henceforth the cloud user must trust the CSP or employ the effective system in the cloud infrastructure to improve the security of data in the cloud [3]. Several standards for cloud security are developed and suggested by standard bodies like the National Institute for Standards and Technology (NIST) [4]. However, the Multi-user cloud environment demands

distribution and sharing with integrating applications that are available either internally or externally. These aspects pose a critical challenge over the integrity, authenticity of cloud data [5].

Many novel algorithms are proposed for securing the data in the multi-user cloud environment to confirm the data integrity with privacy preserving approaches [6] and several other approaches. One among the effective approach is to provide either the symmetric or asymmetric group key. Using the keys, the data owners can encrypt the data and store it in the cloud. The user decrypts the data after downloading to access it. In symmetric keys, both cryptic keys are similar and it is dissimilar in unsymmetrical keys [7]. Even with cryptic keys, the data security is under threat with possible key leakage and cyber-attacks in a multi-user cloud environment [8]. The role-based access approach that is established with the ciphertext policy and attribute-based encryption are subjectable to data leakage with some access permissions [9].

For secure data management in the cloud environment, a novel Enigma Diagonal Encryption Standard Access Control Model (EDES-ACM) is proposed. The model of EDES-ACM is implemented to examine the environmental data in the effective manner and providing a better Unbreakable Secure Data Sharing Environment in Cloud Computing. The proposed EDES algorithm is the improved form of symmetric Advanced Encryption standards with enigmatic diagonal based key generation. The proposed model secures the cloud data effectively by ensuring data security with proper revocation mechanism.

The following contributions are provided to secure the cloud data through the proposed EDES-ACM model:

1) The data security is achieved through very fine-grained access control in Multiuser sharing in a cloud environment.

2) The new mathematical approach for EDES-ACM Cryptographic System is proposed to secure the data.

3) The proposed model is provided with the effective revocation mechanism for revoking the user who performs any unauthorized activities.

4) The proposed cloud set up is protected from the Dictionary Attack, Brute force Attack, and many more.

The present paper is structured in the following order as: Work related to data security in the cloud is discussed in Section 2. The preliminaries for the EDES-ACM frameworks are deliberated in Section 3. The architecture of the EDES-ACM framework with the system and security model in Section 4. Section 5 deliberates the security framework in the EDES-ACM system. Section 6 provides the algorithm and Section 7 details the implementation for the system prototype and its performance and the final section concludes the proposed work with the future proposal.

II. RELATED WORKS

A survey on the performance of some security algorithms in providing security to the cloud is discussed. The work deliberated various clouds that includes the private, community and hybrid cloud. Some research works are selected and related to the encryption mechanism and security of clouds. The comparative studies showed that the performance of AES is better than that of RSA [10]. A chaos-based hybrid AES encryption algorithm was proposed for secure cloud data. An assessment of hybrid algorithms was developed previously revealed that only some portions of the algorithm is employed for chaotic systems. However, in the proposed system many parts of the AES algorithm were involved in chaos structuring [11]. The AES area optimization was accomplished by establishing the mapping among the path hardware of lower data and transformations through the architecture of the iterative loop. The outcomes obtained showed that the proposed architecture of AES is appropriate for its usage in resource constrained systems [12]. An AES algorithm is modified and proposed with an increase in the number of rounds (16) and key size (320 bits) with Polybius square. The main benefit was increased security in high speed data transmission and its drawback was more computational time [13].

An assignment scheme with a hierarchical key was recently proposed [14]. Contrasting with some established schemes that generally encloses a single secret key and few public information to process the decryption process, the proposed system did not share any public information as it is based on access policy chain partitions. A system was developed with a large attribute universe-based access control with the unconstrained flexible attributes of system and space that outsourced cloud decryption. The outcome showed that the developed scheme enhanced the backward secrecy through an effective revocation mechanism but lacked performance in the cryptic mechanism [15]. With decryption outsourcing, a new multi-authority ciphertext policy ABE (MA-CP-ABE) scheme was developed. The implemented outcomes showed that the scheme was efficient, scalable and securable. The scheme was supportable and adaptive for monotone linear secret sharing scheme access policy [16]. A group management system based on lazy-revocation was designed to produce effectual activities of a group that had to endure attacks of collusion when increasing computational costs considerably. The outcome demonstrated an effective achievement towards secure auditing [7].

The combined AES and blowfish algorithm is proposed to secure the cloud with short message service that provides confidentiality, verification and authentication for data. The proposed scheme provided the security for cloud data without a cloud service provider. The main drawback is that the proposed scheme is applicable only for text data [17]. A third-party auditor is employed for enhancing the security of the cloud through implementation of the AES algorithm. The auditor controls the data access of the user whereas the AES encryption technique protects the data during transfer. The AES algorithm in the proposed model ensures the availability of cloud during the storage of huge data in it [18].

From the literature study, it was observed that the issue of fine-grainedness, security and data confidentiality for sharing the cloud data still remained unresolved in multiuser clouds. Another issue is proper method to consider direct attribute / user revocation in data sharing for resource limited users in cloud computing. The frameworks with AES provided a more secure cloud environment than other cryptographic techniques.

III. PRELIMINARIES

A. Bilinear Mapping

Bilinear Mapping is the most common technique employed in generation of security keys and digital signature in the cloud for securing it against any attacks or data infringement [19]. Consider the additive group G_k and a multiplicative cyclic group of Prime Orders, G_p, then the bilinear map that exists between G_k and G_l is given as,

1) Bilinear – c,d $\in Z_q^*$ and a,b $\in G_k$, e (ca,db) = e(a,b)cd for all

2) Non-degenerate-the existence is possible at only one point $(G_k, G_l) \neq 1$

3) Computable- to estimate e (c,d) for any c,d $\in G_k$

B. Digital Signature Generation

For generating the digital signature, the computational Diffie-Hellman technique is employed. Among the many variants, the Inverse Decisional Diffie-Hellman (IDDH) is implemented to generate the digital signature. From the large cyclic group G with the group P (prime order), the following two distribution are involved in signature generation.

1) IDDH triple- $g, g^s, g^{s^{-1}}$, where $s \in Z_q^*$.

2) Random triple- g, g^s, g^t, where s, t $\in Z_q^*$ (s and t are selected randomly form the uniform random).

C. Advanced Encryption Standard

AES does all its computations on bytes rather than bits. It is generally employed with three different key sizes that includes 128 bits, 192 bits and 256 bits with total rounds of 10, 12 and 14 rounds respectively. For encryption it uses byte substitution, shift rows, mix columns and add round key respectively. For decryption it the reverse of the encryption process is performed. The plain text is converted into the cipher-text through the number of round Nr specified for different bit keys. The steps of AES cryptography as in Fig. 1.
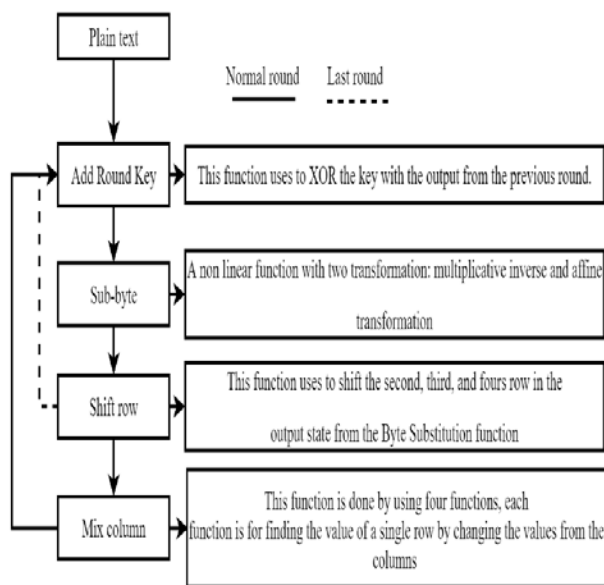
Fig. 1. Steps in AES Cryptology.

IV. PROPOSED EDES-ACM CLOUD ARCHITECTURE

The framework to secure the data in the cloud environment with the Enigma Diagonal Encryption Standard as given in Fig. 2. The proposed approach is established to ensure the security of the cloud through the effective access control mechanism. The control mechanism is established with effective group and data management in the proposed framework. The group management is established with the robust group signature with revocation scheme. The data security is achieved through the EDES algorithm in which the data is encrypted with the symmetrical key before storage into cloud. All the digital group signatures and EDES keys are stored in the key manager and is used to validate the user during the access into the cloud and data.
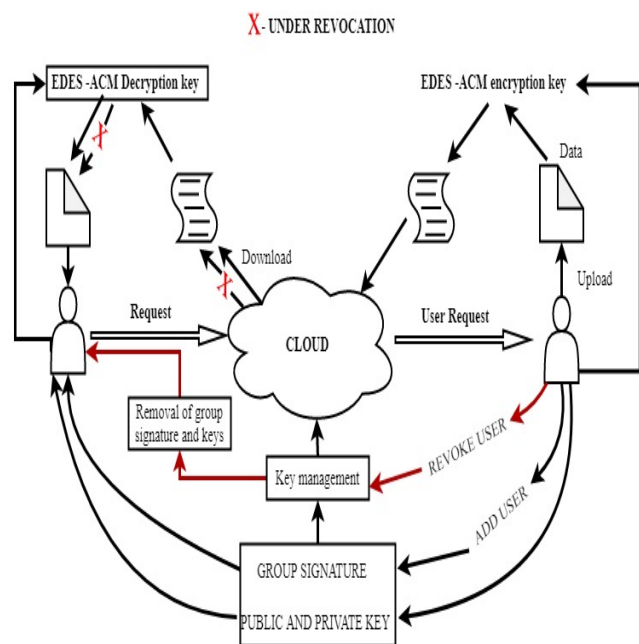

Fig. 2. Proposed EDES-ACM Framework.

Let us consider the user A and B with two different roles process the request to join the cloud environment in the organization. The cloud authority processes the user request and forward it to the data owners for approval. The data owner will generate the two distinct digital group signature and send them to the users through the cloud. Similarly, the generated group signature is stored in the cloud key management. The group owner uploads the data into the cloud and before storing it in the cloud the process of encryption is carried out on the data through the proposed EDES-algorithm. The EDES algorithm is the enhanced version of AES algorithm based symmetric key with 512-byte size that has a total of 10 rounds. After the termination of AES algorithm. The obtained key matrix is subjected to the enigmatic diagonal formulation in which the key is generated with diagonal elements predominantly either in direct order (first to last element) or in reverse order randomly. The final key obtained is shared with the users in the cloud. The users can access the cloud any time, at any place through their distinct digital group signature and private key to access the data.

V. SECURITY MODEL IN PROPOSED EDES-ACM

The main goal of the proposed EDES-ACM framework is to secure the data in the cloud environment in an effective way. Let us again consider the two users A and B who are added to the cloud with different roles. The user's activity is continuously monitored in the cloud environment with the group manager. The group manager reports the activity of the user to the data owner over each session. When user A has involved in any unauthorized activity like tampering the data or any perform any activity that are not defined in the role, the group manager tracks it and reports the owner. The data owner revokes the digital group signature for the user A and the key data management is updated. If user A tries to access the cloud again, the signature is invalid and hence the data is secured in the cloud with improved confidentiality.

In other case, when there is an attack from the user end like DDoS attack, the data in the cloud is under severe threat. In the proposed model, the data is effectively encrypted with the EDES algorithm. Even if the attacker gets the data from the cloud, the decryption is not possible without the EDES key. This results in invalidation of the attacker and the signature through which the access performed is revoked from the key management. It ensures the security of the data over its integrity and privacy preservation.

VI. SYSTEM ALGORITHM

The system algorithm to secure the data in the cloud environment:

- **Setup (1^λ, N)**: The cloud environment is set up is established with the security parameter λ and the maximum receiver capacity of size N. Using the security parameter, the public and the master key are generated in the cloud and are designated as PU and MA respectively.

- **AccUser (IU, IG)**: Every user in the cloud initially process the request for adding them in the cloud. The group manager GM obtains the identity of the user IU

and list it in the group with identity IG based on the role list and its parameter (RL and RLP).

- **GSgen (IU, MA):** Based on the user identity IU, the cloud authority employs the master key MA to generate the digital group signature (DGS) with the validating message V.

- **RM (MA, RUH, S):** Every user in the cloud is provided with some role which follows some hierarchy RUH in the role set S for securing the secrecy of the group through Public parameters set PS.

- **ReUser (IU, IG, DGS):** When the user is tracked for any unauthorized activities, the user identity IU is obtained along with the group identity and digital signature. The RL and RLP of the user are retrieved and digital signature is removed from the key management and it is updated.

- **KeyGen (MA, IU, AP, MK):** The access policy AP of every user is defined in the cloud and it is used along with the MA and IU. The Key matrix MK is obtained through the EDES algorithm with a random diagonal prioritization technique. The final private key PR is provided to the user. Let the final matrix after 10 rounds of AES is given in Fig. 3.

Fig. 3. Enigmatic Diagonal Technique.

The forward or reverse order is selected randomly and its sample keys are given in Table I.

TABLE I. ORDER OF TWISTING

Forward Order	0,5,10,15,1,2,3,4,6,7,8,9, 11,12,13,14	AFKPBCDEGHIJL MNO
Reverse order	15,10,5,0,14,13,12,11,9,8, 7,6,4,3,2,1	PKFAONMLJIHGE DCB

- **EC (M, SU, AP):** The data owner employs the EDES algorithm along with the Access policy AP and the user group set SU to encrypt the data with message M to get EnD.

- **DC (IU, PR, EnD):** Once the user downloads the data with the IU, the PR is provided to decrypt the encrypted data EnD to get the message M.

Algorithm 1: Digital Group Signature and user control

Input: λ,N,IU, IG,
Output: Digital group signature (DGS)
1: Execute the set-up algorithm with λ and N
2: Generate the master key MA, PU
3: When the user processes the request
4:        Get S from Z_q^*
5:        generate the DGS
6: Distribute the DGS to the user
7: Add the user with IU based on IG
8: Track the activities of user
9: When the user performs unauthorized activity
10:        revoke the user DGS
11: user signature is invalid
12: end

Algorithm 2: Key generation and Cryptic mechanism

Input: MA, IU, AP, M, SU, 4x4 matrix, sbox-4x4 matrix,KM, constant matrix – CM, rejandael scheduler RS,
Output : private key. EnD, DnD
1: get A
2: $A \leftarrow Sub\_Bytes(KM, S_{Box})$
3:        $A(i,j) \leftarrow S_{Box}(p,q)$
4: $A \leftarrow Shiftrows(A)$
5:if i=1 or j=1or i=j
6:        A (i,j) ← A (i,j)
7: else If i<j
8:        A (i,j) ←A (i+1,j-1)
9: else if i > j
10:        A(i,j) ←A(i-1,j+1)
11: $A \leftarrow Mix\_Column(A, CM)$
12: for i → 1 to 4
13:        $A(:,i) \leftarrow CM \times A(:,i)$
14: $A \leftarrow Add\_Round\_Key(A, KM)$
15:        $A(:,i) \leftarrow A(:,i) \oplus KM(:,i)$
16: Random (forward, backward)
17: Get the private key PR
18: Using PR encrypt the data (EnD) and upload to the cloud
19: Download the data and decrypt it (DnD)
20: end

VII. RESULTS AND DISCUSSION

The proposed security framework over the data in the cloud is implemented using Java and the data are accumulated in the dropbox. The interfaces in the cloud are achieved through the JAX-WS web services hosted with Apache Tomcat. The SQL database is updated on the server side for storing the data. A Java supported internet browser forms the client- side. The remaining basis configuration is to use the Intel Processor with a memory of 16 GB at 2.45 GHz with a disk of 1TB capacity.

A. Digital Group Signature

Whenever the user processes the request to join the group in the cloud, they are provided with the digital group signature is provided to thee. In the proposed model, the Inverse Decisional Diffie Hellman (IDDH) technique is employed to generate the signature that is provided to the user and stored along with the user identity in the cloud key management. The verification is carried out when the user attempts to access the cloud.so in general there are two parametric that are involved in analyzing the group signature generation i.e. generation time and proofing time as in Fig. 4.

B. Uploading and Encryption

Every data owner in the cloud upload the data into the after encrypting it with the proposed EDES algorithm. Bothe the uploading time and the encryption time increases with the increase in the file size. The time taken for uploading is less compared to that of the encryption time. Hence the cost involved in encrypting the file is given as the **(RL+1) M**. the time taken for uploading and encryption of data through the EDES algorithm is given in Table II and the corresponding plots are given in Fig. 5 and 6.
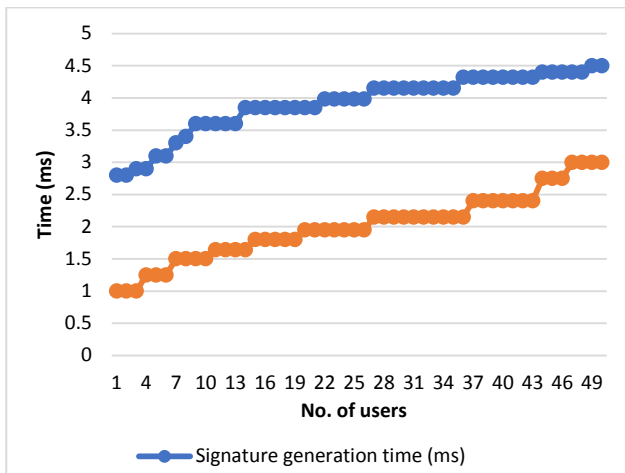
Fig. 4.   Generation Time and Proofing Rime of the IDDH.

TABLE II.   UPLOADING AND ENCRYPTION TIME FOR DIFFERENT FILES

File size (kb)	Upload time (ms)	Encryption time (ms)
10	1	75
20	1.2	80
30	1.3	87
40	1.6	100
50	1.6	115
60	1.95	135
70	2.5	160
80	2.5	190
90	2.5	190
100	3.1	235
500	3.15	280
1024	3.5	335
2048	3.5	380
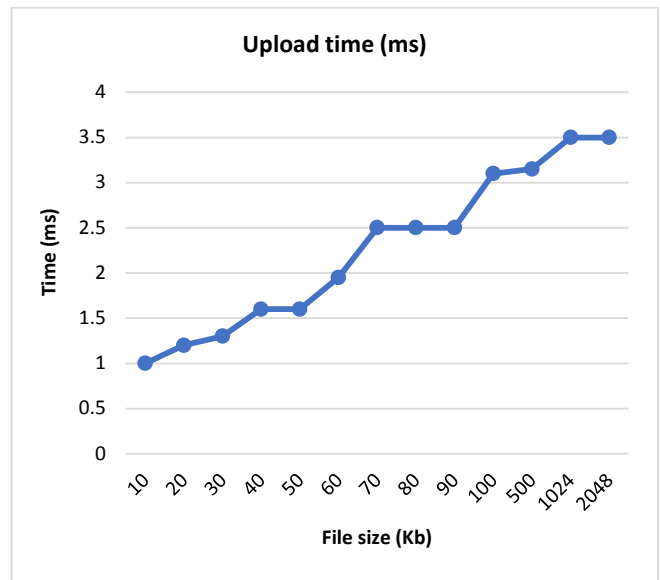
Fig. 5.   Uploading of Data through the EDES Algorithm.
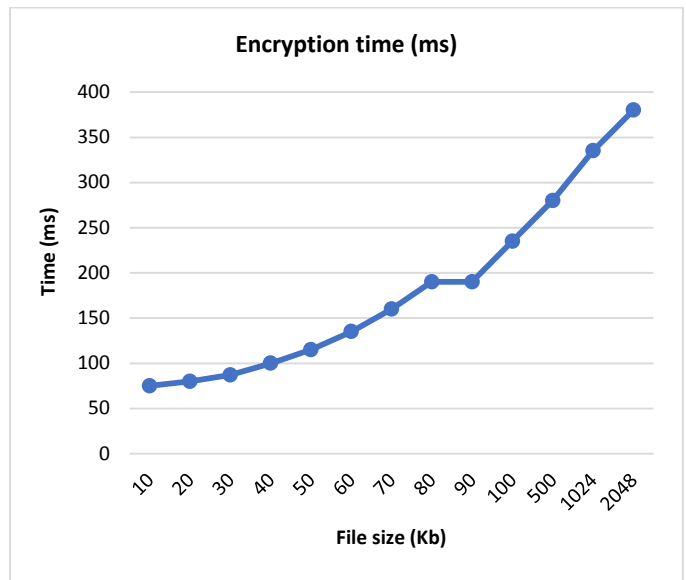
Fig. 6.   Encryption of Data through the EDES Algorithm.

C. Decryption and Download Time

The encrypted data in the cloud is downloaded to the user based on their request. The user employs the decryption key to access the information from the downloaded data. But unlike uploading the downloading time of encrypted data take more time than that involved in decrypting it. The cost involved in decrypting the file $2k(P+E) +3M$ where P and E are the pairing and exponential operators and M is the multiplicative operator. The value of K defines the complexity of the access mechanism to access the data in the cloud. The time taken for downloading and decrypting the file is given in Table III and its corresponding plot are given in Fig. 7 and 8.

The comparison between the existing and the proposed model is given in Table IV.

TABLE III.   DOWNLOADING AND DECRYPTION TIME FOR DIFFERENT FILES

File size (kb)	Decryption time(ms)	Download time (ms)
10	620	950
20	620	950
30	635	950
40	635	985
50	635	985
60	665	985
70	680	1020
80	680	1020
90	705	1065
100	705	1065
500	750	1108
1024	750	1108
2048	800	1139

Fig. 7.   Time for Decrypting the File.

Fig. 8.   Time for Downloading the File.

TABLE IV.   COMPARISON ON COST AND PARAMETER SIZE

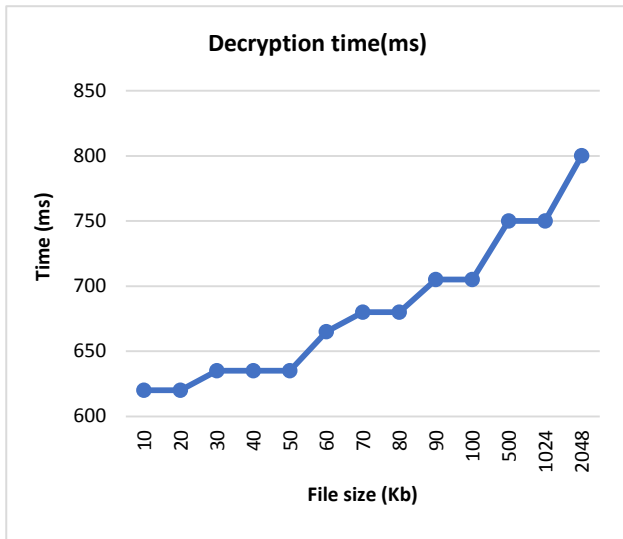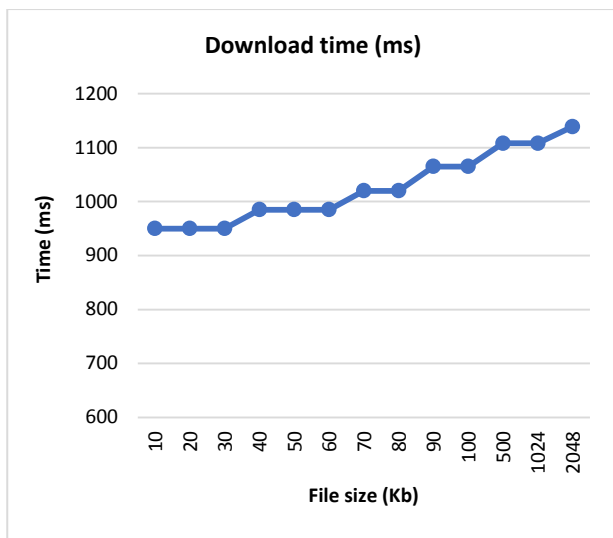| Parameter/ frameworks	HW [20]	ADBS [21]	Proposed EDES-ACM
Encryption cost	$(l_a + 1)$ M	$(l_a + 2)$ M	$(RL+1)$ M
Decryption cost	$3kP + 2kE + 3kM$	$3kP + 2kE + 3kM$	$2k(P + E) + 3M$
Parameter size	$(3k+1)l_G + l_{G_T}$	$(3k+2)l_G + l_{G_T}$	$(3k+3)l_G + l_{G_T}$

Where l_a is the attribute list, LR is the role list, k is the access structure complexity, l_G and l_{G_T} represent the size of element G and G_T. M, P and E are the multiplication, exponential, and pairing operators.

The proposed EDES-ACM framework is effective with reduced cost for both encrypting and decrypting the files than the existing models of HW [20] and ADBS [21]. The size of parameters involved in the study is larger in the proposed EDES-ACM framework than in the existing models. This showed that the proposed framework is effective with reduced computational cost with large parameter size.

VIII. CONCLUSION

The novel framework of EDES-ACM is proposed to secure the cloud environment and its data. The proposed scheme employed the improved version of the AES algorithm to ensure the security of the data. Similarly, the group signature is generated through the IDDH technique and is provided to the user. The key matrix obtained at the end of the AES round is subjected to the enigmatic diagonal technique with either forward or backward approach that are selected randomly for multiusers. Initially, the cloud is organized by the cloud authority with the public key and the master key which are placed in the cloud key management along with the private keys and signatures. The data owner encrypts the data with the EDES encryption key. The data user decrypts with the private key after the validation. The group manager monitors the activity of the user and generates the report on user activities to the data owner. The data owner can revoke the user under any suspicious activity in the cloud environment. The proposed EDES-ACM framework is evaluated on the security metrics that involves the time and cost-based analysis. The results from the analysis showed that the time of signature generation and proofing increase with users and the time for cryptic mechanism increases with file size. The computation cost and parametric size is found to be effective compared with the existing models.

In future the proposed framework can be enhanced with dual encryption techniques with blockchain technology to ensure the security and accumulation of data in effective manner. Aside parallel computation can be performed over the segmented data to reduce the time and cost involved proposed framework. The signature in the proposed framework can involve some secondary operation to make it further more robust.

REFERENCES

[1] Shen, Yaosheng, Ding Wang, and Ping Wang. "Revisiting Anonymous Two-Factor Authentication Schemes for Cloud Computing." In

International Conference on Cloud Computing and Security, Springer, Cham, 2018 pp. 134-146.

[2] Kalaiprasath, R., R. Elankavi, and Dr R. Udayakumar. "Cloud. security and compliance-A semantic approach in end to end security." International Journal Of Mechanical Engineering And Technology (Ijmet) 8, no. 5 2017: 987-994.

[3] Hoeppner, Joseph A. "A Comparison of Cloud Computing Database Security Algorithms." (2015).

[4] Sandhu, Rajinder, and Inderver Chana. "Cloud Computing Standardization Initiatives: State of Play." International Journal of Cloud Computing and Services Science (IJ-CLOSER) 2, no. 5,2013: 351.

[5] Wegberg, Gregor, Hubert Ritzdorf, and Srdjan Capkun. Multi-User Secure Deletion on Agnostic Cloud Storage. ETH Zurich, 2017.

[6] Yu, Yong, Man Ho Au, Giuseppe Ateniese, Xinyi Huang, Willy Susilo, Yuanshun Dai, and Geyong Min. "Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage." IEEE Transactions on Information Forensics and Security 12, no. 4 2016: 767-778.

[7] Tian, Hui, Fulin Nan, Chin-Chen Chang, Yongfeng Huang, Jing Lu, and Yongqian Du. "Privacy-preserving public auditing for secure data storage in fog-to-cloud computing." Journal of Network and Computer Applications 127,2019: 59-69.

[8] Wang, Zhiwei. "Provably secure key-aggregate cryptosystems with auxiliary inputs for data sharing on the cloud." Future Generation Computer Systems 93, 2019: 770-776.

[9] Mao, Xianping, Xuefeng Li, Xiaochuan Wu, Chuansheng Wang, and Junzuo Lai. "Anonymous attribute-based conditional proxy re-encryption." In International Conference on Network and System Security, Springer, Cham, 2018, pp. 95-110.

[10] Kanthale, Akash, and S. P. Potdar. "Survey on Cloud Computing Security Algorithms." vol 5: 2015-2017.

[11] Zeghid, Medien, Mohsen Machhout, Lazhar Khriji, Adel Baganne, and Rached Tourki. "A modified AES based algorithm for image encryption." International Journal of Computer Science and Engineering 1, no. 1 2007: 70-75.

[12] Shaji, Neenu, and P. L. Bonifus. "Design of AES architecture with area and speed tradeoff." Procedia Technology 24 2016: 1135-1140.

[13] Rao, B. Nageswara, D. Tejaswi, K. Amrutha Varshini, K. Phani Shankar, and B. Prasanth. "Design of modified AES algorithm for data security." International Journal For Technological Research In Engineering 4, no. 8, 2017: 1289-1292.

[14] Crampton, Jason, Naomi Farley, Gregory Gutin, Mark Jones, and Bertram Poettering. "Cryptographic enforcement of information flow policies without public information via tree partitions 1." Journal of Computer Security 25, no. 6, 2017: 511-535.

[15] Fu, Xingbing, Xuyun Nie, Ting Wu, and Fagen Li. "Large universe attribute based access control with efficient decryption in cloud storage system." Journal of Systems and Software 135, 2018: 157-164.

[16] Li, Qi, Jianfeng Ma, Rui Li, Ximeng Liu, Jinbo Xiong, and Danwei Chen. "Secure, efficient and revocable multi-authority access control system in cloud storage." Computers & Security 59, 2016: 45-59.

[17] Akhil, K. M., M. Praveen Kumar, and B. R. Pushpa. "Enhanced cloud data security using AES algorithm." In 2017 International Conference on Intelligent Computing and Control (I2C2), IEEE, 2017.pp. 1-5.

[18] PIUS, U.T., ONYEBUCHI, E.C., CHINASA, O.P., and ADOBA, E.F. "A Cloud-Based Data Security System Using Advanced Encryption (AES) and Blowfish Algorithms" Journal of Scientific and Engineering Research, 5 (6) 2018, (pp. 59-66).

[19] Huang, Qinlong, Yixian Yang, and Jingyi Fu. "Secure data group sharing and dissemination with attribute and time conditions in public cloud." IEEE Transactions on Services Computing (2018).

[20] Hohenberger, Susan, and Brent Waters. "Online/offline attribute-based encryption." In International workshop on public key cryptography, Springer, Berlin, Heidelberg, 2014.pp. 293-310.

[21] Li, Jin, Yinghui Zhang, Xiaofeng Chen, and Yang Xiang. "Secure attribute-based data sharing for resource-limited users in cloud computing." Computers & Security 72 2018: 1-12.

Detecting Health-Related Rumors on Twitter using Machine Learning Methods

Faisal Saeed[1], Mohammed Al-Sarem*[3]
Essa Abdullah Hezzam[4]
Department of Information System
College of Computer Science and Engineering,
Taibah University, Medina, Saudi Arabia

Wael M.S. Yafooz[2]
Computer Science Department
College of Computer Science and Engineering,
Taibah University, Medina, Saudi Arabia

Abstract—**Nowadays, the huge usage of internet leads to tremendous information growth as a result of our daily activities that deal with different sources such as news articles, forums, websites, emails and social media. Social media is a rich source of information that deeply affect users by its useful content. However, there are a lot of rumors in these social media platforms which can cause critical consequences to the people's lives, especially if it is related to the health-related information. Several studies focused on automatically detecting rumors from social media by applying machine learning and intelligent methods. However, few studies concerned about health-related rumors in Arabic language. Therefore, this paper is dealing with detecting health-related rumors focusing on cancer treatment information that are spread over social media using Arabic language. In addition, it presents the process of creating a dataset that is called Health-Related Rumors Dataset (HRRD) which will be available and beneficial for further studies in health-related research. Furthermore, an experiment has been conducted to investigate the performance of several machine learning methods to detect the health-related rumors on social media for Arabic language. The experimental results showed the rumors can be detected with an accuracy of 83.50%.**

Keywords—Health-related misinformation; cancer disease; fake information; Twitter; classification formatting

I. INTRODUCTION

Tremendous amount of information are generated as a result of our daily activities and from different sources such as news articles, forum, websites, emails and social media. Therefore, information spread quickly, especially through social media such as Facebook, Twitter, Instagram and others. Hence, social media is a rich source of information that deeply affect users by its contents. This content can be useful for the needs of many users in different areas such education, politics, economics, advertisement, health care, shopping and others. At the same time, there are a lot of information which can be false (rumor) [1][2][3].

Social networks are established in order to connecting people, enhancing relationship and sharing useful information [4]. Recently, it becomes the communication channel for education, advertisement and many other activities. There are a lot of benefits of it in marketing [5] [6], and other professional purposes [7] [8]. In despite of advantageous provided from social networks, the quality of information is low, especially on news and health care information [1] and [2]. Thus, anyone using a social media is able to write self-content as advice or recommendation even without a-prior knowledge and spread such information to many people in minutes [9]. This information could be related to medical treatment and health-related issues [2] [10] [11] [12]. In addition, social media users widely rely on themselves to obtain medical advices from social media. Therefore, the creditability of such information is very important.

Information on social media lacks to quality, credibility and trust-ability as emphasized in health-related misinformation [13] [14] [15]. This misinformation/rumors could have critical consequences to the people's life, especially if it concerns on health information that can lead to health risks [16]. In the existing studies of detecting rumors in health-related information, a little attention has been given to cancer-related information using Arabic language. Therefore, the purpose of this paper is to apply several machine learning methods for detecting health-related rumors aiming cancer treatment over social media using Arabic language. In addition, a dataset for cancer information treatment called Health-Related Rumors Dataset (HRRD) has been created. HRRD has been collected from Twitter, classify by domain experts into true and false information. Then, different preprocessing methods were applied on the dataset such as stemming, tokenization, feature extraction and oversampling. Furthermore, several machine learning methods have been applied and evaluated using different metrics.

This paper organizes as follows: Section II demonstrates related studies. The methods and materials are presented in Section III. While, Section IV explains the results and discussion. Finally, conclusion of this paper is highlighted in Section V.

II. RELATED WORKS

There are few studies focusing on rumors on social media such as [17] [18] [19] [20] [21] [22], while limited studies were conducted to detect rumors about health-related information. Such studies can be classified into focusing on correctness and reliability of medical precipitations [23] [24] [25] effects global health and health literacy [13] [14] [15] [26] [27], and detecting health-related rumors [28] [29] [30].

Soon et al. [23] and Zhang et al. [28] highlighted the issue of health-related rumors by identifying the consequences and benefits of the personal's perceptions through online platform. While, the studies in [13] [14] [15] [26] [27] [31] emphasized

on the importance of checking and verifying online perceptions and credibility of information by health professionals and physicians in domain, otherwise it will be harmful to user's health. In the other way, the authors in [29] [30] [32] presented rumor detection methods by detecting the health-related misinformation using extracting and identifying the fake features. In [30] and [32], Health-related Misinformation Detection framework was developed in order to detect unreliable and reliable health-related information.

Zhang et al. [33] applied logistic regression model for distinguishing between true and false health-related rumors. For this purpose, 453 health rumors from Chinese website were collected and analyzed. The results showed that lengths of rumors headlines, statements and presence of pictures within the context are the most distinctive indicators of false rumors, whereas rumors that contain numbers, hyperlinks and source cues are more likely to be true.

On the other hand, the authors in [34] studied human behavior regarding travel to these areas affected by Zika virus. They have combined content analysis with several machine learning techniques in order to identify first-person reactions and change of travel-related decisions during the Zika outbreak. For this purpose, 29,386 English-language tweets were collected. Only 2000 English-language tweets were annotated and labeled by two annotators and then out of them, 400 tweets were used for training binary logistic regression classifier. The classifier's performance was evaluated using Precision, Recall and F1-score. The best F1-scores were 0.65 for travel change decision, 0.63 for travel consideration and 0.92 for identifying the first-person reactions. In [35], it was dealt with the Zika virus outbreak and gathered around 30 million tweets posted around the world. They incorporated health professionals and crowdsourcing methods to capture and annotated health-related rumors, and used several machine learning techniques including naïve Bayes, random forest and random decision tree to classify the tweets. The data set consists of 3,343 labeled tweets, in which 1786 were rumors. Regarding to the performance of the classifiers, the best achieved results were yielded when random tree was employed with precision of 0.946 and recall 0.944. In [36], they examined questionable health-related information that are posting on Twitter, in particular these tweets related to cancer treatments. For this purpose, they studied 3,212 Twitter users who posted unverified information about cancer treatment. A total of 215,109 tweets about rumor topics were harvested. Then, rigorous filter criteria were applied to exclude irrelevant tweets and users accounts from the data set. At the end, only 4,000 tweets remained, total of which 2,890 were labeled as information about cancer and 1,110 tweets were labeled as non-related to the cancer topic. The logistic regression using n-grams features was employed on this dataset and showed good results.

In addition, the authors in [37] examined 1.5 million tweets mentioning obesity and diabetes epidemics. The main purpose of this study was to assess the quality of information circulating in these conversations, as well as the behavior and information needs of the users engaged in it. The results showed that 41% of the circulated obesity-related tweets and 50% diabetes were posted by non-governmental or academic institution. Furthermore, other studies focused on creating automatically a health misinformation dataset harvested from an online health discussion forum such as [38]. Also, [39] analyzed vaccine rumors in news and social media by developing a dashboard platform that has two networks visualization: the user-as-nodes and tweets-as-nodes. To demonstrate the robustness of the system, a total of 875,088 tweets and 4,020 news articles about vaccine-related topics were collected. It was found that this tool is useful only for tracking the most influenced accounts who post frequently such news or tweets. Similarly, [40] modeled the trustworthiness and reliability of online information using deep learning technique, in particular, convolutional neural network (CNN). The applied model was used to generate a recommendation of trusted medical articles with average veracity score of 78.32%.

III. MATERIALS AND METHODS

The main methodology of this study is illustrated in Fig. 1. It briefly shows the main four phases of conducting this study, which include dataset generation, data preprocessing, applying machine learning methods and evaluation the model.

A. Phase 1: Dataset Generation

The dataset generated for this study is called Health-Related Rumors Dataset (HRRD), which includes a collection of tweets that are related to rumors on cancer disease/treatment. To the best of our knowledge, there is no available dataset for health-related rumors on social media for Arabic language. The phases of generating this dataset includes five steps, which are identifying the keywords, extracting the tweets using Twitter search APIs, extracting the tweets manually, screening the tweets, and labeling the tweets. In this study, the collected tweets are related to cancer symptoms, causes, prevention, treatment and awareness on tweeter that were written in Arabic language. These five steps are described in detail as follows:
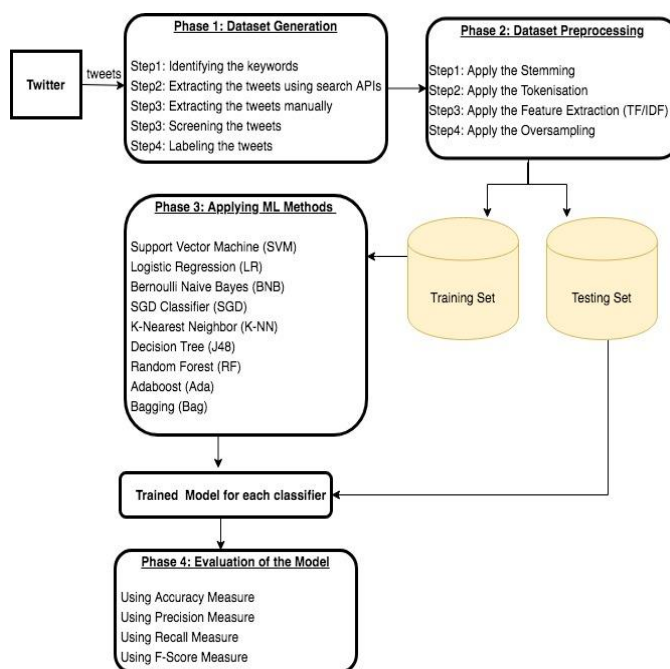

Fig. 1. Methodology of the Study.

Step 1: Identifying the keywords

Several keywords were used in order to automatically or manually extract tweets regarding to cancer disease in Arabic language. These include: cancer disease, cancer causes, cancer treatment, fighting cancer, awareness about cancer, campaign about cancer, warning about cancer, health and cancer, avoid cancer and information about cancer.

Step 2: Extracting the tweets using Twitter search APIs

Using the keywords mentioned in the previous step, 18,684 tweets were automatically extracted from Twitter using Twitter search APIs. These APIs are based on the REST architecture which allow to access Twitter data such as tweets and the user profile information. However, due to the limitation on tweets extraction, the user can only perform a limited number of requests daily. Therefore, additional tweets were extracted manually, as described in the next step.

Step 3: Extracting the tweets manually

In this step, tweets extracted manually due to limitations of extracting tweets using Twitter search APIs which also retrieve huge number of irrelevant tweets, additional tweets were extracted manually using the above keywords to provide more relevant tweets to the dataset. A total of 180 tweets were extracted manually.

Step 4: Screening the tweets

The extracted tweets in step 1 and 2 were screened manually to exclude any irrelevant tweets, which were posted by product sellers, companies, fake/untrusted accounts and others. The total number of tweets was reduced tweets to 175 tweets.

Step 5: Labeling the tweets

In the process of labeling, the extracted tweets were divided into four groups and sent to domain experts (medical doctors) to label the tweets into three options: rumors= yes, no and not sure. The first group were answered by nine experts, while the rest were answered by seven experts only. The majority voting was used to find the final label for each tweet. The results of labeling were (yes: 31, no: 41, 87: not sure, 16: the decision cannot be made because of equal voting). Then, the tweets with labels: "not sure" and "no label" were combined (103) for re-labeling. Also, additional tweets were extracted manually and included to this group (33 tweets). These tweets were divided into three groups and sent to domain experts including oncologists. The first group was answered by seven experts, the second group was answered by four experts and the last group was answered by two experts. The majority voting also was applied here to combine the votes and label the data. In case of any not-sure answers or equal number of votes, more weights were given to the oncologist's answers. The total number of labeled tweets were 208, which include: yes: 128, no: 80. The distribution of the classes for the final dataset is shown in Fig. 2.
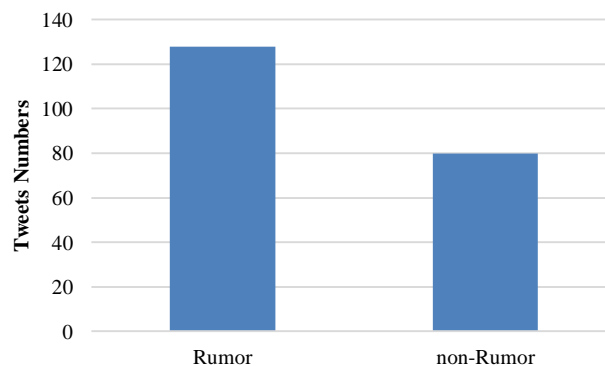

Fig. 2. Distribution of Classes in HRRD Dataset.

B. Phase 2: Data Preprocessing

Python 3.6 with Windows 10 operating system were used for preprocessing the dataset and conduct the experiments. Several libraries were installed including: NLKT for stemming of Arabic texts. In addition, the raw data were tokenized and represented using unigram, bi-gram, trigram, 4-gram and 5-gram. The feature extraction was performed using TF-IDF. The impact of these different preprocessing methods was investigated for detection of health-related rumors in Arabic language. In addition, as shown in Fig. 2, the dataset is unbalanced. Therefore, oversampling method was applied on the minority class in order to provide balanced dataset. The impact of oversampling method also was investigated.

C. Phase 3: Machine Learning Methods

To detect the health-related rumors for Arabic language in social media (Twitter), several machine learning methods were used which include Support Vector Machine (SVM), Logistic Regression (LR), Bernoulli Naive Bayes (BNB), SGD Classifier, K- Nearest Neighbor (K-NN) and Decision Tree (J48). In addition, three ensemble machine methods were used which are Random Forest (RF), AdaBoost (Ada), and Bagging (Bag).

To apply the machine learning methods, the dataset was split into training set (70%) and testing set (30%). Then, several evaluation metrics were applied to measure the performance of detecting the health-related rumors for Arabic language in social media. The details of these measures are described in the next subsection.

D. Evacuation Metric

The evaluation metrics were used for evaluating classification methods that were combined with different preprocessing methods. These includes: Precision, Recall, F-1 score and Accuracy. The definition of these measurements is illustrated as follows:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1\ score = \frac{2*(Precision*Recall)}{Precision+Recall} \tag{3}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

where TP is true positive; TN is true negative; FP is false positive, and FN is false negative.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments have been conducted on two stages *without and with applying oversampling* for the dataset. In each stage, five steps were done by applying different tokenization methods (1-5 gram). The accuracy of each classifier was reported, and the precision, recall and F-score values for rumor class and for the two classes (weighted value) were presented. The best value(s) of each evaluation criterion was highlighted (bold).

Tables I to V shows the performance of the nine machine learning methods for different tokenization methods (1-5 gram)

before applying the oversampling. The results showed that the best accuracy was obtained using SGD classifier (76.19%) for bigram method. For detecting the rumor class, the best precision (0.80) was obtained by Adaboost classifier with 5-gram method, while both BNB and RF obtained the best recall values (1.0) using all tokenization methods. Furthermore, the best F-score obtained for this class was (0.82) by SGD classifier with bigram method. On the other hand, the best weighted precision (for the two classes) was obtained by KNN classifier (0.80) with trigram method, while the best recall and F-score values were obtained using SGD classifier (0.76, 0.75 respectively) with bigram. The experiments reported the good performance of SGD classifier to detect health-related rumors using unbalanced dataset (before oversampling).

TABLE I. THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITHOUT OVERSAMPLING AND USING UNIGRAM)

Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	**74.60%**	**0.76**	0.81	0.78	0.74	**0.75**	**0.74**
LR	**74.60%**	0.75	0.83	**0.79**	0.75	**0.75**	**0.74**
BNB	66.67%	0.63	**1.00**	0.77	**0.79**	0.67	0.60
RF	57.14%	0.57	**1.00**	0.73	0.33	0.57	0.42
SGD	**74.60%**	0.75	0.83	**0.79**	0.75	**0.75**	**0.74**
KNN	61.90%	0.62	0.86	0.72	0.62	0.62	0.58
J48	63.49%	0.57	0.59	0.58	0.63	0.63	0.63
Ada	68.25%	0.70	0.78	0.74	0.68	0.68	0.68
Bag	66.66%	0.67	0.81	0.73	0.66	0.67	0.66

TABLE II. THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITHOUT OVERSAMPLING AND USING BIGRAM)

Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	71.43%	0.72	0.81	0.76	0.71	**0.71**	0.71
LR	71.43%	0.71	0.83	0.77	0.71	**0.71**	0.71
BNB	65.08%	0.62	**1.00**	0.77	0.78	0.65	0.57
RF	57.14%	0.57	**1.00**	0.73	0.33	0.57	0.42
SGD	**76.19%**	0.72	0.94	**0.82**	**0.79**	**0.76**	**0.75**
KNN	71.43%	0.67	0.97	0.80	0.77	0.71	0.68
J48	65.08%	0.65	0.83	0.73	0.65	0.65	0.63
Ada	63.49%	0.71	0.61	0.66	0.65	0.63	0.64
Bag	69.84%	**0.73**	0.75	0.74	0.70	0.70	0.70

TABLE III. THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITHOUT OVERSAMPLING AND USING TRIGRAM)

Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	**73.02%**	**0.72**	0.86	0.78	0.73	0.73	**0.72**
LR	71.43%	0.70	0.89	0.78	0.73	0.71	0.70
BNB	65.08%	0.62	**1.00**	0.77	0.78	0.65	0.57
RF	57.14%	0.57	**1.00**	0.73	0.33	0.57	0.42
SGD	68.25%	0.67	0.89	0.76	0.70	0.68	0.66
KNN	74.60%	0.70	0.97	**0.81**	**0.80**	**0.75**	**0.72**
J48	68.25%	**0.72**	0.72	0.72	0.68	0.68	0.68
Ada	57.14%	0.65	0.56	0.60	0.58	0.57	0.57
Bag	68.25%	0.71	0.75	0.73	0.68	0.68	0.68

TABLE IV.   THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITHOUT OVERSAMPLING AND USING 4-GRAM)

Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	**73.02%**	0.71	0.89	0.79	0.74	**0.73**	**0.72**
LR	71.43%	0.70	0.89	0.78	0.73	0.71	0.70
BNB	65.08%	0.62	**1.00**	0.77	**0.78**	0.65	0.57
RF	57.14%	0.57	**1.00**	0.73	0.33	0.57	0.42
SGD	69.84%	0.67	0.94	0.78	0.74	0.70	0.67
KNN	71.43%	0.67	0.97	**0.80**	0.77	0.71	0.68
J48	58.73%	0.68	0.53	0.59	0.61	0.59	0.59
Ada	71.43%	**0.78**	0.69	0.74	0.72	0.71	**0.72**
Bag	71.43%	0.74	0.78	0.76	0.71	0.71	0.71

TABLE V.   THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITHOUT OVERSAMPLING AND USING 5-GRAM)

Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	**71.43%**	0.70	0.89	0.78	0.73	**0.71**	**0.70**
LR	69.84%	0.68	0.89	0.77	0.71	0.70	0.68
BNB	63.49%	0.61	**1.00**	0.76	**0.78**	0.63	0.54
RF	57.14%	0.57	**1.00**	0.73	0.33	0.57	0.42
SGD	69.84%	0.67	0.94	0.78	0.74	0.70	0.67
KNN	69.84%	0.66	0.97	**0.79**	0.76	0.70	0.66
J48	52.38%	0.62	0.42	0.50	0.55	0.52	0.52
Ada	61.90%	**0.80**	0.44	0.57	0.69	0.62	0.61
Bag	61.90%	0.67	0.67	0.67	0.62	0.62	0.62

In the second stage, oversampling method was applied and the five tokenization methods (1-5 gram) were used. The performance of detecting the health-related rumors using machine learning was consistently improved. Tables VI to X show the performance of the nine machine learning methods with oversampling method. The results showed that the best accuracy was obtained by RF classifier (83.50%) with 4 and 5-gram, and by using SGD classifier (83.50%) using 4-gram method.

For detecting the rumor class, the best precision value (0.83) was obtained by Bag (with unigram) and LR (with 4 and 5-gram), while the best recall and F-score values (1.0 and 0.86, respectively) was obtained by BNB classifier using 3, 4 and 5 - gram).

For the weighted average values, the best precision and recall value was obtained by BNB (0.87, 0.83 respectively) with trigram method. In addition, other classifiers obtained superior performance for recall such as SGD and RF classifiers. The best F-score value was obtained by RF (0.83) using 4 and 5–gram methods.

To compare the accuracy of all machine learning methods using all tokenization methods with and without oversampling, the results were summarized in Fig. 3 to 7. The results showed the consistent enhancements obtained when oversampling was used for all machine learning methods. The best accuracy was obtained by RF using 4 and 5-grams.

TABLE VI.   THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITH OVERSAMPLING AND USING UNIGRAM)

Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	71.84%	0.79	0.63	0.70	0.73	0.72	0.72
LR	73.79%	0.81	0.65	0.72	0.75	0.74	0.74
BNB	79.61%	0.74	**0.94**	**0.83**	**0.82**	0.80	0.79
RF	**80.58%**	0.79	0.85	0.82	0.81	**0.81**	**0.81**
SGD	71.84%	0.80	0.61	0.69	0.74	0.72	0.72
KNN	57.28%	0.63	0.44	0.52	0.59	0.57	0.57
J48	79.61%	**0.84**	0.76	0.80	0.80	0.80	0.80
Ada	76.70%	0.76	0.81	0.79	0.77	0.77	0.77
Bag	**80.58%**	0.83	0.80	0.81	0.81	**0.81**	**0.81**

TABLE VII.   THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITH OVERSAMPLING AND USING BIGRAM)

Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	74.76%	**0.82**	0.67	0.73	0.76	0.75	0.75
LR	75.73%	**0.82**	0.69	0.75	0.77	0.76	0.76
BNB	**81.55%**	0.75	**0.98**	**0.85**	**0.85**	**0.82**	**0.81**
RF	**81.55%**	0.77	0.93	0.84	0.83	**0.82**	**0.81**
SGD	78.64%	0.79	0.81	0.80	0.79	0.79	0.79
KNN	64.08%	0.76	0.46	0.57	0.68	0.64	0.63
J48	76.70%	0.77	0.80	0.78	0.77	0.77	0.77
Ada	78.64%	0.78	0.83	0.80	0.79	0.79	0.79
Bag	78.64%	0.78	0.83	0.80	0.79	0.79	0.79

TABLE VIII.   THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITH OVERSAMPLING AND USING TRIGRAM)

Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	74.76%	**0.82**	0.67	0.73	0.76	0.75	0.75
LR	75.73%	**0.82**	0.69	0.75	0.77	0.76	0.76
BNB	**82.52%**	0.75	**1.00**	**0.86**	**0.87**	**0.83**	**0.82**
RF	80.58%	0.75	0.94	0.84	0.83	0.81	0.80
SGD	**82.52%**	0.80	0.89	0.84	0.83	**0.83**	**0.82**
KNN	62.14%	0.73	0.44	0.55	0.65	0.62	0.61
J48	80.58%	0.76	0.93	0.83	0.82	0.81	0.80
Ada	74.76%	0.79	0.70	0.75	0.75	0.75	0.75
Bag	76.70%	0.78	0.78	0.78	0.77	0.77	0.77

TABLE IX.   THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITH OVERSAMPLING AND USING 4-GRAM)

Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	75.73%	0.82	0.69	0.75	0.77	0.76	0.76
LR	76.70%	**0.83**	0.70	0.76	0.78	0.77	0.77
BNB	81.55%	0.74	**1.00**	0.85	**0.86**	0.82	0.81
RF	**83.50%**	0.77	0.98	**0.86**	**0.86**	**0.83**	**0.83**
SGD	**83.50%**	0.79	0.93	0.85	0.84	**0.83**	**0.83**
KNN	63.11%	0.74	0.46	0.57	0.66	0.63	0.62
J48	67.96%	0.72	0.63	0.67	0.69	0.68	0.68
Ada	79.61%	0.78	0.85	0.81	0.80	0.80	0.80
Bag	73.79%	0.76	0.72	0.74	0.74	0.74	0.74

TABLE X.   THE PERFORMANCE OF MACHINE LEARNING METHODS FOR DETECTING HEALTH-RELATED RUMORS (WITH OVERSAMPLING AND USING 5-GRAM)

| Classifier	Acc.	For Rumor Class			For Two classes (Weighted Avg.)		
		Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- Score*
SVM	75.73%	0.82	0.69	0.75	0.77	0.76	0.76
LR	77.67%	**0.83**	0.72	0.77	0.78	0.78	0.78
BNB	81.55%	0.74	**1.00**	0.85	**0.86**	0.82	0.81
RF	**83.50%**	0.77	0.98	**0.86**	**0.86**	**0.83**	**0.83**
SGD	79.61%	0.79	0.83	0.81	0.80	0.80	0.80
KNN	63.11%	0.74	0.46	0.57	0.66	0.63	0.62
J48	78.64%	0.76	0.87	0.81	0.79	0.79	0.78
Ada	77.67%	0.76	0.83	0.80	0.78	0.78	0.78
Bag	74.76%	0.73	0.81	0.77	0.75	0.75	0.75
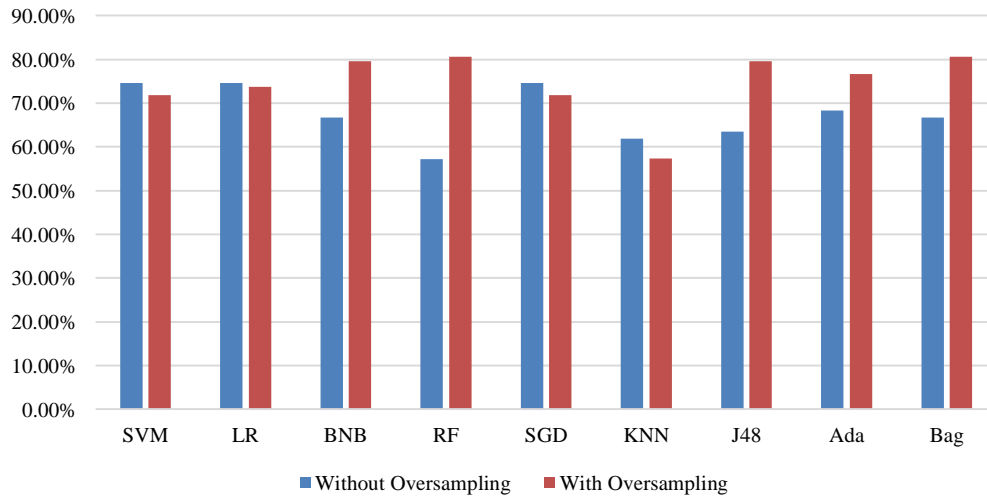
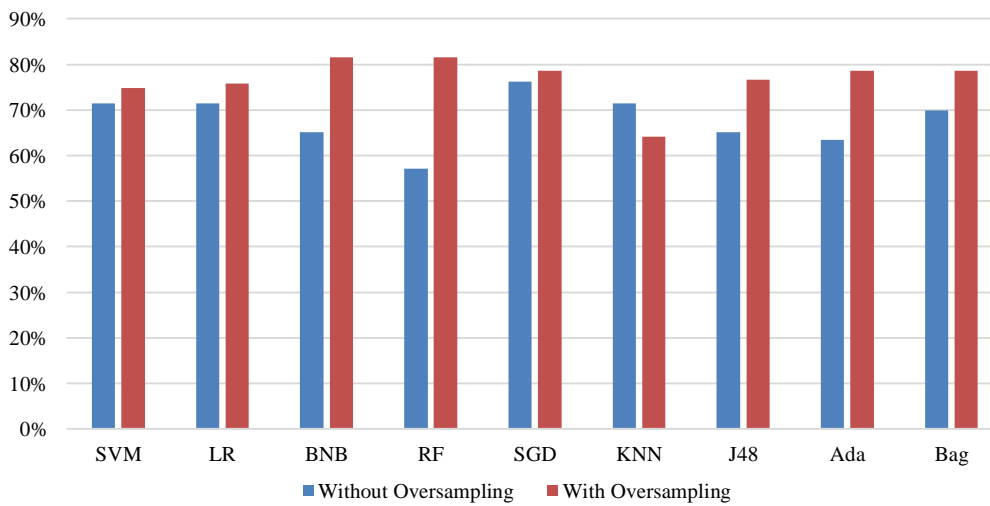Fig. 3.   Accuracy of Classifiers with Unigram.

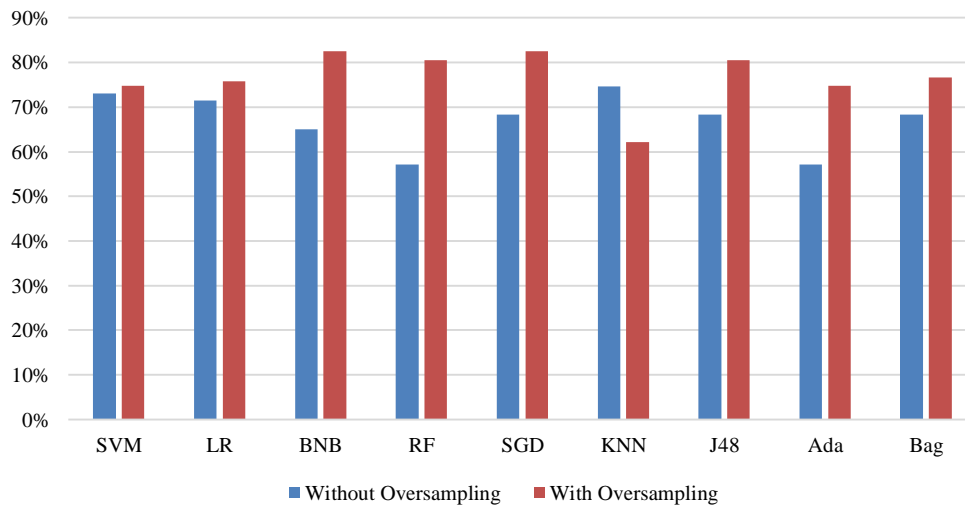Fig. 4.   Accuracy of Classifiers with Bigram.

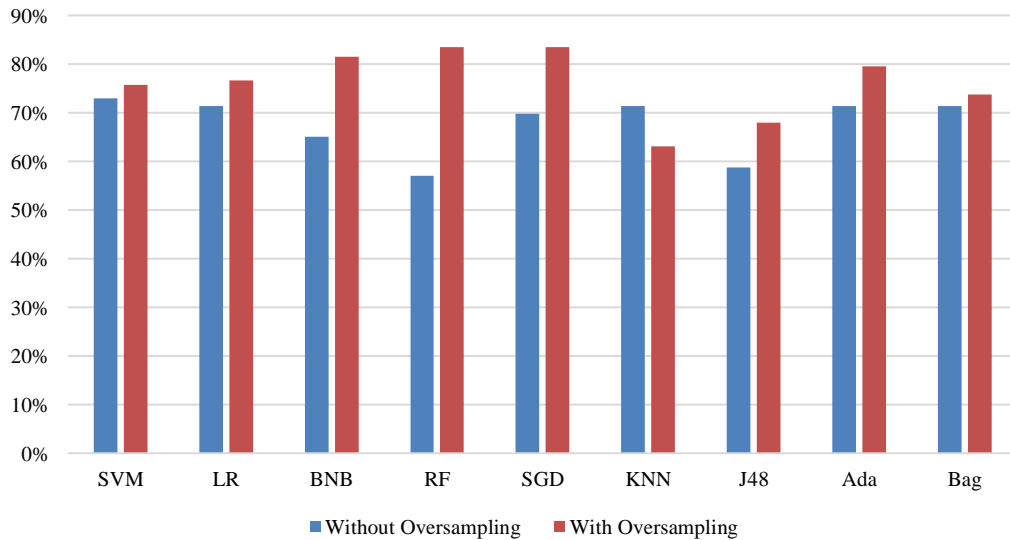Fig. 5.   Accuracy of Classifiers with Trigram.

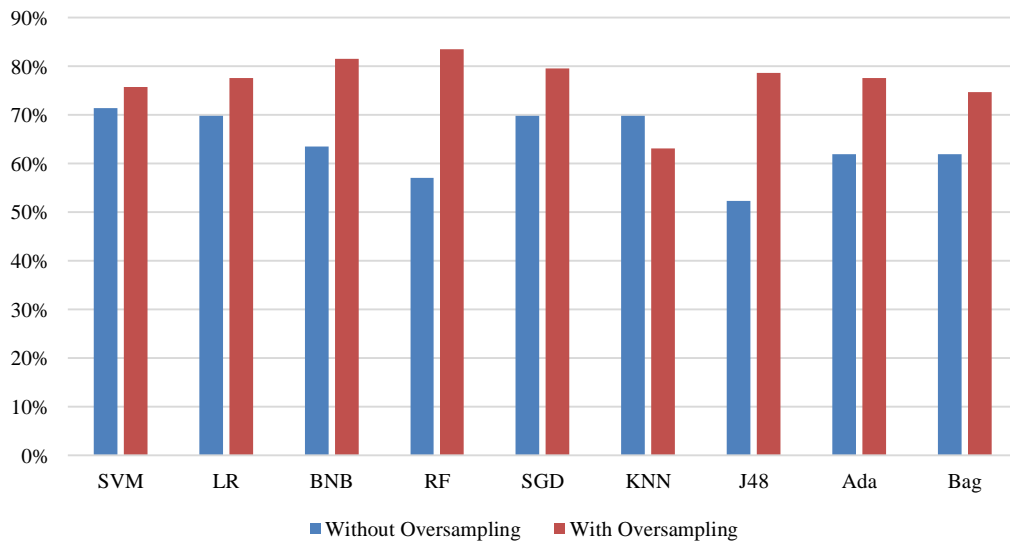Fig. 6. Accuracy of Classifiers with 4-gram.

Fig. 7. Accuracy of Classifiers with 5-gram.

V. CONCLUSIONS AND FUTURE WORKS

This study investigated the performance of several machine learning methods to detect the health-related rumors in social media for Arabic language. The dataset (HRRD) was generated by extracting tweets regarding cancer disease from Twitter using Arabic language. The experiments were conducted by applying several preprocessing methods such as stemming, tokenization and oversampling. Then, several machine learning methods were applied. The experimental results showed that when the data is balanced (using oversampling method), the performance of machine learning methods clearly improved. The best accuracy was obtained by random forest classification (83.50%) using 4 and 5 gram as tokenization methods. Therefore, this study recommends using random forest to detect the health-related rumors in social media written in Arabic language. This study opens the door for other

researchers to work on health-related rumors in Arabic and also provide the HRRD dataset available that can be also beneficial for further studies in health-related research. In future work, other machine learning methods can be applied with different preprocessing methods. In addition, the dataset can be enriched by including more tweets on cancer disease from social media.

REFERENCES

[1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

[2] Zhao, Y., & Zhang, J. (2017). Consumer health information seeking in social media: a literature review. Health Information & Libraries Journal, 34(4), 268-283.

[3] Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. arXiv preprint arXiv:1804.08559.

[4] Kardos, P., Leidner, B., Pléh, C., Soltész, P., & Unoka, Z. (2017). Empathic people have more friends: Empathic abilities predict social

network size and position in social network predicts empathic efforts. Social Networks, 50, 1-5.

[5] Martínez-López, F. J., Anaya-Sánchez, R., Molinillo, S., Aguilar-Illescas, R., & Esteban-Millat, I. (2017). Consumer engagement in an online brand community. Electronic Commerce Research and Applications, 23, 24-37.

[6] Chiang, I. P., Wong, R., & Huang, C. H. (2019). Exploring the Benefits of Social Media Marketing for Brands and Communities. " International Journal of Electronic Commerce Studies", 10(2), 113-140.

[7] Utz, S., & Breuer, J. (2016). Informational benefits from social media use for professional purposes: Results from a longitudinal study. Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 10(4).

[8] Nisar, T. M., Prabhakar, G., & Strakova, L. (2019). Social media information benefits, knowledge management and smart organizations. Journal of Business Research, 94, 264-272.

[9] Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., ... & Sandhu, M. (2019). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Political Behavior, 1-23.

[10] Viviani, M., & Pasi, G. (2017). Credibility in social media: opinions, news, and health information—a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(5), e1209.

[11] Sharma, M., Yadav, K., Yadav, N., & Ferdinand, K. C. (2017). Zika virus pandemic—analysis of Facebook as a social media health information platform. American journal of infection control, 45(3), 301-302.

[12] Alhayan, F., Pennington, D. R., & Ayouni, S. (2018, April). Measuring passive engagement with health information on social media. In 2018 21st Saudi Computer Society National Computer Conference (NCC) (pp. 1-6). IEEE.

[13] Yang, M. (2019, July). Health information literacy of the older adults and their intention to share health rumors: an analysis from the perspective of socioemotional selectivity theory. In International Conference on Human-Computer Interaction (pp. 97-108). Springer, Cham.

[14] Gu, R., & Hong, Y. K. (2019). Addressing Health Misinformation Dissemination on Mobile Social Media.

[15] Trethewey, S. P. (2020). Strategies to combat medical misinformation on social media.

[16] Caulfield, T., Marcon, A. R., Murdoch, B., Brown, J. M., Perrault, S. T., Jarry, J., ... & Rachul, C. (2019). Health misinformation and the power of narrative messaging in the public sphere. Canadian Journal of Bioethics/Revue canadienne de bioéthique, 2(2), 52-60.

[17] Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K. F. (2015, October). Detect rumors using time series of social context information on microblogging websites. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 1751-1754).

[18] Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. British journal of political science, 47(2), 241-262.

[19] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, ''Rumor detection with hierarchical social attention network,'' in Proc. 27th ACM Int. Conf. Inf. Knowl. Manage., 2018, pp. 943–951. doi: 10.1145/3269206.3271709.

[20] Thakur, H. K., Gupta, A., Bhardwaj, A., & Verma, D. (2018). Rumor detection on twitter using a supervised machine learning framework. International Journal of Information Retrieval Research (IJIRR), 8(3), 1-13.

[21] Al-Sarem, M., Boulila, W., Al-Harby, M., Qadir, J., & Alsaeedi, A. (2019). Deep Learning-Based Rumor Detection on Microblogging Platforms: A Systematic Review. IEEE Access, 7, 152788-152812.

[22] Alkhodair, S. A., Ding, S. H., Fung, B. C., & Liu, J. (2020). Detecting breaking news rumors of emerging topics in social media. Information Processing & Management, 57(2), 102018.

[23] Soon, J. J. Q., Banerjee, S., & Chua, A. Y. K. (2017). Analyzing medical personnel's perceptions of online health rumors.

[24] Chua, A. Y., & Banerjee, S. (2015, December). Analyzing users' trust for online health rumors. In International Conference on Asian Digital Libraries (pp. 33-38). Springer, Cham.

[25] Chua, A. Y., & Banerjee, S. (2018). Intentions to trust and share online health rumors: An experiment with medical professionals. Computers in Human Behavior, 87, 1-9.

[26] Zhou, J., Liu, F., & Zhou, H. (2018). Understanding health food messages on Twitter for health literacy promotion. Perspectives in public health, 138(3), 173-179.

[27] Kasemsap, K. (2017). Analyzing the role of health information technology in global health care. In Handbook of research on healthcare administration and management (pp. 287-307). IGI Global.

[28] Zhang, Z., Zhang, Z., & Li, H. (2015). Predictors of the authenticity of Internet health rumours. Health Information & Libraries Journal, 32(3), 195-205.

[29] Sicilia, R., Giudice, S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2018). Twitter rumour detection in the health domain. Expert Systems with Applications, 110, 33-40.

[30] Liu, Y., Yu, K., Wu, X., Qing, L., & Peng, Y. (2019). Analysis and Detection of Health-Related Misinformation on Chinese Social Media. IEEE Access, 7, 154480-154489.

[31] Armstrong, P. W., & Naylor, C. D. (2019). Counteracting health misinformation: a role for medical journals?. Jama, 321(19), 1863-1864.

[32] Li, Y., Zhang, X., & Wang, S. (2017). Fake vs. real health information in social media in China. Proceedings of the Association for Information Science and Technology, 54(1), 742-743.

[33] Zhang, Z., Zhang, Z., & Li, H. (2015). Predictors of the authenticity of Internet health rumours. Health Information & Libraries Journal, 32(3), 195-205.

[34] Daughton & Paul, 2019 Daughton, A. R., & Paul, M. J. (2019). Identifying Protective Health Behaviors on Twitter: Observational Study of Travel Advisories and Zika Virus. Journal of medical Internet research, 21(5), e13090.

[35] Ghenai & Mejova, Ghenai, A., & Mejova, Y. (2017). Catching Zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on Twitter. arXiv preprint arXiv:1707.03778.

[36] Ghenai & Mejova, Ghenai, A., & Mejova, Y. (2018). Fake cures: user-centric modeling of health misinformation in social media. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 58.

[37] Mejova, 2018 Mejova, Y. (2018, April). Information Sources and Needs in the Obesity and Diabetes Twitter Discourse. In Proceedings of the 2018 International Conference on Digital Health (pp. 21-29). ACM.

[38] Alexander Kinsora, Kate Barron, Qiaozhu Mei, and VG Vinod Vydiswaran. 2017. Creating a Labeled Dataset for Medical Misinformation in Health Forums. In Healthcare Informatics (ICHI), 2017 IEEE International Conference on. IEEE, 456–461.

[39] Patty Kostkova, Vino Mano, Heidi J Larson, and William S Schulz. 2016. Vac medi+ board: Analysing vaccine rumours in news and social media. In Proceedings of the 6th International Conference on Digital Health Conference. ACM, 163–164.

[40] Samuel, H., & Zaïane, O. (2018). MedFact: Towards Improving Veracity of Medical Information in Social Media Using Applied Machine Learning. Lecture Notes in Computer Science, 108–120. doi:10.1007/978-3-319-89656-4_9.

A Novel Low Power, Minimal Dead Zone Digital PFD for Biomedical Applications

Sudhakiran Gunda[1]

Research Scholar, Department of Electronics and
Communication Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, AP, India

Ernest Ravindran R. S[2]

Assistant Professor, Department of Electronics and
Communication Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, AP, India

Abstract—Chronic diseases and rising aging populations are the major reasons towards the usage of low power, low noise, life time performance Biomedical Implantable Devices. Efficient architectural designs will be responsible for the requirements set out above. This paper focuses on the ADPLL DPFD architecture for implantable biomedical devices. For high performance DPFD, the dead zone, lock in time is a seldom limitation to ADPLLs. In the present paper, a new approach to design a dead zone free with fast and high locking time and low phase noise DPFD is considered to be a challenge. This can be accomplished by carefully controlling the reference and feedback clock frequencies of the phase detector with the proposed NIKSTRO/SURAV latch based sense amplifier. The proposed architecture was developed and simulated using 45nm technology and it is observed that it provides a 20ns dead zone with 4.8mW of power consumption at the rate of 1.8GHz, while the lock in time for the proposed method is 340ns with moderate phase noise. It is also noted that the designed one showed better results when compared to the existing ones.

Keywords—Biomedical Implantable Device (BIMD); Digital Phase Frequency Detector (DPFD); Digital Controlled Oscillator (DCO); Sense Amplifier Based Flip-flop (SAFF); NIKSTRO or SURAV

I. INTRODUCTION

Archeological research reveals that the Greek civilization used instruments to study the human body in order to understand human anatomy and to treat healthy and pathological conditions. This idea has placed roots for the growth of a biomedical tree. In addition to this, the technical advancements throughout medical sciences have always played an important role by making remarkable advances in health care resulting in emerging a field called biomedical engineering. The new science and technology of biomedical engineering have contributed to the manufacture of cutting-edge biomedical implantable over the last five decades, helping to improve clinician's know-how to improve the human anatomy [1]. A more precise diagnosis, which can be achieved by highly technical biomedical devices / BIMD's, is necessary for medical professionals to prescribe an effective cure. These BIMDs range from sensors, GES and cardiac pacemakers, ICD, to DBS, nerve (PNS, SCS), and bone stimulators.

While a variety of biomedical implants exist for many applications, each IMD consists mainly of an electronic system and battery [2]. Because of the IMD area and size limits, a Chip Specific System (AS-SoC) system is currently covering main portions of the IMDs. The main functionality of these devices is to monitor and analyze body physiological signals, to deliver the drugs needed precisely if necessary, to resurrect the malfunctioning organ or body part, for transmission of the diagnostic data, to receive the external commands, to stimulate the body's organ while it is not functioning properly, thus transceiver is the most important component in BIMDs. Conventional devices have been used for short-range magnetic IMDs that are easily affected by EM way interference resulting in transmission imprecision [3]. In order to provide the safety measures, whole ball of wax the medical applications should be carried out at Medical Implant Communication Services (MICS) ranging from 401 MHz to 406 MHz (intra range is 402 MHz to 405 MHz). The key building block in BIMDs is PLLs, but the conventional analog PLL's need a wider silicon area to accommodate LC oscillators, charge pump and RC LPFs and therefore not easily portable to other technology nodes. To overcome the analog PLL drawbacks, All Digital PLLs (ADPLLs) have been proposed. For detailed information on, How ADPLLs subtle the PLLs & Digital PLLs (DPLLs)? block diagram of PLL, categories of DPLLs, reader has suggested to read [3-8].

II. ALL DIGITAL PHASE LOCKED LOOP

The advancements in CMOS technology scaled down the supply voltages $\leq$ 1 V, making the traditional analog PLL design for designers in current deep-submicron CMOS processes very challenging. Nevertheless, short channel CMOS process has preferred digital circuits and is therefore highly focused on digital circuits today. All these distinct factors lead to undergo a change in velocity of the growth of ADPLLs in which all the sub-blocks of the conventional analog PLL were replaced by their comparable/equivalent digital blocks. The general ADPLL block diagram is shown in Fig. 1 consisting of Digital Phase Frequency Detector (DPFD), loop filter and Digital Controlled Oscillator (DCO).
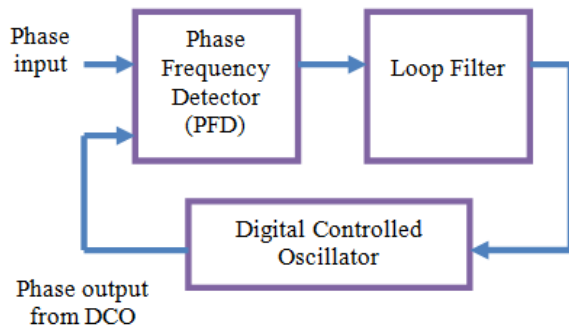
Fig. 1.   General Block Diagram of ADPLL.

As all ADPLL blocks are digital, a converter is needed for designers' final analysis. An alternative approach in the current paper is also listed like Time to Digital Conversion (TDC) based ADPLL, which translates time intervals to a digital value to reduce the number of converters in ADPLL rather than using DPFD [4]. ADPLL's block diagram can now be modified slightly by using TDC in Fig. 2. Therefore the alternative ADPLL contains mainly the Time to Digital Converter (TDC), phase error detector (PED), the digital loop filter (DLF) and the computer controlled oscillator (DCO) with additional circuitry based on the requested application.
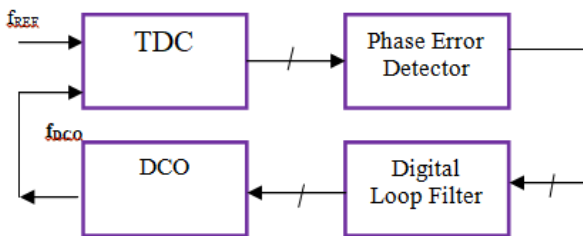
Fig. 2.   Block Diagram of ADPLL with TDC.

III. Literature Survey

An alternative solution is important to overcome the design challenges of digitally driven deep-submicrometer CMOS processes. One of the solutions to dealing with the above listed problem is the time mode signal processing (TMSP). It operates and processes analog signal information through digital blocks, in which voltage (V) and current (I) variables are replaced with the appropriate time differences between two rising edges as the time variables, the large and power-hungry analog blocks can be substituted with logical circuits. In addition, digital elements as an analog circuit replacement, allows the use of digital synthesis and testing techniques. This represents significant step forward to bridge the gap between analog and digital design, which for the analog design community is extremely important. There are two types of converters known as DTC and TDC to offer the requirements mentioned. A TDC converts the time interval between the two clock edges to a digital number. Whereas the DTC is quite opposite to the TDC. For the detailed principle of operation of TDC [5, 6]. In ADPLL, the frequency comparator known as the Phase Frequency Detector (PFD) is replaced by the TDC. Various types of TDCs are available; some of them are taken into account and shown as appropriate in this paper.

Buffer delay line TDC (BDLTDC) [6], which can transmit signal via chain of buffers and flip-flops to output buffers. Once the reset signal is activated the delay line is sampled by flip-flops. Complete implementation is in digital, so it is very simple and easy to operate, but having lower resolution since it has the delay of one buffer cell.

Inverter delay line TDC (IDLTDC) [6], the signal transmission can take place through a chain of inverters and flip flops which can increase the resolution twice in comparison with BDLTDC but is therefore limited in applications by the advancements in CMOS technology.

Time to Amplitude Converter TDC (TACTDC) [7], is the combination of TAC and ADC (Analog to Digital Converter), the entire operation depends on the charging mechanism of a capacitor on the output section of the TAC. The capacitor value is then converted into a digital value with the help of ADC. Very simple to implement is the major advantage for ASICs, but they have a large dead time and the limited resolution for ADC operation, a high power dissipation for 50 ps resolution [7]. Because of its complete analogy in its structure designing, it has no CMOS implementation.

Vernier Delay Line TDC (VDLTDC) [6, 7] can measure the delay line with sub gate resolution mechanism. This architecture delays both START and STOP signals with uneven time periods (delay of STOP signal is slightly greater than the START signal) which is the START signal that leads the STOP signal. The operation explains that the resolution depends on the two delay lines rather than the delay element compared to BDLTDC, IDLTDC and TACTDC. VDLTDC provides not only high resolution, but also high power consumption and silicon area since it is composed of 2 buffers and one flip flop in its designing.

According to TDC's report, it drastically reduced the number of ADPLL converters required, but the remaining technical limitations were severely affected [3-7]. These results make TDC based ADPLLs were not suitable for bio-medical applications, but ADPLLs are the major functional unit in BIMDs, and there is a need for significant improvement in this area. Another major requirement for these types of devices is not only provision of ultra-low power consumption to safeguard a longer battery life, but also the provision of better circuit topologies for smaller silicon-area. One of the best ways to achieve ultra-low power consumption is to assertively scale down the supply V_{DD}, because the transition power dissipation of digital architectures has squared dependence on V_{DD} with constant capacitance and frequency for the same technology (CV^2F). Conversely, this destructive reduction in V_{DD} leads in low driving capability and reduced circuit speed due to threshold voltages of MOS devices. To overcome this, the working operation of MOS devices should be very close to the week/middle inversion region [9, 10]. Nonetheless, the mixed signal processing circuits functioning in these regions still shows the signs of low driving and switching speeds. In order to acquire low power consumption, less silicon-area and portability which are extremely required for the BIMDs, the use of TDC based ADPLLs should be bypassed.

The PFD [8] consists of two D flip - flops, a NAND gate and a reset path with a fixed delay element. The design process requires a large area of silicon and affects the dead zone due to the fixed delay cells.

A further signal called "dir" is given in PFD [10], which shows the leading and lagging information for clock phases. Additional circuits compared to the design shown in [8] are needed for the construction of this kind of PFD. D flip flops with Strollo's latches, which produce glitches at the PFD output. The overall architecture is complex and high power consumption. PFD-based ADPLL is therefore the best solution for biomedical applications, since it provides good qualitative and quantitative measurements for the required operation.

A new ADPLL consisting mainly of the digital phase Frequency detector (DPFD) rather than the TDC is introduced and discussed in this paper. A new approach to building the DPFD with SAFFs was considered. This paper is organized as, Section 1 deals with the introduction and the genesis of BIMDs, Section 2 presents the glance about the devices used in ADPLL with their pros and cons, following the literature survey of TDCs & PFDs in Section 3. Section 4 deals with SAFF and its operation with different latches, Section 5 provides information on the evolution and operation of proposed latch, Section 6 provides a description of the design of DPFD with the proposed SAFF, finally a comparative conclusion as set out in Section 7.

IV. SENSE AMPLIFIER BASED FLIP FLOP (SAFF)

Daily use of electronic devices in various applications in different fields has led researchers to work on new ideas. SAFF is one of the proofs of the requirements set out above, since it has been used in digital systems where high performance, high reliability, high speed and low energy consumption are the primary criteria. The limitations of timing elements, storage elements such as (latches and flip-flops) and clock loading make SAFF the key choice for most applications in the digital era. The above parameters are partly or entirely depending on the clocking mechanism. Due to the fact that clock distribution network and the impact of the clock skew, setup time, hold time and their relationship in the design consumes 20 to 45% of the total power consumption of the digital system [11], imposed on new latches and flip flop designs.

A. SAFF Mechanism

The standard flip-flops consist of pulse generators followed by a latch as shown in Fig. 3: the schematic representation appears to be the same as the master-slave flip-flop, but varies in its operation.

From Fig. 3, the pulse generator sets the operation of the slave latch by generating pulses of sufficient duration on the basis of input data and clock signal considerations. Depending on the specific time of increase of the clock signal and the time of dropping, the pulse generator stage is more sensitive to the edges than to the depths. This may give rise to an ambiguous situation for the operation of a flip flop under certain conditions, mainly in terms of reliability and robustness of operation. Thus, some design methodologies such as IBM's LSSD [11] struggled to allow use of flip flops.
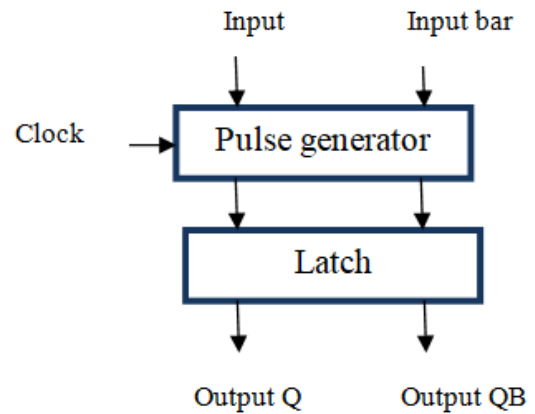

Fig. 3. Traditional Flip Flop Block Diagram [11].

SAFF also has master and slave blocks identical to flip flop, except the master block is a sense amplifier, and the slave block is a key, as seen in Fig. 4: Sense amplifier block is the name given because it is sensed with a clock signal and additional differential inputs and produces complemented inputs to the slave latch. Slave latch is designed as an SR (Set-Reset) latch that has been ignited by either SB or RB (but not both) created by the master block [11-16].
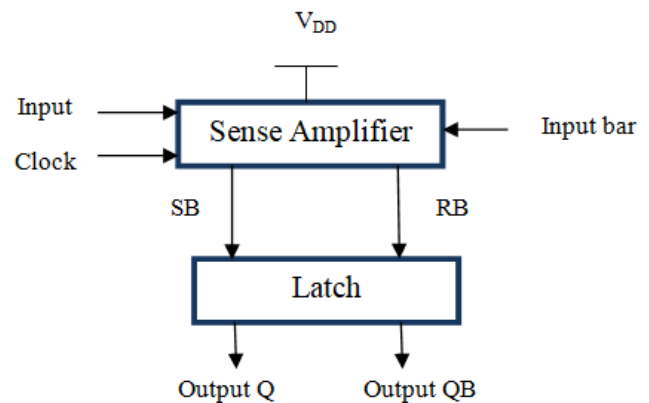

Fig. 4. General Structure of SAFF.

The entire design of SAFF portrays the flip flop process as the Sense amplifier block produces uniform transitions from logic zero to logic one on any output based on the leading edge of the clock distribution. During the clock operation, the output is not affected by any successive changes to the input data. The output state generated can now be processed and the slave latch can be held until the next leading edge of the clock signal arrives. Once the clock is inactive, both outputs of the master block assume high logic [11-16].

B. Operation of Sense Amplifier

The operation of the Sense Amplifier stage is mainly controlled by the clock signal as follows:

When the clock is in an active low state: then the PMOS transistors PM1 and PM4 are ON, the output nodes SB and RB are precharged to logic high (H). Both states make NMOS transistors NM1 and NM2 ON and charge their sources to VDD-V_{thNMOS}, since there is no clear path to the ground due to the clocked NMOS transistor NM5 being OFF. The common NA node (NM3, NM4 and NM5) will also precharge either

NM3 or NM4 to VDD-V_{thNMOS}. Thus, all the MOS capacitances in the differential tree were precharged before the clock signal enters the logical high level.

When the clock is in the active high state (H): the NMOS transistor NM5 is completely ON-conditioned, creating a direct path to SB and RB towards the ground, conditioned on the status of the NM3 or NM4 transistors. The detailed SB is discharged by either NM1, NM3 & NM5 when D is logic H by turning NM2 OFF and PM3 ON or NM1, NM6, NM4 & NM5 when DB is logic L which is the longest delay path for discharging (not preferable for high-speed applications). The smallest path for RB is through NM2, NM4 and NM5 when DB is logic H by turning NM1 OFF and PM1 ON. Further alterations of input data should not affect the status of SB and RB after these initial changes [11-16]. The complete operation of sense amplifier can be described in Table I.

Another important observation in Fig. 5 is that the inputs are decoupled from the outputs of the sense amplifier stage forming the basis for the flip flop operation of the circuit [12]. All transistors were designed using 45 nm technology, considering that all PMOS (PM1 to PM4) had a W / L ratio of 4 units and NMOS (NM1 to NM6) had 2 units. The resulting increase time and fall time in the simulated waveforms is approximately 40ns and therefore the total propagation delay is almost 20ns. In order to obtain a DB signal, the data D signal is inverted with a delay element having the same W / L ratio for PMOS and NMOS.

At the high clock pulse, the output of the sense amplifier forces to low, and this will be floating low if the data changes during the high clock pulses. Although the data is modified, the NM6 transistor provides a path to the ground that prevents the possible charging of the low output of the sense amplifier stage due to leakage currents. These leakage currents cannot be ignored in the low power design. The other important observation of the NM6 transistor is that it irregularly charges and discharges the entire differential section of the master stage during each clock cycle without any reference to the data input at the leading edge of the clock as shown in Fig. 6(a). This simultaneous charging and discharging slow-down operation of the sense amplifier must be minimized in the design for the prevention of NM6, as shown in Fig. 6(b).

C. Operation of Slave Latch

The symmetrical NAND (traditional structure) [11] or NOR logic can be used to create a slave latch. Slave latch treats SB as a set and RB as a reset input in traditional SAFF. Q, QB depends on SB, RB as SB sets Q to H, which forces QB to L, and RB sets QB to H, which forces Q to L. Therefore, one of the outputs of the slave latch will always lead the other if the latch is implemented using CMOS logic NAND gates contributes to a SAFF performance limitation.

A new slave latch that should have symmetry functionality is required to solve the question of the conventional implementation. In the technical market there are various types of latches to select the right one. The best latches for power consumption circuits include the latches such as Nikolic's [11-13], Kim's [17], Strollo's [18]. Comparative parameters such as rise delay, fall delay, power consumption

and power delay product (PDP) descriptions of these latches are shown in Table II. From the table, the latch of Strollo's is a bit faster than the rest of the latch in case of a fall delay, which makes PDP too low compared to others. But in case of rise delay, Nikolic's latch comes first after Kim's and Strollo's latches. In terms of power consumption, Strollo's consumption is about 7.5% lower than Nikolic's, as shown in Table II.
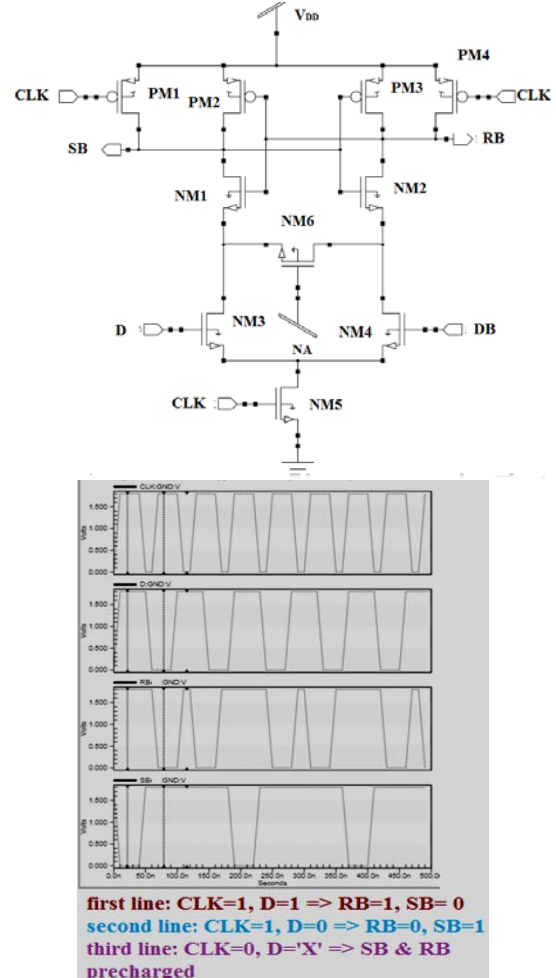

first line: CLK=1, D=1 => RB=1, SB= 0
second line: CLK=1, D=0 => RB=0, SB=1
third line: CLK=0, D='X' => SB & RB precharged

Fig. 5. Sense Amplifier in SAFF and its Simulation Waveform (Simulated using 45nm Technology).
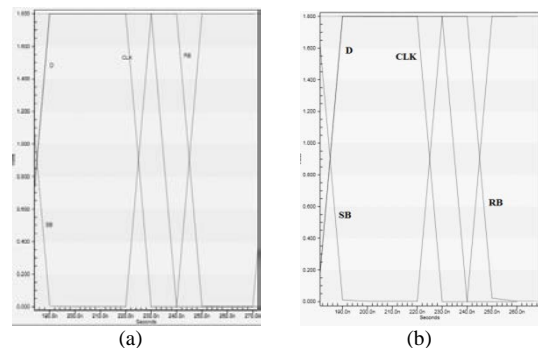

| (a) | (b) |

Fig. 6. Simulation Waveform of a Sense Amplifier (Over a Period of Sample Provides CLK-Q Delay and CLK-D Delay for both Transitions) (a) Including NM6, (b) Excluding NM6.

TABLE I.   TRUTH TABLE OF SENSE AMPLIFIER

CLK	D	DB	SB	RB
0	X	X	PRECHARGED	PRECHARGED
1	0	1	1	0
1	1	0	0	1

TABLE II.   PERFORMANCE COMPARISON OF SAFFs WITH DIFFERENT LATCHES

Parameter	Nikolic's [11]	Kim's [17]	Strollo's [18]
Number of transistors	28	26	24
Rise delay(ps)	265	221	285
Fall delay(ps)	236	231	161
Power consumption (µW)	480	425	444
Power Delay Product (PDP) (fJ)	127	98	99
Glitch free output[18]	Yes	No	Yes
Ratio less	Yes	No	Yes
Q and QB delay independence	Yes	Yes	Yes
Dual outputs	Yes	Yes	Yes

V. PROPOSED LATCH

A. Design of Proposed Latch

This paper, based on the new concept-a hybrid solution-will lead to the best architecture with the nominal parameter values through Nikolic's latch design principle and Strollo 's logic functionalities. The hybrid-symmetric structure was constructed based on Table III of the SAFF truth table. It has been constructed in 4 sub-parts with 2 pull-up (PUN1 and PUN2) and 2 pull-down (PDN1 and PDN2) parts. These sections are cross-coupled as PDN1 × PUN2 and PDN2 × PUN1 and this orientation results in a small design transistor requirement compared to conventional latch.

The output Q(next state output) of the proposed latch can be obtained from the CMOS pull down network (PDN1) given in (1) and the CMOS pull up network (PUN1) given in (2).

$$((CLK.DB+QB)SB) \tag{1}$$

$$((CLKB+DB) QB+R) \tag{2}$$

Similarly, for QB(next state output) obtained from CMOS, pull up the network (PUN2) given in (3) and pull down the network (PDN2) given in (4).

$$((D+CLKB)Q+S) \tag{3}$$

$$((CLK.D+Q)RB) \tag{4}$$

It is quite obvious from the Boolean equations that these were derived from the NIKOlic design theory and from STROllo Boolean equations so that the name of the proposed latch was described in the form of the NIKSTRO latch or the SURAV latch (because it was designed/proposed by SUdhakiran G and ernest RAVindran R S).

B. Operation of NIKSTRO/SURAV Latch

As the proposed architecture working with differential signals then from Table III, when output Q rises to logic high, then the signal R remains at logic low and RB is in logic high even though they have an inverter delay. In the pull down section of proposed latch that is equivalent to Strollo's latch, the transistors NM1, NM2, NM3 provides the discharging path to the output Q as soon as the clock transitions are from low to high and the QB discharges through NM6, NM7, NM8 so that one delay gate is needed for the proposed architecture [18].

TABLE III.   TRUTH TABLE OF SAFF

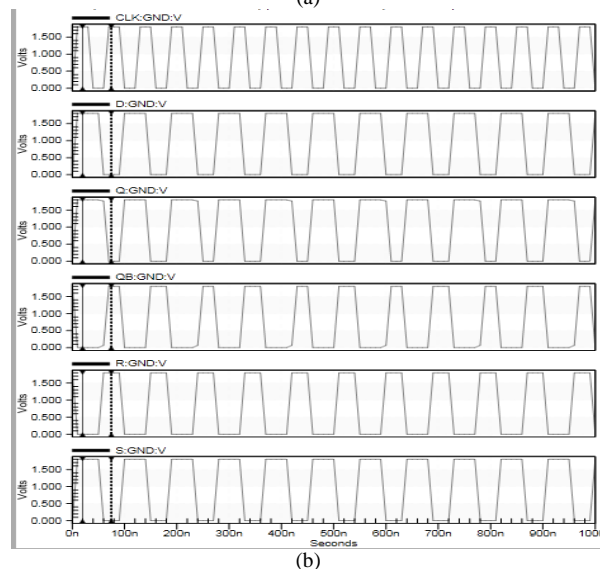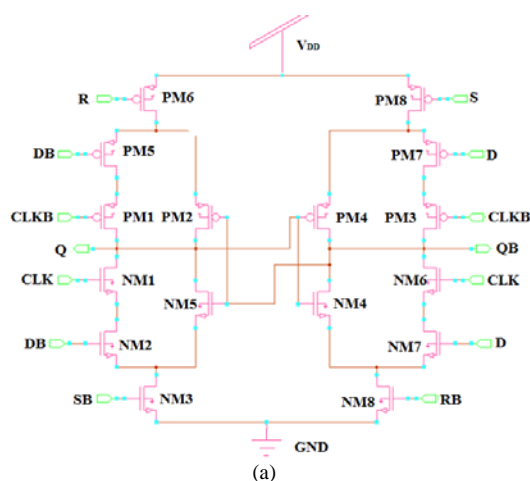| CLK	D	DB	SB (from SA)	RB (from SA)	S	R	Q	QB
0	X	X	1	1	0	0	NC	NC
1	0	1	1	0	0	1	0	1
1	1	0	0	1	1	0	1	0

(a)

(b)

Fig. 7. (a) Proposed Latch (Nikstro Latch) Structure (b) Waveforms of Nikstro Latch.

The transistors PM2, NM5 and PM4, NM6 are constitute a cross coupled inverter structure which provides the present states, outputs to the next state inputs for the operation of the latch. In the pull up section of SURAV Latch, the additional transistor PM5 is to avoid the glitches at the output. If the clock is logic low, then the MOS transistors PM1, PM3, PM6, PM8, NM3, NM8 are ON and NM1, NM6 are OFF, based on these values the cross coupled inverter sections are also in ON such that the architecture retains the logic values. When CLK is high the output Q of SURAV latch charges through PM6, PM5 and PM1which are controlled by R, DB and CLKB respectively. Similarly QB charges through PM8, PM7 and PM3, which are controlled by S, D and CLKB, respectively. Fig. 7(a) and (b) represents the proposed NIKSTRO structure and its simulated waveforms. During the simulation complement signals can be generated using an inverter. Table IV specifies the transistor aspect ratio in the Surav latch.

TABLE IV. W/L RATIOS OF MOS TRANSISTOR IN THE PROPOSED LATCH

MOS transistors	W/L ratio
PM1 – PM8	4
NM1 – NM8	2
Inverters (additional) PMOS NMOS	4 2

C. Algorithm for Designing and Operating NIKSTRO/SURAV Latch

Step 1: Inputs for the NIKSTRO latch have been obatined from the Sense Amplifier section.

Step 2: According to the truth table obtained from Step1 the slave latch pull down network implemented by Strollo's Logic.

Step 3: According to the truth table obtained from Step1 the pull up network of slave latch implemented according to the Nikolic's latch in concerning with Step 2.

Step4:A hybrid slave latch is developed after Step 3 and is named as NIKSTRO / SURAV latch.

Step 5: Operation of slave latch is:

    if (CLK=1) **then**

        if (D=1) **then**

            Q=1 and QB=0 (intermediate signals S=1 and R=0)

        else

            Q=0 and QB=1(intermediate signals S=0 and R=1)

    else

        Q= previous output and QB=previous output

**Step 6**: The operation of the proposed latch is exactly the same as the normal latch

D. Sense Amplifier based Flip Flop with NIKSTRO Latch

This is also called a flip-flop due to its function because it uses latch in the construction of SAFF. From Table III, it is very clear that the latch can capture and sustain the transition to the next rising edge of the CLK signal. The two outputs of the sense amplifier stage are considered to be normal when the CLK reaches a true low level. The whole idea therefore turns out to be flip-flop. The designed and simulated SAFF is shown in Fig. 8(a) and (b), respectively. Fig. 8(a) is complex to understand, the condensed and rearranged diagram shown in (c) is very similar to Fig. 4.

The NIKSTRO latch SAFF will only be ready for real-time applications when it operates efficiently with synchronous and asynchronous signals. Nikstro SAFF with synchronous signals discussed in the Section V. Now this is the right time to disclose the operation of the proposed SAFF with asynchronous signals-preset and clear. The circuit diagram shown in Fig. 9(a) illustrates how the asynchronous preset (PRST) and clear (CLR) signals connected to the proposed SAFF in Fig. 8(a). Fig. 9(b) represents the simulated waveforms of the proposed SAFF with asynchronous signals (considered PRST= 0) affecting the Q and QB output glitches and having a high processing time of 102ns.
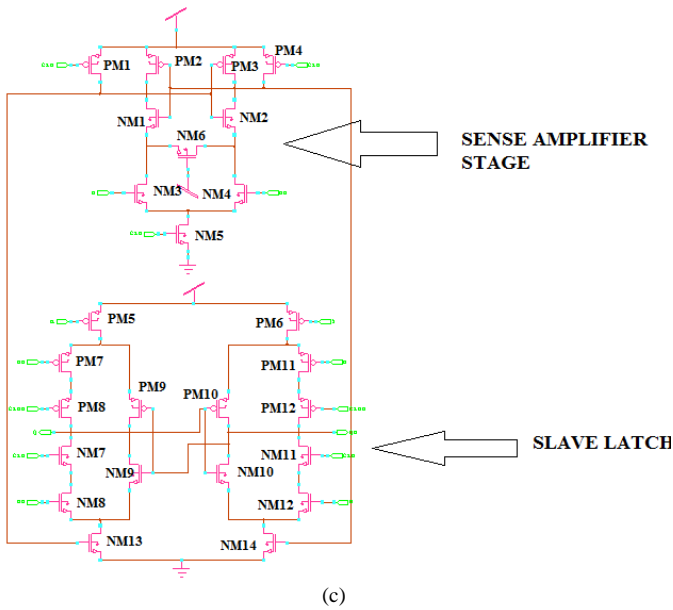

(a)

(b)

(c)
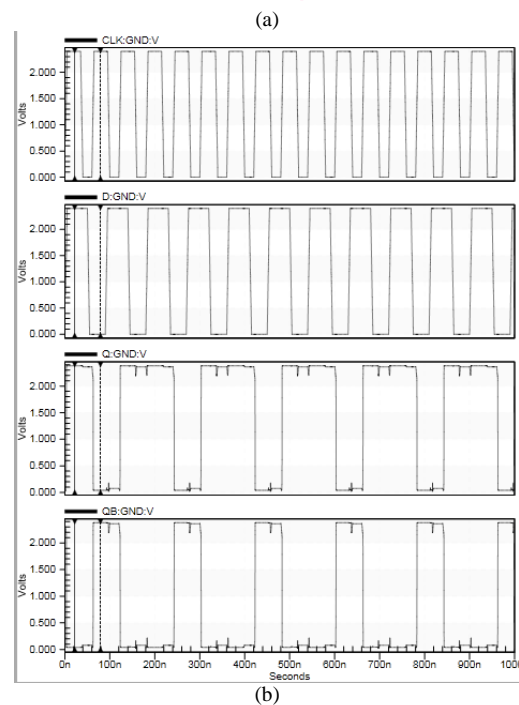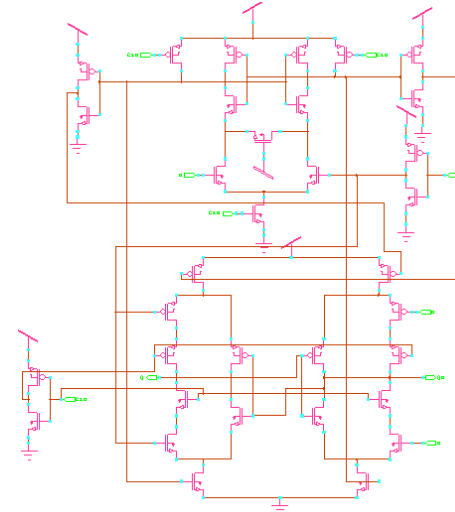
Fig. 8. Proposed SAFF with NIKSTRO Latch (a) Design (b) Simulated Waveforms (c) Simplified SAFF with NIKSTRO Latch.
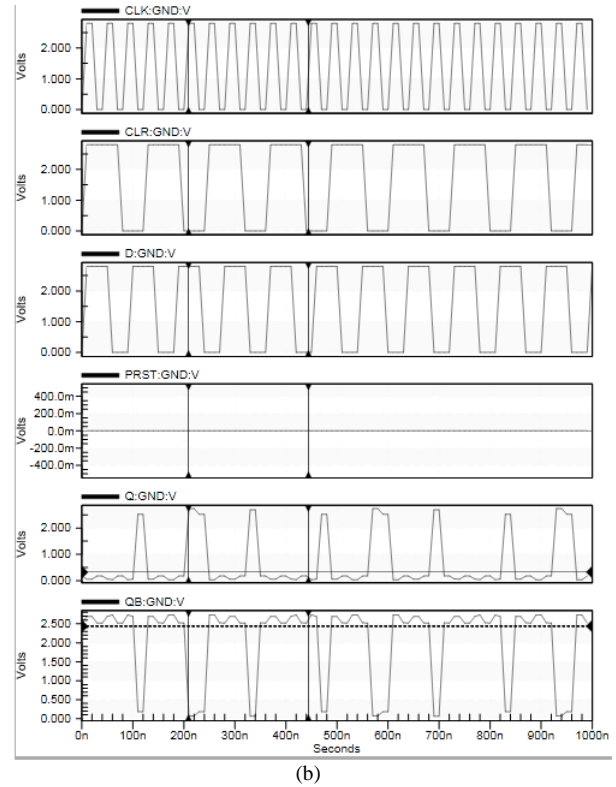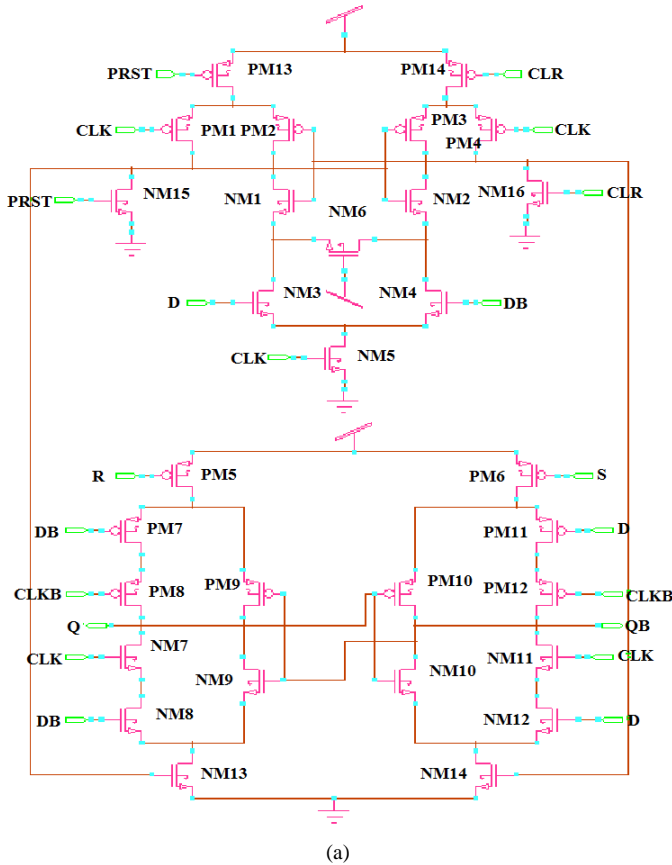

(b)

Fig. 9. Proposed SAFF with Asynchronous Signals Preset and Clear (a) Design (b) Waveforms (with Preset Signal Value Zero).

VI. DIGITAL PHASE FREQUENCY DETECTOR (DPFD)

ADPLL's overall performance is contingent on each block's output contribution. The main blocks in ADPLL are DPFD, Digital Loop Filter (DLP), Digital Control Oscillator (DCO) as shown in Fig. 10.

Parameters such as phase noise minimization, low noise and fast ADPLL lock can mainly be obtained from DPFD. This can only be accomplished by properly organizing and observing the input reference signals (pf_{REF}) and the signal produced by the DCO (pf_{DCO}). In reference to the frequency and phase of the input reference signal and the DCO feedback signal DPFD produces the output signal (UP and DOWN). The difference between UP and DOWN signals increases the phase noise of the DPFD and modulates the output that alters the entire activity of the ADPLL [20]. The conventional PFD must provide fast locking, high operating speed and low dead zone [21, 22]. The only way to achieve the above properties for ADPLL is to concentrate and properly manage the dead zone condition. The main goal of this paper is to reduce the dead zone situation to the minimum possibility. It can be achieved by simultaneously increasing the output signals of the DPFD when it is locked. The ideal PDF diagram is shown in Fig. 11.
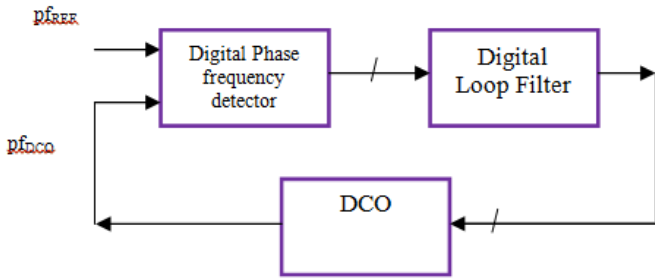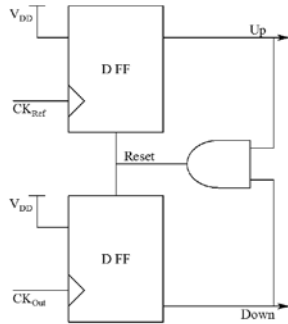

(a)

Fig. 10. General Block Diagram of ADPLL [19].

Fig. 11. Conventional Digital Phase Frequency Detector [20-22].

Old PFD methods used to reduce the dead zone, by keeping the fixed delay element [21-23] or variable delay element [20] in the reset path. But these procedures trigger disruptions in the control voltage of the oscillator leading to the reference spurs at the ADPLL output [20-24]. This paper focuses on a novel approach to provide the dead zone free and low phase noise architecture for DPFD. Initially, the produced UP and DOWN DPFD signals are low, once the clock of any flip flop rises, the outputs are enabled accordingly. This can be continued until the output of the other flip flop is going to high. When the two output signals are high, the entire circuit is reset. The DPFD performance of charging and discharging depends on the clock frequency of flip flops. Ideally, there should be a linear relationship, but it basically varies. The phase difference changes when the frequency changes since they are related as shown in (5) [20].

$$\Delta\varphi = 2\pi\left[\frac{t_{pf_{ref}} - t_{pf_{DCO}}}{\max\left(t_{pf_{ref}}, t_{pf_{DCO}}\right)}\right] \tag{5}$$

Where $t_{pf_{ref}}$ and $t_{pf_{DCO}}$ are the time periods of the input reference signal and DCO feedback signal.

These incoming signals determine the set or reset condition of DPFD. If ADPLL is in locked state, then the phase difference between these two signals is much less. During this condition if any of the flip flops received the raising edge, then it will set and produce the high output. Meanwhile, if another flip flop detects the rising edge then it will also set and produce the high output. Thus the two outputs of DPFD are high which causes a reset signal for the flip flops and makes DPFD disable for the phase difference detection. This phenomenon is called as a dead zone in which the DPFD is unable to detect the phase differences which are smaller than the dead zone leads increase in overall phase noise of

ADPLL. The conventional PLLs have a divider in its feedback path which relates the phase noise spectral density which has a strong relation between phase noise and the delay width (T_D) of the delay element. Therefore, phase noise cannot be decreased below a certain amount without reducing the dead zone [20]. The novel approach which uses a low power, high robust architecture is introduced here to reduce the dead zone of ADPLL. The proposed DPFD is implemented using the Sense Amplifier based flip flops (SAFF) shown in Fig. 9. The implementation block diagram and its simulated waveforms are shown in Fig. 12(a) and (b) with preset signal is set to zero since preset requirement is not required in DPFD operation.

For the designing of PFD, the W/L ratios can be taken from Table V for SAFFs and for CMOS inverter and NAND gate the W/L ratio for PMOS is taken as 4 and for NMOS 2 it is based on the drain to source current I_{DS} is given in eq. 6.

$$I_{DS} = \frac{k}{2}\frac{W}{L}\left(V_{GS} - |V_{T_n}|\right)^2(1 + \lambda V_{DS}) \tag{6}$$
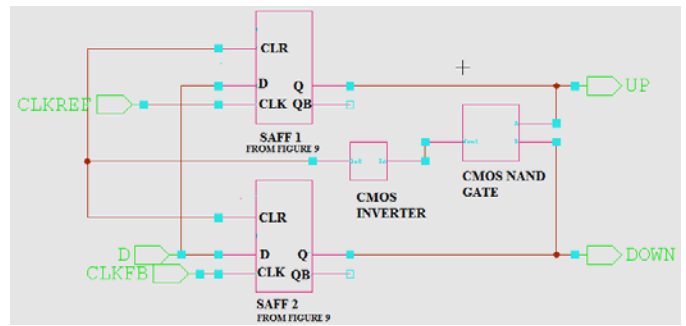
(a)
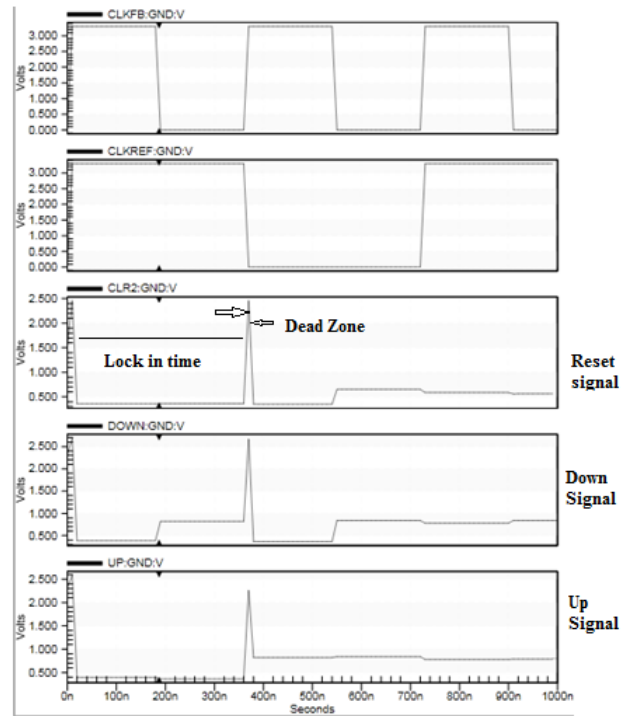
(b)

Fig. 12. Proposed DPFD with Preset Signal Setting to Zero (a) Block Diagram (b) Waveforms.

Where k is a process gain factor $= \frac{\mu\epsilon}{t_{ox}}$ μ,$\mathcal{E}$ and t_{ox} values depends on technology used, these values are different for PMOS and NMOS transistors.

$$\mu = \frac{average\ carrier\ drift\ velocity}{electric\ field}$$

λ – Is the empirical channel length modulation co-efficient.

The propagation delay in the mentioned architectures is calculated as.

$$\sum_{i=1}^{n} t_{pi} \qquad (7)$$

n- Number of delay elements in the stages

t_{pi} - total propagation delay

The low to high transition (t_{pLH}) at the output can be derived as (similar approach is followed for high to low transition also (t_{pHL})). Therefore:

$$t_{pHL} = \frac{C_L \frac{V_{OH}}{2}}{\frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_{n(p)}\left(V_{OH}-V_{T_{n(p)}}\right)^2} = t_{pLH} \qquad (8)$$

where $C_L = C_G + C_P$

$V_{OH} = V_{DD}$

Equation 8 can be modified from equation 6 as follows

$$t_{pHL} \approx \frac{C_L}{k_n V_{DD}(1+\lambda V_{DD})} \qquad (9)$$

$$t_{pLH} \approx \frac{C_L}{k_p V_{DD}(1+\lambda V_{DD})} \qquad (10)$$

where k_n & k_p are the process gain parameters of NMOS & PMOS respectively

Therefore the propagation delay of an inverter is given as

$$t_p = \frac{1}{2}(tpHL + tpLH)$$

From equation 9 & 10 the propagation delay of an inverter is given as

$$t_p = \frac{C_L}{V_{DD}(1+\lambda V_{DD})}\left(\frac{1}{k_n} + \frac{1}{k_p}\right) \qquad (11)$$

According to the number of delay stages used in the designing, the total propagation delay is calculated by using (7).

The clock to data and output delays provided by the circuit shown in Fig. 8(a) are: the clock to data (CLK – D) delay is recorded as 30ns for low to high transition and for high to low transition it took 15ns. Similarly for clock to Q (CLK – Q) the low to high and high to low transitions reported as 40ns and 22ns respectively. These delays are approximately equal to the delays caused by the voltage variable delay element PFD architecture [20] and the theoretical calculations done using (7) & (11).

VII. PERFORMANCE ANALYSIS

A. Proposed SAFF

The mixed transistor size was used in Fig. 9, details of which are shown in Table V, in order to achieve the right results. The figures in the table are the transistor's width / length ratio to the technology of 45 nm. The minimum transistor size is indicated by the * symbol.

TABLE V. W/L RATIOS OF MOS TRANSISTOR IN THE PROPOSED SAFF (L=45NM)

MOS transistors	W/L ratio
PM1,PM4	8
PM2,PM3	2.32*
PM5-8,PM11,PM12	11
PM9,PM10	2.32*
NM1, NM2	12
NM3,NM4	16
NM5	2.32*
NM6	24
NM7,NM8,NM11-14	11
NM9,NM10	2.32*

B. SAFF Performance Comparison

SAFF performance comparison with existing ones is given in Table VI and Fig. 13.

TABLE VI. PERFORMANCE COMPARISON OF SAFFS WITH DIFFERENT LATCHES

Parameter	Nikolic's [11]	Kim's [17]	Strollo's [18]	NIKSTRO (proposed)
Technology	-	-	250nm	45nm
Number of transistors	28	26	24	26 (figure 8(c))
Rise delay(ps)	265	221	285	296
Fall delay(ps)	236	231	161	180
Power consumption (µW)	480	425	444	132
Power Delay Product (PDP) (fJ)	127	98	99	32
Glitch free output[18]	Yes	No	Yes	Approximately yes (figure 8(c))
Ratio less	Yes	No	Yes	Yes
Q and QB delay independence	Yes	Yes	Yes	Yes
Dual outputs	Yes	Yes	Yes	Yes

Fig. 13. Comparison of SAFF with different Latches.

C. Proposed DPFD

Fig. 12(b) shows the transient response of each signal in the DPFD proposed. This will make it very clear that the outputs of flip flops become active when the two clock signals are high. But this state cannot be sustained for a long time; within 20nsecs the clock signals receive other inputs. Therefore the dead zone for the proposed DPFD is 20ns. Another advantage of the proposed one is, it increases the lock-in time to 340ns such that it provides a bandwidth of 8.33 MHz's for locking operation. As the proposed architecture requires quite a bit large silicon area compared to the existing architectures, the stacking transistors have been forced to use the supply voltage between 2.915V to 3.3V. For the purpose of simulation, 3.3V was considered in this paper. As a result, the power consumption is marginally increased and results in a value of 4.8mw.

TABLE VII.   DPFD COMPARISON WITH EXISTING ONES

Para meter	[8]	[10]	[20]	[25]	[26]	[27]	[28]	Prop osed
Design	ADP LL	ADP LL	PFD	ADP LL	AD PL L	AD PL L	PFD	PFD
Techn ology (nm)	130	180	90	90	28	22	130	45
V_{DD}(V)	0.7	1.8	1.8	0.52/ 1	1	-	1.2	2.9 - 3.3
Lock in time (µs)	-	-	<1	0.8 (4 Cycle s)	<1	12	-	0.34
Dead zone (ns)	-	-	36	-	-	-	52	20
Power consu mptio n (mW)	0.84 @412 MHz	22.68 @1.2 6GHz	1 @500 MHz	0.92 @600 MHz	0.64	18.4	0.496 @128 MHz	4.8 @ 1.81 GHz

VIII. CONCLUSION

In the dead zone and the locking range with a little more power consumption, this approach obtained better performance. Dead zone and lock in time are recorded for this design. The lock-in-time for the new architecture is roughly 340ns, which is greater than the fixed delay architecture. Another advantage of the proposed one is that the dead zone area is about 20ns less than the architectures referred in Table VII. For the implementation the wide silicon area is needed as it requires more transistors in the design, which is a constraint on the proposal. A further limit for the proposed architecture is a high phase noise of order -8.9db/Hz. However, the cumulative analysis of all parameters reveals a marginally increased phase noise and power usage, with the design being best achieved with appropriate lock in time & dead-zone modulation. This takes a closer look on the proposed design to use in BIMDs.

REFERENCES

[1] Kateryna Bazaka and Mohan V. Jacob "Implantable Devices: Issues and Challenges", Electronics, 2, 1-34, 2013.

[2] Achraf Ben Amar, Ammar B. Kouki and Hung Cao "Power Approaches for Implantable Medical Devices", Sensors, 15, 28889-28914, 2015.

[3] Arjun Ramaswami Palaniappan, Liter Siek, "A TDC-less all-digital phase locked loop for medical implant applications", Microprocessors and Microsystems 69 (2019) 168–178.

[4] Chung-Cheng Su, Cheng-Chung Lin, and Chung-Chih Hung, "An All-Digital Phase-Locked Loops with a Multi-Delay-Switching TDC", 978-1-5090-3969-2/17/$31.00 ©2017 IEEE.

[5] Roberts, Gordon & Ali-Bakhshian, Mohammad. (2010). A Brief Introduction to Time-to-Digital and Digital-to-Time Converters. Circuits and Systems II: Express Briefs, IEEE Transactions on. 57. 153 - 157. 10.1109/TCSII.2010.2043382.

[6] Chen Yao, "Time to Digital Converter used in ALL digital PLL", a Thesis In System-on-Chip Design, Stockholm, 08, 2011.

[7] R. Machado, J. Cabral and F. S. Alves, "All-Digital Time-to-Digital Converter Design Methodology Based on Structured Data Paths," in IEEE Access, vol. 7, pp. 108447-108457, 2019.

[8] J. Bae, S. Radhapuram, I. Jo, T. Kihara and T. Matsuoka, "A low-voltage design of controller-based ADPLL for implantable biomedical devices," 2015 IEEE Biomedical Circuits and Systems Conference (BioCAS), Atlanta, GA, 2015, pp. 1-4, doi: 10.1109/BioCAS.2015.7348405.

[9] R.G. Dreslinski , M. Wieckowski , D. Blaauw , D. Sylvester , T. Mudge , "Near threshold computing: reclaiming Moore's law through energy efficient integrated circuits", Proc. IEEE 98 (February(2)) (2010) 253–266.

[10] X. Deng, Y. Mo, X. Lin and M. Zhu, "A 0.68-to-1.44 GHz low-jitter all-digital phase-locked loop with a novel PFD and a high resolution DCO in 0.18µm CMOS," 2016 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC), Hong Kong, 2016, pp. 104-107, doi: 10.1109/EDSSC.2016.7785220.

[11] B. Nikolic, V. G. Oklobdzija, V. Stojanovic, Wenyan Jia, James Kar-Shing Chiu and M. Ming-Tak Leung, "Improved sense-amplifier-based flip-flop: design and measurements," in IEEE Journal of Solid-State Circuits, vol. 35, no. 6, pp. 876-884, June 2000, doi: 10.1109/4.845191.

[12] Borivoje Nikolic, San Jose; Wenyan Jia, Milpitas, SENSE AMPLIFIER BASED FLIP-FLOP Texas Instruments Incorporated, Dallas, TeX. USOO6107853A, United States Patent, Aug. 22, 2000.
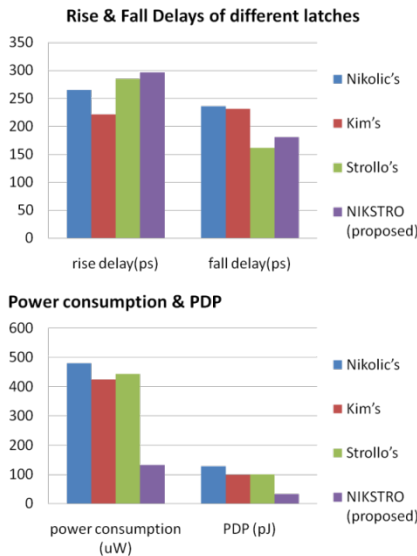
[13] J. Lin, Y. Hwang, C. Wong and M. Sheu, "Single-ended structure sense-amplifier-based flip-flop for low-power systems," in Electronics Letters, vol. 51, no. 1, pp. 20-21, 8 1 2015, doi: 10.1049/el.2014.3922.

[14] Zhijun Pei, Lei Han, Yaxin Wang, "Sense Amplifier Based Comparator Design for SAR ADC", in 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC 2019).

[15] D. Gupta and R. Bhardwaj, "Performance comparison of SR and nikolic sense amplifier based energy resumption flip-flops at 90nm CMOS technology," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018, pp. 146-148, doi: 10.1109/ICISC.2018.8399054.

[16] H. Jiang et al., "Single-Event Performance of Sense-Amplifier Based Flip-Flop Design in a 16-nm Bulk FinFET CMOS Process," in IEEE Transactions on Nuclear Science, vol. 64, no. 1, pp. 477-482, Jan. 2017, doi: 10.1109/TNS.2016.2636865.

[17] Jin-Cheon Kim, Young-Chan Jang and Hong-June Park, "CMOS sense amplifier-based flip-flop with two N-C/sup 2/MOS output latches," in Electronics Letters, vol. 36, no. 6, pp. 498-500, 16 March 2000, doi: 10.1049/el:20000409.

[18] A. G. M. Strollo, D. De Caro, E. Napoli and N. Petra, "A novel high-speed sense-amplifier-based flip-flop," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 13, no. 11, pp. 1266-1274, Nov. 2005, doi: 10.1109/TVLSI.2005.859586.

[19] Y. Ho and C. Yao, "A Low-Jitter Fast-Locked All-Digital Phase-Locked Loop With Phase–Frequency-Error Compensation," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 24, no. 5, pp. 1984-1992, May 2016, doi: 10.1109/TVLSI.2015.2470545.

[20] U. Nanda and D. P. Acharya, "An efficient technique for low power fast locking PLL operating in minimized dead zone condition," 2017 Devices for Integrated Circuit (DevIC), Kalyani, 2017, pp. 396-400, doi: 10.1109/DEVIC.2017.8073978.

[21] S. S. Kuncham, M. Gadiyar, D. K. Sushmitha, K. K. Lad and T. Laxminidhi, "A Novel Zero Blind Zone Phase Frequency Detector for Fast Acquisition in Phase Locked Loops," 2018 31st International Conference on VLSI Design and 2018 17th International Conference on Embedded Systems (VLSID), Pune, 2018, pp. 167-170, doi: 10.1109/VLSID.2018.56.

[22] Y. He, X. Cui, C. L. Lee and D. Xue, "An improved fast acquisition PFD with zero blind zone for the PLL application," 2014 IEEE International Conference on Electron Devices and Solid-State Circuits, Chengdu, 2014, pp. 1-2, doi: 10.1109/EDSSC.2014.7061226.

[23] A. Homayoun and B. Razavi, "Analysis of Phase Noise in Phase/Frequency Detectors," IEEE Transactions on Circuits and Systems I: Regular Papers, vol.60, no.3, pp.529-539, March 2013.

[24] Chen Wu-Hsin, M.E. Inerowicz and Jung Byunghoo, "Phase Frequency Detector with Minimal Blind Zone for Fast Frequency Acquisition," IEEE Transactions on Circuits and Systems II: Express Briefs, vol.57, no.12, pp.936-940, Dec. 2010.

[25] C. Chung, W. Su and C. Lo, "A 0.52/1 V Fast Lock-in ADPLL for Supporting Dynamic Voltage and Frequency Scaling," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 24, no. 1, pp. 408-412, Jan. 2016, doi: 10.1109/TVLSI.2015.2407370.

[26] S. Höppner, S. Haenzsche, G. Ellguth, D. Walter, H. Eisenreich and R. Schüffny, "A Fast-Locking ADPLL With Instantaneous Restart Capability in 28-nm CMOS Technology," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 60, no. 11, pp. 741-745, Nov. 2013, doi: 10.1109/TCSII.2013.2278123.

[27] Y. W. Li, C. Ornelas, H. S. Kim, H. Lakdawala, A. Ravi and K. Soumyanath, "A reconfigurable distributed all-digital clock generator core with SSC and skew correction in 22nm high-k tri-gate LP CMOS," 2012 IEEE International Solid-State Circuits Conference, San Francisco, CA, 2012, pp. 70-72, doi: 10.1109/ISSCC.2012.6176934.

[28] W. Chen, M. E. Inerowicz and B. Jung, "Phase Frequency Detector With Minimal Blind Zone for Fast Frequency Acquisition," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 57, no. 12, pp. 936-940, Dec. 2010, doi: 10.1109/TCSII.2010.2087951.

Comprehensive Interaction Model for Cloud Management

Md. Nasim Adnan[1]

Department of Computer Science and Engineering
Jashore University of Science and Technology
Jashore-7408, Bangladesh

Md. Majharul Haque[2], Abu Sadat Mohammad Yasin[5]
Muhammad Shakil Pervez[6]

Bangladesh Bank, Dhaka, Bangladesh

Mohammad Rifat Ahmmad Rashid[3]

University of Liberal Arts Bangladesh
Dhaka, Bangladesh

Mohammod Akbar Kabir[4]

Department of Economics
University of Dhaka, Dhaka, Bangladesh

Abstract—Cloud computing is readily being adopted by enterprises due to its following benefits: ability to provide better service to customers, improved flexibility, lower barrier to entry for an enterprise, lower maintenance cost on IT service, availability etc. However, the interaction between cloud service provider and customer is not well-defined yet. Understanding of the service offered while approaching cloud computing paradigm and also understanding of the required actions during the period of receiving a cloud service e.g. provision of new resources, scaling up/down, billing, etc. remains a concern for the enterprises. This paper proposes a segregated interaction model to manage the receiving of a cloud service in a hierarchical way.

Keywords—*Cloud computing; cloud management; cloud customer*

I. INTRODUCTION

This Nowadays the concept of a cloud computing is one of the emerging issues in the domain of Information Technology (IT) service industry [1], [2]. Cloud computing [3], [4] is the realization of a long-held desire to offer computing (e.g. networks, servers, storage, and application services) as a service like other utilities e.g. electricity, gas, water etc. It offers a an array of solutions and advantages to business with increased flexibility, scalability and agility with reduced procedural, administrative, hardware and software costs and with higher efficiency [4], [5]. Due to these benefits, the growth in adoption of cloud has been phenomenal [6]. Gartner has the forecast that the cloud computing industry is poised for strong growth through 2014, when worldwide cloud services revenue is projected to reach $148.8 billion [7].

Cloud computing enables us to enroll into subscription-oriented services in a "pay-as-you-go" manner by leasing IT capabilities through the Internet. As a result, new entrepreneurs no longer require investing capital for IT Infrastructure. Instead, they can take the opportunities of provisioning of resources based on usage requirement and pay based on consumption. Thus, entrepreneurs can transfer risks and maintenance aggravation at the cost of converting capital expenses to operational expenses. Cloud computing also gives an illusion of ample resources available on demand for enormous scalability. Cloud computing has been defined by many thinkers and experts in many different ways. Among them, we will use the definition given by National Institute of Standards and Technology (NIST). According to NIST: "Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [8].

A cloud can be public or private. A private cloud is behind the company firewall and is not avaiable to external entity whereas in a public cloud system (Fig. 1), general public or external enterprises can get computing services as utility. This public cloud now triggers the need for cloud service management on the receiving side, which is addressed by this paper. On the other hand, cloud service providers consider service delivary model e.g. Insfrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS) and manage cloud by employing differernt management models e.g. consumer utilization model to maximise their utilization and hence profit [9], [10].

Many people are aware of the concept of cloud but do not have a clear understanding of the management of receiving a cloud service and thus reluctant to adopt the cloud. On the other hand, many people already have shifted to cloud but unable to gain a real benefit of being in the cloud due to absense of a proper approach. This paper offers a comprehensive interaction model for cloud customers covering the enitre interaction scenarios by using a set of use cases, so that users can approach cloud with a systematic way and can realize maximum benefits from cloud.

The remainder of the paper is structured as follows: Section 2 provides an anatomy of services provided by a cloud. A breakdown of a cloud system with a mapping to use cases is provided in Section 3. In Section 4, an interaction model for cloud management is proposed. Finally, Section 5 provides some useful concluding remarks.
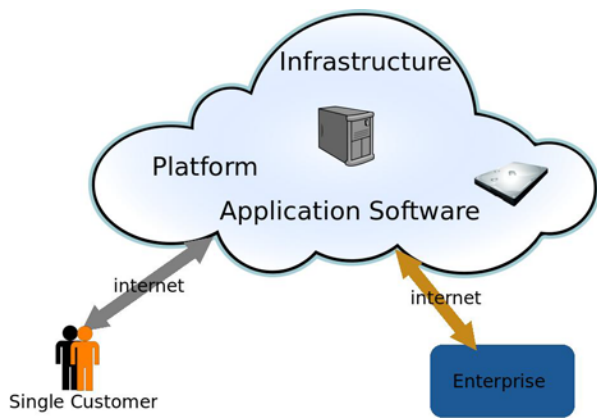
Fig. 1. A Public Cloud.

II. CLOUD SERVICE

To manage a cloud, it is very essential to obtain an understanding of a typical cloud service. A cloud service is similar to a recurring business activity. A cloud user can get one or more services for a period of time, from one or multiple cloud providers. Cloud providers may charge cloud customer for their contract/consumption of resources by hour, month or year. Cloud service is defined by the Open Group as: "A service is a logical representation of a repeatable business activity that has a specified outcome, is self-contained, may be composed of other services" [11].

A cloud service is initiated, when a cloud provider makes an offering of a cloud service and makes it available for potential customers. A service is ended, when the particular service is no longer available for any customer. A typical cloud service has five states in its lifecycle (see Fig. 2):

- **Available:** When a cloud provider makes an offering of a service along with associated cost, policies, etc. and makes it available for potential customers.

- **Contract:** When cloud customer(s) make contract to get a service.

- **Runtime Maintenance:** Cloud provider deploys resources for customer(s) as per contract i.e. makes the resources available for the customer(s) and manages the resources.

- **End of Contract:** When cloud customer(s) terminate the contract for the service.

- **Unavailable:** When the cloud provider(s) or other external event(s) make the service unavailable.

Now, consider a scenario of a business organization, where IT infrastructure is required. If the organization invests on IT infrastructure more than the actual demand, then a part of the capital is wasted, whereas having less of IT capability will hamper business growth. These businesses can take the benefit of cloud by leasing IT capabilities in a systematic approach. Moreover, demand of IT infrastructure varies with time; some business face periodic demand, some face very high demand for a short period of time. Some of the common features of IT demand are shown in Fig. 3. Instead of leasing a constant IT capability, an organization can realize a tailored service(s) by

following a usage pattern from old consumption data in a cloud. To realize the full benefit of being in cloud, users should be aware of all the functionalities within the system boundary and this can be achieved by following a model like we propose in this paper.
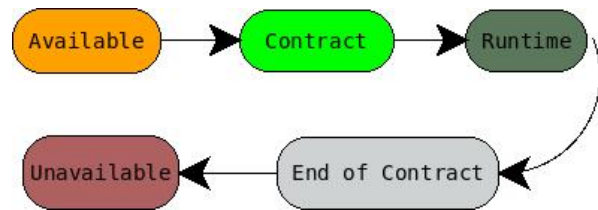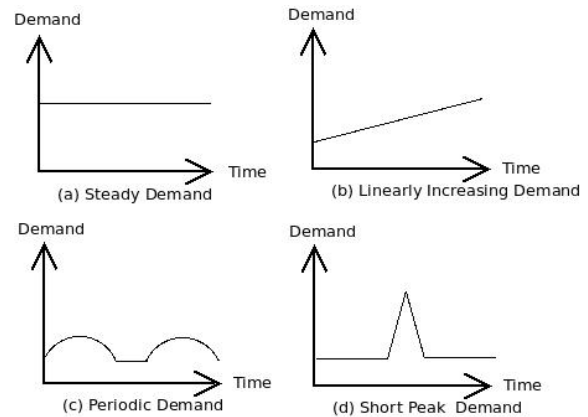

Fig. 2. Lifecycle of a Cloud Service.

Fig. 3. Some Common usage Patterns.

III. USE CASE

Use cases [11] are very useful to capture and communicate the functional characteristics of a system in a clear and easy way without internal details. It is a list of steps defining who does what and thus it describes the interactions between stakeholders and helps to achieve the system goal(s). A complete set of use cases can describe all the different ways to use the system and bounds the system scope. In a use case, a stakeholder is represented as an actor.

1) Actors: In a system, an entity that is able to make decisions is represented as an actor which may not necessarily be a human. In cloud system, there are two high level actors: *Cloud User* and *Cloud Provider*.

2) Cloud user: Cloud user is the representation of a person or company, who consumes cloud-based service(s) and pays in return and triggers the need for effective cloud usage management.

3) Cloud provider: Cloud provider is a representation of a company, who provides cloud-based services (infrastructure, platform, application software, etc.).

In the proposed interaction model, we have two low level actors in each high-level group: Business Manager (for business decision making) and Technical Manager (for technical decision making) as shown in Fig. 4. In the proposed model, we use the term "customer" to represent an entity that pays for receiving cloud services while user is the superset of customer.
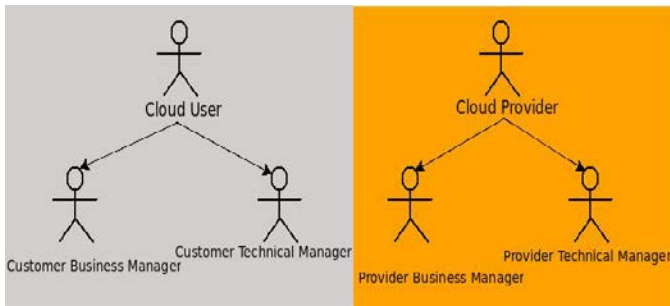
Fig. 4.   Classification of Actors in a Cloud System.

IV. INTERACTION MODEL

We have already defined the actors in the previous section. Now, we will define the roles of the actors to describe all the functions in a cloud system. The system functions can be categorized into three groups: business contract level, operational level and feedback level functions. In business contract level, we will describe the system functions that are attached with legal bindings. The system functions that are actually used to receive a cloud service are grouped into operational level, whereas feedback level functions feed into business contract level so that business contracts can be done or changed in an optimal way. A complete and comprehensive set of use cases are shown in Fig. 5.

A. Business Contract Level Functions

Establish Relationship: It is essential for a cloud provider to be certain of the identity of the cloud customer and provide it with security credentials (Table I).

Establish Contract: Cloud customer and provider should enter into legal bindings before providing or receiving any service. Cloud Service level Agreements (SLA) acts as a key liaison between consumers and providers on renting Anything as a Service (AaaS) [12]. In this stage, an SLA is agreed upon by both parties (Table II). Due to the dynamic nature in cloud, continuous monitoring on Quality of Service (QoS) characteristics is essential to administer SLAs [13]. While preparing SLA, the following points should be considered from the cloud customer point of view:

a) Everything should be written in SLA.

b) There must be a certain guarantee of performance such as service availability, response time, etc. and there must be a financial penalty in presence of failure to meet such performance level.

c) The licensing of required software(s) must be explicitly defined.

d) It is better to get service from multiple vendors to avoid vendor lock-in.

e) The time required to provision resources must be included in SLA.

f) Data ownership, confidentiality, auditability must be defined and the vendor must return data in case of change in vendor.

g) There must be provisions for change (Table III) and terminate (Table IV) the contract.
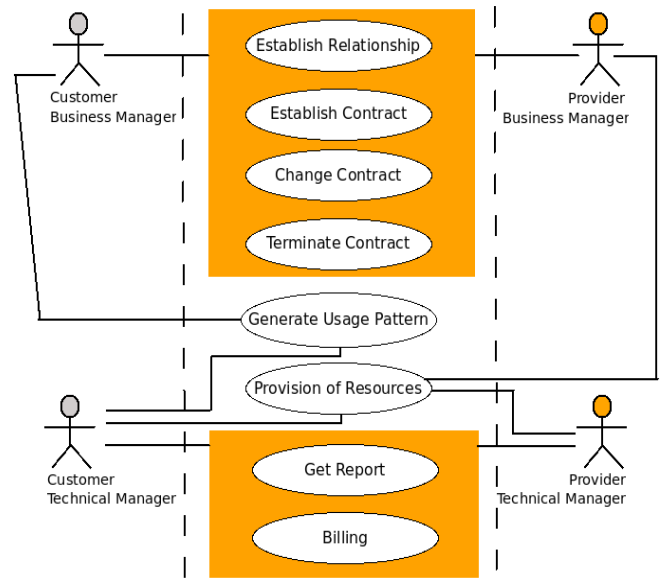

Fig. 5.   A Comprehensive Set of use Cases Defining System Boundary.

B. Operational Level Functions

Provision of Resource: By using this function, a cloud customer can reserve resources for its use within the limit mentioned in the SLA and these resources are kept aside for the particular customer until the resources are released (Table V).

Get report: Cloud customer gets a report on the resource consumption to use it for internal purpose and also these reports are used to find a pattern in its usage (Table VI).

Billing: Cloud customer gets an invoice for billing and makes the payment (Table VII).

TABLE I.   USE CASE TO ESTABLISH RELATIONSHIP

Use Case 1	Establish Relationship	
Description	A potential customer of a cloud-related service wants to establish an identity with a cloud service provider and gets security credential in return.	
Goal in context	The customer have an identity with the cloud provider.	
Actors	• Provider Business Manager (PBM) • Customer Business Manager (CBM)	
Preconditions	None	
Trigger	Business needs of a customer.	
Scenarios	Scenarios Details	Data or Artifacts Exchanged
	CBM Browses the services offered by cloud provider.	Service catalog
	PBM asks for security credential.	None
	CBM provides identity, bank details etc.	None
	PBM provides CBM with the security credntial.	Security token
Exceptions	The process can be done offline or online.	

TABLE II.   USE CASE TO ESTABLISH A SERVICE CONTRACT

Use Case 2	Establish a Service Contract	
Description	A potential customer of a cloud-related service wants to establish a contract with a cloud service provider.	
Goal in context:	The customer enters into a legal binding with the cloud service provider as per Service Level Agreement (SLA) with sufficient technical and business information.	
Actors	• Provider Business Manager (PBM) • Customer Business Manager (CBM)	
Precondition:	Customer Business Manager has established identity with Provider and can login in the interface provided by the cloud service provider.	
Trigger	Business need to start receving of a cloud service.	
Scenarios	Scenarios Details	Data or Artifacts Exchanged
	CBM logs in.	Security credential
	CBM selects the intended service.	None
	PBM informs the associated terms and conditions.	Proposed SLA
	CBM agrees to the proposed SLA and enters into legal bindings.	SLA
Exceptions	It can be done offline or online.	

TABLE III.   USE CASE TO CHANGE CONTRACT

Use Case 3	Change Contract	
Description	Initiated either by cloud customer or provider and agreed by both, their legal responsibilities changes.	
Goal in context	Both the cloud customer and provider agree and make changes to SLA if allowed within the SLA.	
Actors	• Provider Business Manager (PBM) • Customer Business Manager (CBM) • Provider Technical Manager (PTP)	
Precondition	The customer is already in a legal binding with the cloud service provider.	
Trigger	Business need to change the service received.	
Scenarios	Scenario Details	Data or Artifacts Exchanged
	CBM logs in.	Security credential
	Change configurations in SLA and submits for approval by PBM.	SLA
	CBM is notified with the reply from PBM.	SLA
	If PBM agrees, then PTP apply those changes.	None
Exceptions	• CBM and PBM can terminate the existing SLA and create a new one. • PBM can also initiate to change SLA.	

TABLE IV.   USE CASE TO TERMINATE CONTRACT

Use Case 4	Terminate Contract	
Description	Initiated either by cloud customer or provider their contract cease to exit.	
Goal in context:	Both the cloud customer and provider agree to terminate the contract. The resources kept for that particular customer is freed and ready to be allotted to other customers.	
Actors	• Provider Business Manager (PBM) • Customer Business Manager (CBM)	
Precondition	The customer is already in a legal binding with the cloud service provider.	
Trigger	CBM no longer requires the cloud service.	
Scenarios	Scenarios Details	Data or Artifacts Exchanged
	CBM/PBM logs in and request for termination of contract.	Security credential
	Another party agrees.	None
	Contract gets null and void.	None
	There can be provision of penalty for immature termination of SLA.	
Exceptions	CBM/PBM logs in and request for termination of contract.	

TABLE V. USE CASE TO PROVISION OF RESOURCES

Use Case 5	Provision of Resources	
Description	Cloud customer makes a reservation of resources wihtin the limit set by SLA.	
Goal in context	Cloud provider sets aside resources to be consumed by the cloud customer.	
Actors	• Provider Technical Manager (PTM) • Customer Technical Manager (CTM)	
Precondition:	Cloud customer and provider have a SLA and the requested amount and type of resources is permitted according to SLA.	
Trigger	Business requirements to use resources through a cloud service.	
Scenarios	Scenarios Details	Data or Artifacts Exchanged
	CTM logs in.	Security credential
	CTM requests for resources.	Service
	PTM checks against the SLA and makes necessary arrangements to reserve requested resources.	None
	CTM is notified about the result	None
Exceptions	None	

TABLE VI. USE CASE TO GET CONSUMPTION REPORT

Use Case 6	Get Consumption Report	
Description	Cloud provider sends a detailed report of the consumption by the cloud customer.	
Goal in context:	Cloud customer gets detailed informations about the consumption of resources by itself.	
Actors	• Provider Technical Manager (PTM) • Customer Technical Manager (CTM)	
Precondition:	Cloud customer and provider have a SLA.	
Trigger:	To have an updated status of the usage.	
Scenarios	Scenarios Details	Data or Artifacts Exchanged
	CTM logs in and request for consumption report.	Security credential
	PTM provides consumption report by using historic data.	None
	CTM observes the detailed report on the consumption of resources.	None
	None	
Exceptions	CTM logs in and request for consumption report.	

TABLE VII. USE CASE TO BILLING

Use Case 7	Billing	
Description	Cloud provider sends an invoice as a charge for the consumption/contract of the cloud customer and gets payment.	
Goal in context	Cloud customer makes a payment for his consumption/contract.	
Actors	• Provider Technical Manager (PTM) • Customer Technical Manager (CTM)	
Precondition	Cloud customer and provider have a SLA.	
Trigger	Payment for the service received	
Scenarios	Scenarios Details	Data or Artifacts Exchanged
	PTM sends an invoice for billing.	
	CTM takes action for payment.	None
Exceptions	None	

C. Feedback Level Functions

Generate Usage Patterns: In this function, a cloud customer tries to find a pattern of its use against time. With the understanding of its internal demand, a cloud customer can make changes in SLA or get multiple different type of contracts to maximizing its goal(s) (Table VIII).

TABLE VIII.   USE CASE TO GENERATE USAGE PATTERNS

Use Case 8	Generate Usage Patterns	
Description	Customer Cloud Manager tries to generate a pattern of usage from historical data to reduce overprovisioning and also underprovisioning.	
Goal in context:	Cloud customer finds a pattern from earlier consumption data.	
Actors	• Customer Technical Manager (CTM) • Customer Business Manager (CBM)	
Precondition:	Cloud customer has been receiving service from cloud provider.	
Trigger:	Identify usage pattern to reduce cost.	
Scenarios	Scenarios Details	Data or Artifacts Exchanged
	CTM Generates pattern of their consumption.	None
	CBM gets information from CTM so that s/he can update SLA more efficiently.	None
Exceptions	None	

V.   CONCLUSION

In this paper, we have considered the problem of receiving a service in a cloud computing system. We propose a segregated interaction model to manage cloud service in a hierarchical way. The developed interaction model covers all the major functions in the defined system boundary. The presented model identifies the major decision makers for cloud management and their relationship and provides a segregation of system functions to approach the considered problem easily and effectively. At the end, the model analyses the system functions with the help of use cases. In future, data mining services can be incorporated in order to analyze usage patterns more effectively [14],[15],[16],[17].

REFERENCES

[1] Zhang, Q., Cheng, L. Boutaba, R.: Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications, Vol. 1(1), pp. 7–18 (2010).

[2] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: A view of cloud computing. Communications of the ACM, Vol. 53(4), pp. 50–58 (2010).

[3] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: Above the Clouds: A Berkeley View of Cloud Computing. Technical Report, EECS Department, U.C. Berkeley (2009).

[4] Avram, M. G.: Advantages and Challenges of Adopting Cloud Computing from Enterprise Perspective. Procedia Technology, Vol. 12(1), pp. 529–534 (2014).

[5] Yeo, C. S., Venugopal, S., James Broberg, J., Buyya, J. B. R.: Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems, Vol. 25, No. 6, pp. 599–616 (2009).

[6] Xu C. T. S., Xin, M. T. W.: Benefits and Challenges of the Adoption of Cloud Computing in Business. International Journal on Cloud Computing: Services and Architecture, Vol. 6, No. 6, pp.1–15 (2016).

[7] Gartner: Forecast: Public Cloud Services. Worldwide and Regions, Industry Sectors 2009–2014, (2010).

[8] Mell, P., Grance, T.: The NIST Definition of Cloud. National Institute of Standards and Technology, Information Technology Laboratory, Retrievedon September 15, 2011,from http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[9] Definition of SOA, http://www.opengroup.org/soa/soa/def.htm, last accessed 2011/08/15.

[10] Imran, A.: An Overview of Service Models of Cloud Computing. International Journal of Multidisciplinary and Current Research, Vol. 2, pp. 779–783 (2014).

[11] Malan, R., Bredemeyer, D.: Functional requirements and use cases. Bredemeyer Consulting, Retrieved on August 11, 2011 from http://www.bredemeyer.com/pdf_files/ functreq.pdf, (2001).

[12] Saravanan, K., Rajaram, M.: An Exploratory Study of Cloud Service Level Agreements - State of the Art Review. KSII Transaction of Internet and Information System, Vol. 9, No. 3, (2015).

[13] Aljoumah, E., Al-Mousawi, F., Ahmad, I., Al-Shammri, M., Al-Jady, Z.: Cloud Computing Architectures: A Comprehensive Study. International Journal of Grid Distribution Computing, Vol. 8, No. 5, pp. 7–32 (2015).

[14] Adnan M. N.: On Dynamic Selection of Subspace for Random Forest. The 10th International Conference on Advanced Data Mining and Applications (ADMA 2014), pp. 19 – 21 (2014).

[15] Adnan M. N., Islam M. Z.: ComboSplit: Combining Various Splitting Criteria for Building a Single Decision Tree. The International Conference on Artificial Intelligence and Pattern Recognition (AIPR), pp. 17 – 19 (2014).

[16] Adnan, M. N., Islam, M. Z.: ForEx++: A new framework for knowledge discovery from decision forests. Australasian Journal of Information Systems, Vol. 21, pp. 1–20 (2017).

[17] Adnan, M. N., Islam, M. Z.: Forest PA: Constructing a decision forest by penalizing attributes used in previous trees. Expert Systems with Applications, Vol. 89, pp. 389–403 (2017).

Determining the Presence of Metabolic Pathways using Machine Learning Approach

Yara Saud Aljarbou[1]

College of Computer Sciences and Engineering
Taibah University, Madinah
Saudi Arabia

Fazilah Haron[2]

College of Computer and Cyber Sciences
Prince Muqrin University, Madinah
Saudi Arabia

Abstract—The reconstruction of the metabolic network of an organism based on its genome sequence is a key challenge in systems biology. One of the strategies that can be used to address this problem is the prediction of the presence or the absence of a metabolic pathway from a reference database of known pathways. Although, such models have been constructed manually, obviously such a method cannot be used to cover thousands of genomes that has been sequenced. Therefore, more advanced techniques are needed for computational representation of metabolic networks. In this research, we have explored machine learning approach to determine the presence or the absent of metabolic pathway based on its annotated genome. We have built our own dataset of 4978 instances of pathways. The dataset consists of 1585 pathways with each having 20 different representations from 20 organisms. The pathways were obtained from the BioCyc Database Collection. The pathway dataset also consists of 20 features used to describe each pathway. In order to identify the suitable classifier, we have experimented five machine learning algorithms with and without applying feature selection methods, namely Decision Tree, Naive Bayes, Support Vector Machine, K-Nearest Neighbor and Logistic Regression. Our experiments have shown that Support Vector Machine is the best classifier with an accuracy of 96.9%, while the maximum accuracy reached by the previous work is 91.2%. Hence, adding more data to the pathway dataset can improve the performance of the machine learning classifiers.

Keywords—*Metabolic pathway prediction; pathway dataset; metabolic network of organism; machine learning; support vector machine*

I. INTRODUCTION

Constructing a comprehensive model of metabolic reaction networks which occurs within every organism is a key step toward understanding the metabolism of an organism. The availability of metabolic networks as predictive tools is fundamental in many research fields such as metabolic engineering, diagnostic medicine, pharmacology, biochemistry, biology and physiology. An exciting example of using metabolic networks is the screening of disease-specific biomarkers that can be applied for early detection of diseases.

Although several of metabolic network models have been constructed through manual processing, such an approach obviously cannot be used when we have thousands of sequenced genomes. Therefore, more advanced techniques are needed for computational representation of metabolic networks.

The pathway prediction is one of the methods that can help in the reconstruction of an organism's metabolic network from its genome sequence. In the pathway prediction problem: by having the annotated genome of an organism and its reactome, we can predict the set of metabolic pathways present in the organism. In this research, by taking the reactome as predetermined by other methods, we can focus on developing an improved pathway prediction methods.

The pathway prediction can involve predicting novel pathways that have not been previously observed which also called pathway discovery, or predicting pathways that were previously known in other organisms. Our methodology does the latter, predicting pathways from a selected reference database.

Our main motivation is to develop a more accurate method for predicting metabolic pathways and to overcome the limitations of the current existing methods. It is expected that machine learning will help provide effective knowledge from a variety of big data including data about metabolism.

The aim of our research is to improve the accuracy of the metabolic pathway prediction process. The outcome of this research is a comparative study of our chosen machine learning algorithms with the work that have been established in this domain.

The rest of the paper is organized as follow: Section II provides a background knowledge of our research, and presents the existing researches that are related to our research. Section III provides the details about the methodology that has been adopted in this research. Section IV describes the experimental results obtained from our experiment. Section V summarizes the whole research and provides some suggestions on how the work could be further improved.

II. BACKGROUND AND LITERATURE REVIEW

This section provides a background knowledge of our research, and also presents the works that are related to our research.

A. Backgound of Metabolic Pathways

Bioinformatics is a rapid development research area for the scientific community. The bioinformatics research focuses on algorithms, statistical approach, computations approach and developing huge databases to solve problems in the biology

field. One of major research efforts in this field is metabolic pathway prediction.

In biochemistry, the metabolism is a series of chemical reactions that take place within the cells of organisms based on enzymes; which are necessary to ensure that the organism survives [1]. Enzymes in the process play an active role in the reproduction and growth of living organisms and they stimulate the organisms to interact with and respond to their environment. It also said that the term metabolism, referred to all the biochemical processes carried out by the bodies of the biological organisms starting from the production of new tissues based on basic nutrients breaking down of carbohydrates, sugars and fats then turning them into energy for the body to carry out daily activities. There are two main goals for metabolism [2]: the first is gaining energy that enables the cell to perform its functions through demolition reactions, which also known as Catabolism, and the other goal is compounding complex organic compounds that are necessary for the cell through building reactions, which also known as Anabolism.

These two goals rarely achieved by a single chemical reaction. Rather, the dominant rule is to produce energy or synthesize the compounds through a number of successive reactions so that the material from the first reaction is a reactive substance in the second reaction. A set of reactions that transform a specific material into another material called metabolic pathway. A metabolic pathway has many steps, each step begins with a specific molecule and ends with a product, and each reaction is catalyzed by a specific enzyme as shown on the Fig. 1.
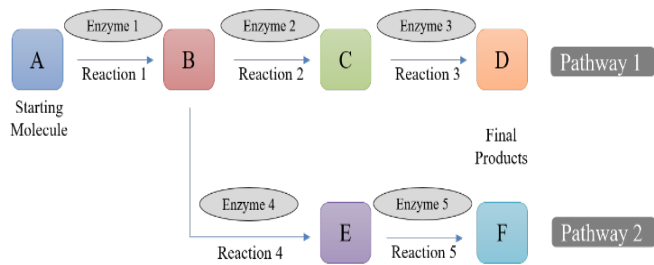

Fig. 1. Metabolic Pathways.

There are three main goals for studies conducted on metabolic pathways [3]: (1) Identification of intermediate compounds and enzymes involved in a series of reactions that positively understand the activation and inhibition of this path. (2) Identify the mechanism of an interaction in a series of reactions that requires the separation of enzymes that help in each reaction. (3) Identify the mechanism of regulating the speed of different reactions.

The principal axiom of Systems Biology is that a system should be also analyzed at the level of interactions of its parts, not only as sum of them.

B. Related Work

There have been many previous works on solving the metabolic pathway prediction problem, such as the computational method that was developed to compare organisms based on genome-wide metabolic pathway analysis

[4], using the WIT (What Is There) database, which is a metabolic pathway profile for each completed genome [5]. These profiles are records of the presence and absence of the various metabolic pathways, and constitute the basis for a comparison of organisms. The developed methodology requires that all the reactions in a pathway to have enzymes in order for the pathway to be consider present.

KEGG project on "pathway maps" based on the information of the genome [6]. KEGG pathway maps encompassed varied metabolic pathways from varied organisms. One of the issues faced in the project is the problem of pathway map prediction rather than the problem of pathway prediction. The description of KEGG's algorithm for map prediction and the accuracy evaluation of that algorithm are no ware to be found.

Matthews1 et al. [7], performed prediction of metabolic pathways based on the information of the genome stored in the Reactome Knowledgebase, which is an online, manually curated resource that provides an integrated view of the molecular details of human biological processes that range from metabolism to DNA replication and repair to signaling cascades [8]. However, the description of their algorithm and the accuracy evaluation of their algorithm are no ware to be found.

In [9] and [10], Kastenmüller et al. developed an outcome similar to the "information content" features used in the predictors of [11], calculating the fraction of reactions present in the pathway, weighted in terms of the unity of the reaction. It is expected that such analyses could be improved by taking advantage of the probabilities of pathway presence.

Al Daoud [12] developed a new algorithm to predict pathway classes and individual pathways for a previously unknown query molecule. His main idea was to use a dense graph, where the enzymes are represented as edges and the compounds as vertices. The weights are assigned to the edges according to the previous known pathways. He applied the shortest path algorithm for each missing enzyme in a pathway. A pathway is considered to be belong to an organism if the total cost between the initial and final compound is higher than a threshold. The validation of their experiments showed that the suggested algorithm is capable to classify more than 90% of the pathways correctly.

None of the above work involve in the predicting the absence or presence of pathway. There is PathoLogic [13], which is a known tool that can be used to predict the presence or the absence of metabolic pathways in sequenced and annotated genomes.

Another highly related work to ours is Dale et al. [11], they developed a machine learning methods for metabolic pathway prediction, their method showed a better performance when compared with the standard methods presented in the hard-coded pathway prediction tool PathoLogic [13], while at the same time allowing easier explanation, tenability, and extensibility of the results. Table I summarizes a comparison between the PathoLogic tool and the machine learning approach.

TABLE I.   COMPARISON BETWEEN PATHOLOGIC AND ML APPROACH

PathoLogic Algorithm [13]	ML Algorithm (Logistic Regression) [11]
Accuracy of 91%	Accuracy of 91.2%
No additional information	Provide probability for each predicted pathway
Requires Experts	No Experts Required
Developed and refined over approximately a decade	Developed with well-designed collection of input features

Machine learning algorithms that has been tested in pathway prediction included logistic regression, decision trees and naive bayes. The goal of this study was to test different machine learning methods for the determination of the presence or the absence of a metabolic pathway based on the pathways information for many organisms presented in the pathway collection MetaCyc and to develop new predictors for determining the presence of a metabolic pathway in newly sequenced organisms. In order to evaluate their methods, Dale et al. have developed the gold standard pathway dataset [11].

III. METHODOLOGY

This section provides the details about the methodology that has been adopted in this research.

The overall research methodology was based on the machine learning process. The methodology is divided into four phases, namely data acquisition, data preprocessing, training, and model evaluation. We decided to use Weka, which is one of the most popular and freely available machine learning tool [14].

The first phase of our methodology was the data acquisition phase, in which we collected the relevant data (pathways information) from the BioCyc databases for the study and computed the values of the features based on their description. The data preprocessing phase was the second phase, in which the collected data was cleaned and integrated such that, the datasets were proper for the process of classification. We also applied feature selection methods on the dataset to select the most effective feature to train our machine learning algorithms. The data from the second phase which is the data preprocessing phase were then passed over to the third phase. The third phase is the training of the machine learning algorithms, which consists of two parts; without feature selection and with the selected feature from second phase. In the fourth phase, the model evaluation and comparison phase, we have tested the classifiers without feature selection and the classifiers that used the selected features by using standard 10-fold cross validation. We also performed a comparative analysis between the different classifiers based on widely used evaluation metrics. Fig. 2 represent the overall methodology of our research. Each phase of our research methodology will be explained in detail in the following sections.
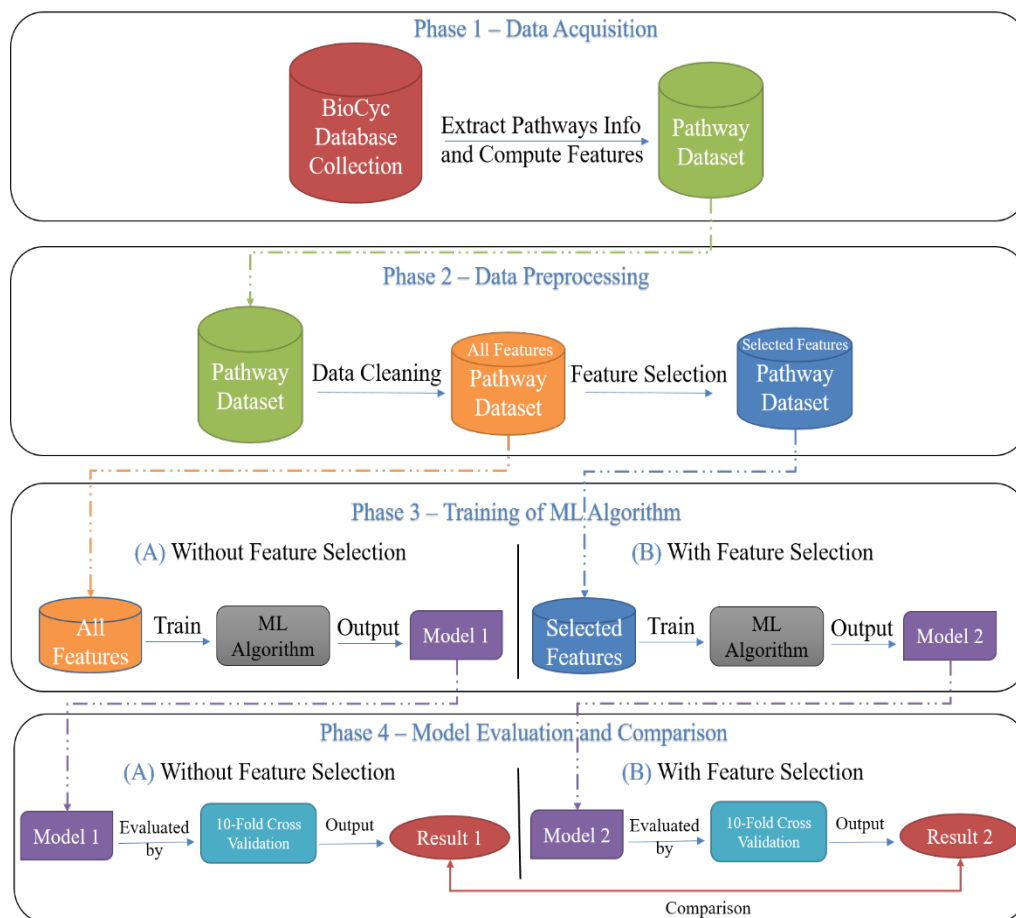

Fig. 2.   Overall Research Methodology.

A. Data Acquisition Phase

In order to train our machine learning methods and validate them against each other, we have constructed a pathways dataset containing known information about the presence or absence of pathways in different organisms. Here we describe the construction and content of the dataset.

First, we accessed the pathways databases from the BioCyc Database Collection, which is a collection of 14560 Pathway/Genome Databases (PGDBs) [15]. We then obtained the list of features and their description that has been used by Dale et al. in [11]. The pathways dataset currently contains 4979 elements that describe pathway absence and presence in 20 organisms as shown in Table AI in the Appendix, along with the corresponding Pathway/Genome Databases (PGDBs) [15]. All our data were mainly derived from PGDBs. Each element or instance of the dataset is a triple of the form (Organism Name, Pathway ID, Is-Present?), as shown in Table II.

TABLE II.　　Attribute Description of Pathway Dataset

Attribute	Description
Organism Name	The organism's name.
Pathway ID	The pathway's ID.
Is-Present?	The presence or absent of the pathway (Label).

Organism's name, is not considered as a factor that determines the presence or the absent of a particular pathway. We only use it to determine the number of instances (different representation) of a pathway. For example, we have 20 different representations of pathway 1 from 20 different organisms.

We used the MetaCyc Metabolic Pathway Database, which is "a curated database of experimentally elucidated metabolic pathways from all domains of life, as a reference for the curated metabolic pathways, MetaCyc contains 2801 pathways from 3123 different organisms" [16]. In order determine the presence or the absent of the pathways in each organism, we applied the two rules. For each organism, the first rule is to mark as positives (present) all the pathways that are present in the database corresponding to each organism [11]. Then, we added as negatives (absent) all pathways in the same databases but has not been annotated in MetaCyc database [11]. As for the second rule, we added as negatives (absent) all the pathways that have no enzymes [4]. As for features that

describe each pathway, we used some of the features that have been defined and used by Dale et al. [11] due to the absent of expert in this domain. Table AII in the Appendix describes the features that we extracted and used in our research.

At the end of this phase, we have a complete version of the pathway dataset. The pathway dataset has 4979 instances and 22 attributes including the ID of the pathways, the label of each pathway (present or absent) and finally the 20 features that describe each pathway. The number of unique pathways in the pathway dataset is 1585 pathways with each having 20 different representations from 20 organisms as compared to the original dataset that has been constructed by Dale et al. [11], which only have 6 organisms.

After describing each part of the pathway dataset, Table III shows a sample of the complete pathway dataset. The "…" in Table III refers to the remaining features. In the sample, we only mentioned two of the 20 features, the Biosynthesis-Pathway and the Num-Reactions features.

B. Data Preprocessing Phase

The data preprocessing stage, were the collected data from the data acquisition phase was integrated, and then the data went through the cleaning process, in which we deleted unfitting entries of the data, such as those that provide unrelated results in the dataset.

We also applied feature selection methods on the pathway dataset in order to get the most effective features in determining the presence or the absent of each pathway in the pathway dataset. Feature selection methods can reduce both the computational complexity and the data in the dataset. Therefore, the dataset can also be more useful and efficient to train the classification algorithms [17]. One of the feature selection methods that we used is Information Gain (IG), "which measures how much 'information' a feature gives us about the class or the label. Features that perfectly partition should give maximal information and produce a higher score than the unrelated features that give us no information about the value of the class or the label" [18]. The other feature selection method that we used is Correlation-Based Feature Selection (CFS), which "evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them, the features subsets that are highly correlated with the class while having low inter-correlation are preferred" [19].

TABLE III.　　Sample of Pathway Dataset

Organism Name	Pathway ID	Features			Is-Present?
		Biosynthesis-Pathway	…	Num-Reactions	
Escherichia coli	CYANCAT-PWY	FALSE	…	3	PRESENT
Escherichia coli	ARO-PWY	TRUE	…	0	ABSENT
Arabidopsis thaliana	PWY-3781	FALSE	…	4	PRESENT
Arabidopsis thaliana	PWY-6754	FALSE	…	2	ABSENT
Bacteroides thetaiotaomicron	GLYSYN-PWY	FALSE	…	1	PRESENT
Bacteroides thetaiotaomicron	PWY-7353	FALSE	…	2	ABSENT

In our research, we have used both of these feature selection methods in order to get the most effective features in determining the presence or the absent of a particular pathway. More details about the experiment will be discussed later in later section.

C. Training of Machine Learning Algorithm Phase

In this research, five commonly used classification algorithms [20], namely, K-Nearest Neighbor, Decision Tree, Naive Bayes, Logistic Regression and Support Vector Machine were evaluated. The first four algorithms were chosen since the purpose of the work is to compare with the work in [11], where they use the same algorithms. Support Vector Machine was chosen because it is also one of the most widely used classification algorithms.

D. Model Evaluation and Comparison Phase

For evaluating the performance of the prediction techniques, we used several performance measures that are widely used. The method for evaluating the classification models was checking the confusion matrix. The confusion matrix contains information about the predicted and the actual classifications that we get from the proposed classifier [21]. The other evaluation metrics assessed for effectiveness measurement were classification accuracy, specificity, and sensitivity or recall [21].

In order to test the models that we got from the training of machine learning algorithm phase, we used the k-fold cross validation. In this test, the dataset is randomly divided into K equal parts, one part is selected to test the model, and the k-1 is the remaining part used to train the model [22]. In our research, we used the standard 10-folds cross validation to test and evaluate our models. The main advantage of this method is that it is simple to implement and does not require much time in the calculation process.

IV. Experimental Results and Discussion

In this section we describe the experimental results that we got from the last three phases of the research methodology, and these are data preprocessing, training of machine learning algorithm, and model evaluation and evaluation.

A. Feature Evaluation and Selection

In the data preprocessing phase, we have applied two feature selection methods in order to reduce both the computational complexity and the data in the pathway dataset.

The first method is information gain, which measures how much "information" a feature gives us about the class or the label, and base on this information, the information gain method assigns scores to the features and rank them based on these scores. Based on these scores, we started to remove the features with the lower score one by one and at each time, we used the remaining features to train the machine learning algorithm and observed its accuracy. The machine learning algorithm that we used in this experiment is the Naive Bayes. The results of this experiment are shown in Fig. 3.

The results showed that by using the feature with the higher score, we reached the highest accuracy which is 92.7%, while when using the two features with highest score, the accuracy

reduced to 89.4%. From this, we can say that the gain information method does not tell us what are the best combination of features to reach higher accuracy, therefore; another feature selection method is needed.

The other feature selection method that we have used in our experiment is the correlation-based feature selection, which evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The best subset of features that has been selected by this method is a combination of the features Deg-Or-Detox-Pathway, Is-Sub-Pathway, Has-Enzymes and Has-Key-Reactions. By using these features to train our machine learning algorithms, we reached a higher accuracy level.

B. Evaluation of Classification Models without Feature Selection

In the first part of the training of machine learning algorithm phase, we constructed five classification models using the pathway dataset for the Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor and Support Vector Machine classifiers. The models were constructed by using all the features in the pathway dataset without applying any feature selection methods. We used the standard 10-fold cross-validation for testing our models in the model evaluation and comparison phase.

Fig. 4 shows the results of the experiment, the accuracy, sensitivity, and specificity rate of the classification models for the Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor and Support Vector Machine classifiers.
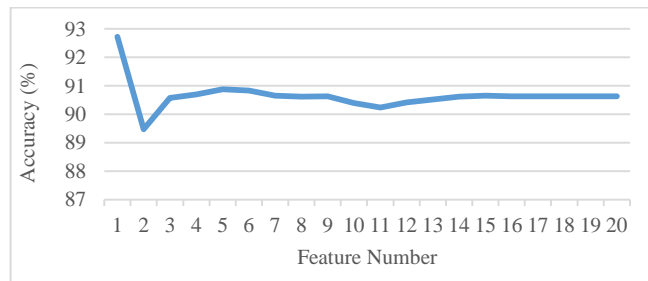

Fig. 3. Accuracy (%) of Naive Bayes with Feature Selection by Information Gain Method.
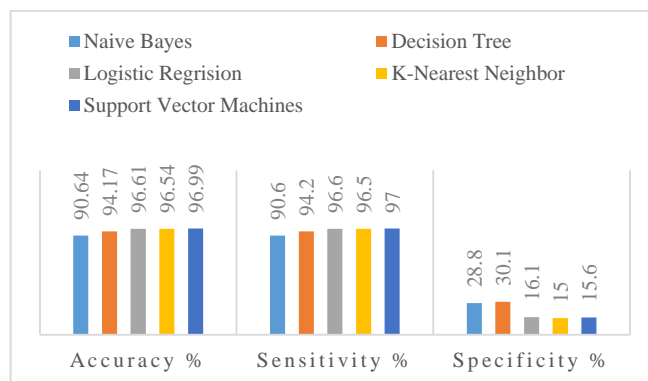

Fig. 4. Classification Models' Performance without Feature Selection.

The experiment results indicated that the accuracy percentage of the pathway classification model that used Support Vector Machine as the classifier gave 96.9%, which is the highest compared to the Naive Bayes, Decision Tree, Logistic Regression and K-Nearest Neighbor. In this experiment, we did not apply any feature selection methods.

The sensitivity, which is the proportion of the present pathways classified as present, we can see that the classifier of the Support Vector Machine gave the highest sensitivity of 97%. As for the specificity, which is the proportion of the absent pathways classified as absent, the classifier that gave the highest specificity of 30.1% was the Decision Tree. Therefore, we can say that among the five classifiers that we have trained and evaluated, the best classifier in predicting the present pathways is the Support Vector Machine and the Decision Tree is the best in predicting the absent pathways.

C. Evaluation of Classification with Feature Selection

In the second part of the training of machine learning algorithm phase, we constructed the same five classification models using the pathway dataset for the Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor and Support Vector Machine classifiers, but this time, we constructed the models by using the selected features by the feature selection methods. We used the standard 10-fold cross-validation for testing our models in the model evaluation and comparison phase.

Fig. 5 shows the results of the experiment, the accuracy, sensitivity, and specificity rate of the classification models for the Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor and Support Vector Machine classifiers, which are trained by the pathway dataset that only contains the selected features.

The experiment results indicated that the accuracy percentage of the pathway classification model that used Support Vector Machine as classifier gave 96.9%, which is the highest compared to the Naive Bayes, Decision Tree, Logistic Regression and K-Nearest Neighbor. The pathway dataset used in this experiment only contained the features selected by the feature selection methods.

The sensitivity, which is the proportion of the present pathways classified as present, we can see that the classifier of the Support Vector Machine gave the highest sensitivity of 97%. As for the specificity, which is the proportion of the absent pathways classified as absent, the classifier that gave the highest specificity of 30.1% was the Decision Tree. Therefore, we can say that among the five classifiers that we have trained end evaluated, the best classifier in predicting the present pathways is the Support Vector Machine and the Decision Tree is the best in predicting the absent pathways.

D. Classification Models Comparison with and without Feature Selection

In the model evaluation and comparison phase, a comparative analysis of the constructed models with and without feature selection was performed.
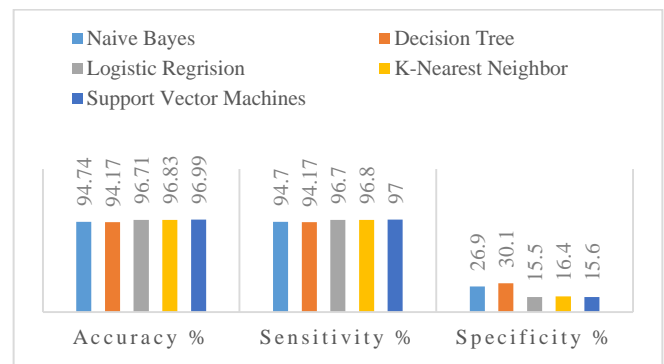

Fig. 5. Classification Models' Performance with Feature Selection.

The results show that by applying a proper feature selection method, some of the classifiers gets higher results than their corresponding classifiers without feature selection in terms of the accuracy level and while others got the same accuracy level. However, none of the classifiers produces results that are lower. The best result for pathway prediction in terms of accuracy, was given by the Support Vector Machine classifier, which obtained 96.9% accuracy, followed by K-Nearest Neighbor with 96.826 accuracy, Logistic Regression with 96.7% accuracy, Naive Bayes reaching the 94.7% accuracy, and Decision Tree with 94.1% accuracy level.

However, in the experiments without using any feature selection methods, The best result for pathway prediction in terms of accuracy, was given by the Support Vector Machine classifier which obtained 96.9% accuracy, followed by Logistic Regression with 96.6% accuracy, K-Nearest Neighbor with 96.5% accuracy, Decision Tree with 94.1% accuracy, and Naive Bayes reaching the 90.6% accuracy level. Fig. 6 shows a graphical representations of the results.

E. Comparative Analysis of our Models and the Existing Models

In this section, a comparative analysis between the models constructed by the pathway dataset that we build and the models constructed by the pathway dataset in [11] was performed. The machine learning classifiers used by Dale et al. [11] are Naive Bayes, Decision Tree, Logistic Regression and K-Nearest Neighbor, but they did not include the results for the K-Nearest Neighbor classifier. Therefore, we only compared the accuracy of the first three classifiers with the accuracy given by the classifiers constructed by our pathway dataset. A graphical representation of the comparison results based on the accuracy level of our classification models and the existing classification models shown in Fig. 7.

From Fig. 7, we can see that the models constructed by our pathway dataset out-performed the models in [11]. The Logistic Regression classifier in both experiments gave us the highest accuracy compared to the Decision Tree and the Naive Bayes classifiers.

In our research, we also constructed Support Vector Machine classifier, which gave us a higher accuracy equal to 96.9%, while the K-Nearest Neighbor classifier gave slightly lower, that is 96.8%.
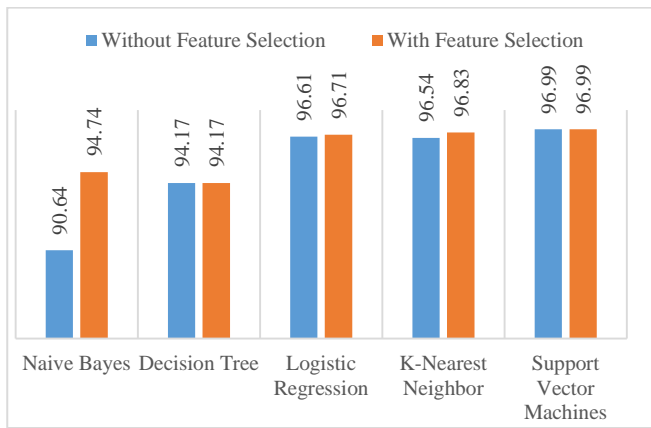
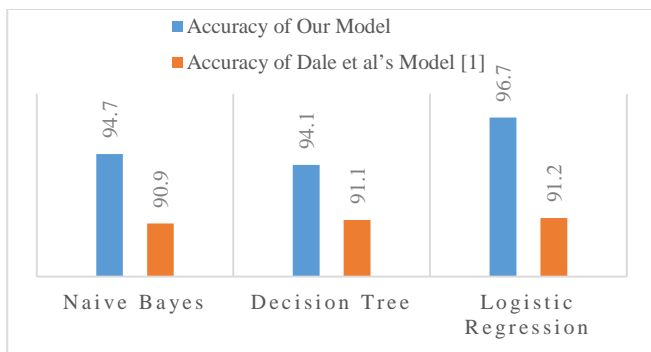Fig. 6.   Classification Models' Accuracy (%) with and without Feature Selection.

Fig. 7.   Accuracy (%) of our Models and the Existing Models.

It is to be noted that in Dale et al.'s [11], although they have discussed the used of K-Nearest Neighbor, however, they did not specifically report any results of the method. Therefore, we have excluded K-Nearest Neighbor from our comparison as shown in Fig. 7.

The maximum accuracy we got from our experiments was 96.9%, which is higher than the maximum accuracy 91.2% obtained by Dale et al. using similar machine learning algorithms [11] and also the earlier work of Dale et al.'s using Pathologic with 91% accuracy [13]. We believe that, by using an efficient pathway dataset with high number of different representation for each pathway (adding more organisms), and with well-defined set of features could improve the performance of the machine learning classifiers.

V.   Conclusion and Future Work

In this research, we have presented a comparative study of our chosen machine learning algorithms with the work that have been established in the metabolic pathway prediction field. Our study was conducted over determining the presence or the absence of a metabolic pathway. We found that only few works addressed the problem of determining the presence or the absence of a metabolic pathway. After that, we started to build a pathway dataset in order to train and evaluate the machine learning algorithms. Our methodology is built upon the four machine learning phases: data acquisition, data preprocessing, training the machine learning algorithms and model evaluation and comparison. Our results shows that the maximum accuracy we got from our experiments was 96.9% given by the Support Vector Machine classifier with and without feature selection methods.

As a future work, there are some points that can be taken into consideration in order to improve this research and these are: (1) Build a pathway dataset with more than 20 organisms in order to increase the number of representation for each pathway. (2) Include all the features that has been defined by Dale et al. or define more features that have better description for each pathway. (3) Develop a deep learning-based prediction for determining the absence or presence of a metabolic pathway.

Contributions

Our work is built upon the work of Prof. Peter D. Karp, the director of Bioinformatics Research Group, SRI International, California, USA, formerly was a Stanford Research Institute to support innovative ideas. Prof. Karp is a pioneer in metabolic pathways studies. He devised a hard-coded algorithm to predict the absence or presence of a pathway using a software tool called Pathologic. He and his colleagues' later studied selected machine learning algorithms (K-Nearest Neighbor, Decision Tree, Logistic Regression and Naive Bayes) to compare with the results of Pathologic. Our work extends Prof. Karp's work by using a new dataset of 20 organisms (as compared to six by them) and also we used an additional algorithm, which is the Support Vector Machine. Our results show an improvements in accuracy as compared to Karp's and his team.

Acknowledgment

References

[1]   Raval, "Introduction to Biological Networks," Introd. to Biol. Networks, 2018, doi: 10.1201/b14987.

[2]   T. Theorell, "Anabolism and catabolism - Antagonistic partners in stress and strain," Scand. J. Work. Environ. Heal. Suppl., no. 6, pp. 136–143, 2008.

[3]   B. Junker and S. Falk, Analysis of Biological Networks. A John Wiley & Sons, Inc., Publication, 2008.

[4]   L. Liao, S. Kim, and J. Tomb, "Genome Comparisons Based on Profiles of Metabolic Pathways," 2002, pp. 469–476.

[5]   R. Overbeek, "WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction," Nucleic Acids Res., vol. 28, no. 1, pp. 123–125, 2002, doi: 10.1093/nar/28.1.123.

[6]   S. Okuda et al., "KEGG Atlas mapping for global analysis of metabolic pathways.," Nucleic Acids Res., vol. 36, no. Web Server issue, 2008, doi: 10.1093/nar/gkn282.

[7]   L. Matthews et al., "Reactome knowledgebase of human biological pathways and processes," Nucleic Acids Res., vol. 37, no. SUPPL. 1, 2009, doi: 10.1093/nar/gkn863.

[8]   "Home - Reactome Pathway Database." https://reactome.org/ (accessed Apr. 20, 2019).

[9]   G. Kastenmüller, J. Gasteiger, and H. W. Mewes, "An environmental perspective on large-scale genome clustering based on metabolic

capabilities," Bioinformatics, vol. 24, no. 16, pp. 56–62, 2008, doi: 10.1093/bioinformatics/btn302.

[10] G. Kastenmüller, M. E. Schenk, J. Gasteiger, and H. W. Mewes, "Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes," Genome Biol., vol. 10, no. 3, 2009, doi: 10.1186/gb-2009-10-3-r28.

[11] J. M. Dale, L. Popescu, and P. D. Karp, "Machine learning methods for metabolic pathway prediction Supplementary Material Features Used in Machine Learning Predictors," BMC Bioinformatics, pp. 1–13, 2009.

[12] E. Al Daoud, "A new algorithm for Predicting Metabolic Pathways," Int. J. Eng. Sci. Invent., vol. 5, no. 8, pp. 20–24, 2016, [Online]. Available: www.ijesi.org.

[13] S. M. Paley and P. D. Karp, "Evaluation of computational metabolic-pathway predictions for Helicobacter pylori," Bioinformatics, vol. 18, no. 5, pp. 715–724, 2002, doi: 10.1093/bioinformatics/18.5.715.

[14] I. Witten, E. Frank, and M. Hall, Data Mining Practical Machine Learning Tools and Techniques, vol. 54, no. 2. 2011.

[15] P. D. Karp et al., "The BioCyc collection of microbial genomes and metabolic pathways," Brief. Bioinform., no. June, pp. 1–9, 2017, doi: 10.1093/bib/bbx085.

[16] R. Caspi et al., "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases,"

Nucleic Acids Res., vol. 44, no. D1, pp. D471–480, 2016, doi: 10.1093/nar/gkv1164.

[17] I. Guyon, "An Introduction to Variable and Feature Selection," vol. 3, pp. 1157–1182, 2003.

[18] B. Azhagusundari and A. S. Thanamani, "Feature Selection based on Information Gain," Int. J. Innov. TechnolB Azhagusundari Antony Selvadoss Thanamani. 2013. Featur. Sel. based Inf. Gain. Int. J. Innov. Technol. Explor. Eng. (IJITEE), 2(2)18–21.ogy Explor. E, vol. 2, no. 2, pp. 18–21, 2013, doi: 2278-3075.

[19] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," no. April, 1999.

[20] Jecinta Morgan, "Differences Between Supervised Learning and Unsupervised Learning | Difference Between," 2018. http://www.differencebetween.net/technology/differences-between-supervised-learning-and-unsupervised-learning/ (accessed Apr. 18, 2019).

[21] Aditya Mishra, "Metrics to Evaluate your Machine Learning Algorithm," 2018. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 (accessed Apr. 20, 2019).

[22] H. Jiawei, K. Micheline, and P. Jian, Data Mining, Concepts and Techniques. Elsevier Inc, 2012.

APPENDIX A

ORGANISMS' LIST AND DATABASES

Organism	Database	Version
Escherichia coli K-12 MG1655	EcoCyc	22.6
Arabidopsis thaliana	AraCyc	13.0
Bacteroides thetaiotaomicron VPI-5482	BtheCyc	20.1
Candida albicans SC5314	CalbiCyc	12.0
Chlamydomonas reinhardtii	ChlamyCyc	5.0
Cyanidioschyzon merolae strain 10D	CyanidioCyc	20.0
Emiliania huxleyi CCMP1516	EmilianiaCyc	20.0
Candidatus Evansia muelleri	EvaCyc	1.0.1
Cryptosporidium hominis TU502	HominisCyc	20.5
Homo sapiens	HumanCyc	20.5
Lactobacillus rhamnosus GG	LactorhaCyc	20.5
Candidatus Portiera aleyrodidarum	PabtqvlcCyc	1.0.1
Plasmodium berghei ANKA	PbergheiCyc	20.5
Phaeodactylum tricornutum CCAP 1055/1	PhaeoCyc	20.0
Prevotella copri DSM 18205	PrecopriCyc	20.5
Toxoplasma gondii ME49	ToxoCyc	20.5
Trypanosoma brucei	TrypanoCyc	10.0.1
Saccharomyces cerevisiae S288c	YeastCyc	20.5
Danio rerio	ZfishCyc	18.0
Amycolatopsis mediterranei S699	Amed713604Cyc	19.0

THE TYPES AND DESCRIPTION OF THE 20 FEATURES

Feature	Type	Description
Has-Orphan-Reaction	Boolean	True if the pathway has an orphan reaction.
Has-Spontaneous-Reaction	Boolean	True if the pathway has a spontaneous reaction.
Energy-Pathway	Boolean	True if the pathway is an energy pathway.
Deg-Or-Detox-Pathway	Boolean	True if the pathway is a degradation pathway or a detoxification pathway.
Detoxification-Pathway	Boolean	True if the pathway is a detoxification pathway.
Degradation-Pathway	Boolean	True if the pathway is a degradation pathway.
Biosynthesis-Pathway	Boolean	True if the pathway is a biosynthetic pathway.
Is-Variant	Boolean	True if the pathway is a variant pathway.
Is-Sub-Pathway	Boolean	True if the pathway belongs to any super pathways.
Multiple-Reaction-Pathway	Boolean	True if the pathway has more than one reaction.
Single-Reaction-Pathway	Boolean	True if the pathway has only one reaction.
Num-Reactions	Numeric	Number of reactions in the pathway.
Has-Enzymes	Boolean	True if there are enzymes catalyzing reactions in this pathway.
Num-Enzymes	Numeric	Number of enzymes catalyzing reactions in this pathway.
Enzymes-Per-Reaction	Numeric	Number of enzymes catalyzing reactions in this pathway, divided by number of reactions.
Has-Key-Reactions	Boolean	True if the pathway has key reactions.
Num-Output-Compounds	Numeric	Number of (primary) output compounds of the pathway.
Num-Input-Compounds	Numeric	Number of (primary) input compounds of the pathway.
Num-Input/Output-Compounds	Numeric	Number of (primary) input or output compounds of the pathway.
Num-Initial-Reactions	Numeric	The number of reactions in the pathway that have no predecessors in the pathway.

Intelligent and Scalable IoT Edge-Cloud System

Shifa Manihar[1], Ravindra Patel[3]
Department of Computer Science
UIT RGPV, Bhopal, India

Tasneem Bano Rehman[2]
Department of Computer Science
Sage University, Bhopal, India

Sanjay Agrawal[4]
Department of Computer Science
NITTTR, Bhopal, India

Abstract—**Scalability is an utter compulsory for the success of the IoT's unprecedentedly growing network. The operational and financial bottlenecks allied with growth can be overwhelming for those peeping to integrate IoT solutions. As the IoT technology proceeds, so is the scale of operations desired to arrive at a wider target region. Breakdown may take place not because of device's ability to scale, but due to data scale. As more devices are being incorporated, more data/information will be amassed, stored, processed, and scrutinized. The volume of this collection simply cannot be managed from a single edge device by deploying vertical approach. When starting small, it's important to peep into the future and anticipate growth. Companies that can't adapt to unpredictable market changes will fold without the right IoT architecture in place. Therefore, a scalable IOT framework has been proposed in the paper, which will provide load balancing or scalability by deploying the provisions of horizontal scalability for the system. The framework will be utilizing SOM for the purpose of classifying applications (whether delay sensitive or delay insensitive), so that proper decisions can be made based on the incoming data (typically signals) and if edge gets over flooded with the data, then edge is scaled to instigate the other edge for computing additional requests . The proposed system is termed as intelligent because its algorithm empowers the edge to take decision and classify applications based on the type of requirement of the application.**

Keywords—Scalability; internet of things; self organizing map; edge; horizontal scalability

I. INTRODUCTION

Internet of Things (IoT) was first introduced to the community in 1999 for supply chain management [1], and then the concept of "making computer sense information without the aid of human intervention" was widely adapted to other fields such as healthcare, home, environment, and transports [2], [3]. A prediction has been made by Analysts Firms that there will be around 2020 billion of active connected products. Now with IoT, we will arrive in the post-cloud era, where there will be a large quality of data generated by things that are immersed in our daily life, along with lots of applications deployed at the edge to consume these data.

IoT is an environment that encompasses the objects (living and non-living) communicating with each other by means of Internet. The basic purpose of IoT is to induce intelligence into any object and providing it with the decision making capability. Here we lay emphasis on the capacity building capability of any device. We do not need to have storage within the object itself. The whole system relies on offloading the computing and storage on to the network by the IoT devices.

Edge computing refers to the technologies empowering computation to be performed at the edge of the network. Here we define "edge" as any computing and network resources along the path between cloud data centers and data sources. For example, a smart phone is the edge between body things and cloud, a gateway in a smart home is the edge between home things and cloud, a micro data center and a cloudlet [4] is the edge between a mobile device and cloud. The main logic behind edge computing is that computing should happen in close proximity of mobile devices or sensors. Data is increasingly produced at the edge of the network; therefore, it would be more efficient to also process the data at the edge of the network. Previous work such as micro datacenter [5], [4], cloudlet [6], and fog computing [7] has been introduced to the community because cloud computing is not always efficient for data processing when the data is produced at the edge of the network.

Acting as a common interface or middleware for unprecedentedly increasing number of disparate data from variety of IoT devices, edge computing has gained a tremendous impetus by many researchers from the last decade. It not only provide a platform for heterogeneous and unpredictable data from n number of IoT devices while communicating with the cloud but also act as a computational hub which encompasses intelligence for decision making and encounters the bottleneck of latency and power consumption. Performing computation at the end devices had the ramifying effect on the global power consumption by the IoT devices. Now the computational burden of IoT devices has been offloaded to edge servers lying to the vicinity of the IoT devices and hence alleviating the energy conservation. As per the survey conducted by IEA based on 4E Agreement, the standby power consumption by the IoT devices and their respective edges globally is estimated to be 46 TWh by 2025 [8]. Thus recent researches are laying great emphasis on conserving this standby or idle power consumption by IoT devices by enhancing new technological developments. We often encounter the problem of limited storage on edge servers as contradictory to unlimited storage at the cloud. This bottleneck of edge has attracted the users towards the proliferating use of cloud. But this issues of limited storage at edge can be tackled if data approaching edge server is compressed beforehand, hence enhancing the capacity of the edge to accommodate more number of applications. Edge computing incorporates the intelligence, processing power and communication capabilities of an edge gateway or appliance directly into devices [19].

Scale, by definition, refers to "the capability of a system, network, or process to handle a growing amount of work

(service requests), or its potential to be enlarged in order to accommodate that growth" [9]. The Scalability is the phenomenon which means accommodating and servicing the ever increasing network traffic without creating a burden on an edge server. Scalability can be two dimensional. One dimension involves deploying more and more resources to partition the incoming service requests among the multiple servers to handle the requests. This is termed as horizontal scalability. This dimension involves partitioning the tasks based on the number of incoming requests. The second dimension involves partitioning the incoming requests based on action without deploying extra resources. This is termed as vertical partitioning. This paper emphasizes on the first dimension of the scalability.

The rest of the article is organized as follows: Section I briefly discuss about some of the existing scalable based edge IoT systems. In Section II, we describe our proposal of an intelligent edge system in details. In Section III, implementation and simulation of the algorithm is discussed. Finally, the paper is concluded in Section IV.

II. RELATED WORKS

Many researchers have been working on the proliferation of edge computing and have proposed various fog computing layered architectures and paradigms. Shreshth Tuli et al. [20] proposed a framework, known as EdgeLens, adapts itself to the application or user requirements to provide high accuracy or low latency modes of services. They tested the performance of the software in terms of accuracy, response time, jitter, network bandwidth and power consumption and show how EdgeLens adapts to different service requirements. Gusev [10] has suggested the concept of 'Everything as a Service', and has emphasized on the significance of the distribution of smaller servers in front of the central server and in the vicinity of the user, hence promoting the concept of edge computing. Wei Yu [11] has classified various edge computing architectures and compared their influence on the performance of various parameters of IoT networks such as network latency, storage, bandwidth occupation, energy expenditure, and the overhead incurred and various security issues such as availability, integrity, availability and confidentiality.

Jalali [8] has compared various literatures which have been proposed to curtail energy consumption in the IoT devices and has proposed various fog computing techniques which can alleviate power consumption in various IoT devices. Various factors such as access network technologies; idle power consumption of IoT devices, application type, virtualization and network management which lead to higher power consumption has been discussed. Case study on the effect of using Fog computing along with microgrid to reduce energy consumption has been proposed. Fei Li [12] in his work has proposed IoT PaaS architecture promoting vertical scalability by utilizing computing resources and middleware services on the cloud. Domain mediation which is deployed for domain specific control application by the solution providers has been suggested. Two use cases have been put forward in building management domain to show the results.

Sarkar [13] has proposed a layered architecture for distributed environment that utilize scalability which is coupled with cognitive capabilities that promoted intelligent decision making. In his work, an usage control policy model has been proposed to endorse security and privacy in the distributed environment. Sarkar [14] has also proposed a distributed architecture which emphasized interoperability, heterogeneity and scalability, security and privacy of the IoT devices, and named such architecture as Distributed Internet like Architecture (DIAT). In this layered architectures, each layer is classified based on the similar function, forming the hierarchical structure of functionalities.

Onoriode [15] has compared various existing architectural frameworks for integrating various heterogeneous IoT devices and has put forward a viable solution based on micro service. The proposed IoT integration framework takes advantage from an cognitive API layer that makes use of an external service assembler, service auditor, service monitor and service router module to direct service publishing, subscription, decoupling and service amalgamation within the architecture. Ju Ren [16] has proposed a Transparent Computing based IoT architecture to build scalable IoT platforms. With the help of case study, he has build scalable lightweight wearables deploying the proposed architecture.

Kajal [17] has worked for compressing the data especially video frames using Self Organizing Map (SOM) and stored the output feature using Hopfield networks. The frames are preprocessed by making it to pass through the SOM that outputs the helpful features diminishing redundant and irrelevant tenets [17]. Hopfield network is deployed in turn, to store various output patterns. The compressed data can be restored by increasing the dimensions of the frame.

III. LEARNING BASED SCALABLE IOT SYSTEM (LSI)

Taking into account the issue of idle power consumption by IoT devices, a framework has to be proposed that can be deployed to reduce the idle time of edge servers. Also as mentioned above, one of the basic tasks of deploying the edge server is to minimize the latency when addressing several simultaneous requests by IoT devices. Delay sensitive applications are time sensitive thus possess high priority, need to be addressed at first instance. This paper proposes a mechanism that provides intelligence (a learning mechanism) to the edge server by exploiting which the type of application can be classified whether delay sensitive or delay insensitive. This paper also addresses the issue of scalability when edge server gets flooded with delay sensitive applications taking into account horizontal scalability.

Objects (e.g. sensors) are provided with IP address, which aid communication with each other. Various sensors send numerous and variety of incompatible data to the edge for processing. In this paper, the proposed system employ Self organizing map for classifying the delay sensitive data and the delay insensitive data. The delay sensitive data is immediately forwarded to the edge server for further processing. The SOM not only clusters the type of data but also compresses the incoming data, hence enabling the edge server to accommodate more data as edge is limited by its storage capacity, and therefore data compression is of vital importance to edge computing. Whenever the edge gets enormous delay sensitive data, its throughput declines, therefore this paper laid

emphasis on providing additional edge servers, which at sleep initially. It activates only when data overflow at the server 1. This additional server can also be deployed to handle delay sensitive data. This server comes to play only at high peak times, but at normal times, it is just set to sleep. This can be accomplishes through the use of Aneka Auto-scaling provision. Through this mechanism load balancing has been achieved without much affecting the energy consumption rate. The comparison of LSI with the already existing researches and algorithms has been listed in the Table I.

TABLE I. COMPARISON OF VARIOUS IoT ARCHITECTURES

IoT Features	DIAT [14]	Transparent Computing Architecture [16]	IoT PAAS [12]	LSI
Horizontal Scalability	Yes	Yes	Not Known	Yes
Vertical Scalability	No	No	Yes	No
Energy Consumption	Not Known	Yes	Not Known	Yes
Response Delay	Not Known	Yes	Not Known	Yes
Context Aware Service Support	Yes	Yes	Not Known	Yes
Intelligence	Yes	Not Known	Not Known	Yes

A. Self Organizing Map (SOM)

With the purpose of data visualization which in turn endorse in realizing high dimensional data by short sizing the dimensions of the data to a map, SOM is of crucial weight and has grabbed the attention of many scholars and researchers. SOM deploys competitive learning paradigm to cluster data by amassing similar data altogether. Thus SOM reduces data dimensions and displays similarities among data. With the aid of SOM, clustering is accomplished through having several units compete or race for the current element. Once the SOM is provided with the data, the network of artificial neurons goes through competitive training. The weight vector of the unit that is found nearest to the current element is proclaimed to be the winning or active unit. While undergoing training, SOM preserves the neighborhood relationship that persists among the input data sets. As it approaches near to vicinity of the input object, the weights of the winning unit are put to adjustment as well as its neighbor's nodes.

In contradiction to other learning technique, SOM do not employ target vector. A SOM learns to classify the training data with no external interference. Normally, Euclidean distance is the most rapidly used method for determining the distances of the input vector from the weight vector. The input vector whose distance is shortest from the weight vector is proclaimed to be the winner. SOM algorithm can be summarized as follows:

1) Initial steps involve the weights to be provided with arbitrary weights.

2) Next we choose any random input vector.

3) We examine each node by evaluating which node's weights are found to be closest to the input vector. We call this winning node as Best matching node.

4) Next we find out the neighbors of the winning node.

5) The winning weight is rewarded with becoming more like the sample vector. The neighbors tend to be more like the sample vector. The closer a node is to the winner node, the more its weights adjusted and the farther away the neighbor is from the winning node, the less it learns.

6) Go to step 2 for next samples.

B. LSI Working

Raw data and/or information in the form of signals are amassed based on whatever is sensed by the Sensors and actuators. The most clear-cut demonstration of a time series signal is based on its time-domain form, and then distances between time series relate to differences between the time-ordered measurements themselves. A traditional approach to the classification of time series problem used in data mining domain, focus on classifying short time series which encode meaningful patterns. This is termed to as instance based classification. Here new time series are classified by matching them to similar instances of time series with a known classification. These incompatible data has to be preprocessed and normalized before being used as input for any further action. Collected sensed data is forwarded to the edge device, where it first encounters Self organizing map (SOM) where the learning is basically relied upon only the input data and such unlabelled data is independent of the desired output. We call such learning mechanism as unsupervised learning. Since SOM tend to respond to several output categories after performing training. Hence an additional mechanism is added, so that SOM can respond to which one unit will respond (this is termed as Competition, often called Winner take all strategy). The number of output unit has been limited to two, either cluster into delay insensitive data or to delay sensitive data. During training, SOM identifies which output unit best matches the input sensed data. The sensed data is preprocessed and input to the SOM situated at the edge. Kohonen feature map is created, which determine the winner unit (whether delay insensitive or delay sensitive data unit). Kohonen feature map can also be trained to compress the sensed input data, so that more number of data can be accommodated at the edge, hence compensating the limited size of the edge server to a limit. Based on the winner unit, data is forwarded to either Queue 1 or Queue 2. Queue 1 is maintained for delay insensitive data, whereas Queue 2 is maintained for delay sensitive data. Kohonen network will undergo two phases: training and testing. During the training phase, the SOM is trained to form the cluster of inputs based on the similarity (or minimum Euclidean distance). During the test phase, we evaluate the performance by measuring to what extent our network classifies delay sensitive and delay insensitive data.

Data stored in the Queue 2 is basically stored for processing at the edge device. A variable counter is maintained on the number of the requests arriving at a time and being served by the edge server. If the counter exceeds 30, then additional edge is initialized and any further request is forwarded to this edge 2 provided Aneka auto-scaling.

Queue 2 is provided with a threshold (T) on its size. Whenever overflow occurs i.e. Queue 2 reaches its threshold (T), an alternative Queue 3 is maintained which will store the further data. With the instigation of Queue 3, an additional edge which is maintained in order to handle the traffic at the edge device 1, is also provoked. This edge server 2 will process the further data arriving after the overflow at the Queue 2. Initially, additional edge device is at sleep, and provoked only after the initialization of the Queue 3. After processing the data at Queue 3, the edge server 2 can be disabled by the Aneka and again goes to sleep. This is done to curtail the extra energy consumption, which would have been expended due to the idle state of the edge server 2 during its free time. Thus, this feature ensures horizontal scalability without much energy expenditure.

Queue 1 is maintained for the delay insensitive data, hence data in the Queue 1 can either be directly forwarded to the cloud for processing, or additionally we provide a mechanism called cycle steal. Sometimes, while processing delay sensitive data at edge server 1 and edge server 2, the processor may be found in an idle state. During such time, data in Queue 1 or training at the SOM can be processed at these edge servers. This mechanism not only enhances CPU throughput, but also saves energy consumption due to idle state at the edge servers.

C. Inputs

This proposed research encompasses the following application areas: Hospital, Surveillance, Organization, Inventory, and Smart Home. Each of these applications is facilitated with the following types of Sensor: Smoke, Temperature, Proximity, Thumb, and Smoke. The utility of each sensor varies according to the application areas, for example the data sensed by smoke sensors is of highest priority and has to be processed immediately; the thumb sensors finds importance for the Organization but is of less importance if found in the Inventory; the temperature sensor is critical for the applications like Hospital and Inventory, etc. It is proposed to process the data sensed by the higher priority sensor (Priority vary according to the application area) in the edge, forwarding the rest data to the cloud. The inputs are basically signals which can be represented in time series.

D. LSI Algorithm

The LSI algorithm is explained in detail as follows:

Here T (threshold) = 30   /*Threshold T is set to be 30 keeping in view the processing capacity of the edge & IoT sensors are low power devices and require less processing power and an Edge can easily process at least 30 requests at a time without causing Overloading. */

If (Sensed_Data == True)

{

Initialize SOM;

SOM← Sensed_Data;   /*SOM evaluates the Euclidean distance of the signals with the already stored ones.*/

If Output(SOM)==1

    Queue1←Sensed_Data;   //Send data to the Cloud.

If Output(SOM)==2

    Counter++;

    If (Counter<=T) /*Check whether counter value <=Threshold */

        Queue2←Sensed_Data;   /*Send data to the Edge Processor 1 */

    else

        Queue3←Sensed_Data;   /* Send data to the Edge Processor 2 */

If Processing(Sensed_Data)==True /*If the processing of the request is completed.*/

    Counter--;

If (Queue3==Empty)//Check if no task at Edge processor 3

    Dissolve(Queue3);   /*Disable the Edge processor 3 if under load.*/

}

This particular system and algorithm can be well illustrated in Fig. 1.

E. Multi-Application use Case

To validate architecture, a composite use case is described in this section. The instances consist of various IoT applications like smart city, smart transportation, smart home, smart healthcare, smart retail and supply chain system, smart biometric system, smart agriculture, smart logistics, smart fitness system etc. The proposed LSI architecture overcomes the barrier of separate heterogeneous IoT applications, and thus achieves a global IoT system. When multiple such sensed data is collected by the sensors and sent to the LSI, the SOM classifies real time data such as smart healthcare, smart transportation, etc. and non-real time data such as smart biometric, logistics, fitness, etc., and pass them either to the edge or to the cloud respectively for further processing.
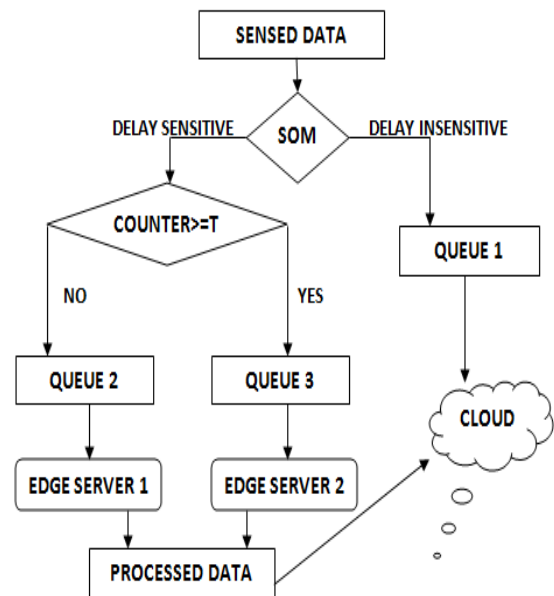

Fig. 1.   LSI Flowchart.

IV. SIMULATION AND RESULT ANALYSIS

This research is simulated in the Aneka platform, which is a software platform for developing cloud computing applications [18]. It is a Pure PaaS solution for cloud computing. It is a framework which provides both middleware for managing and scaling distributed applications and an extensible set of APIs for developing them. Aneka 5.0 can be installed on the Virtual machine which may be configured as per the user requirements and agreed SLA. Aneka is interfaced with Matlab 8.0 which contains the code for implementing the scheduling of the delay sensitive and delay sensitive data. An Aneka server was developed locally with a single master and multiple workers, which create the environment of an edge, and have installed Aneka on the virtual machine created on the AWS cloud. Hence, creating edge-cloud environment. Aneka 5.0 has the provision of Autoscaling, which varies according to our algorithm.

In order to create the edge environment, a master and two workers have been created. The algorithm for SOM is run on the Aneka master computer which installed with Matlab 8.0 in which algorithm for classification implemented, which decides whether to forward the data to its worker computers or to the cloud. Initially only single worker computer is in the processing state, while the other inactive. Whenever the number of requests crosses 30 (the threshold T) to be served at the edge, the Aneka transfers the workload to its other worker along with its first worker. Here in this scenario, only two workers have been created, but many more worker computers can be created for processing the unprecedentedly increasing incoming requests. This research has been performed at a small level. Alternatively a virtual machine has been created with the Aneka PaaS installed in it on the AWS cloud. The classified delay insensitive applications are forwarded to this particular cloud. This cloud is also provisioned with the Autoscaling of VMs as the number of requests gets flooded. But in our research, any such requests flood at the cloud was not witnessed, since performed at the small level. When the workload at the edge gets low, the edge master automatically disables its other worker computer. The master does not perform any task on its own except classification and load distribution.

Initially, the research is simulated by taking number of incoming requests as: 5, 15, 25, 35, 45, 55, 65, and 75. Firstly these entire requests were forwarded to the cloud and their performance was evaluated. Then these requests were processed on the LSI system as proposed above. Then comparison on the following parameters were made and the proposed algorithm is found to prone to succeed not only in achieving the primary characteristics of IoT like heterogeneity, scalability, etc. but also attained the following metrics:

1) Throughput: The amount of service requests processed by the device per unit time. Since LSI incorporates the feature of scalability by providing the facility of additional edge server (Aneka worker), the network throughput is prone to enhance tremendously as multiple requests are being processed simultaneously at edge worker 1, another additional edge worker, and the cloud. Thus number of requests processed per unit time improves.

The average number of requests per second was found to be 63 in case of LSI, and 48 in case of Cloud.

2) Response time/delay: The duration of time required taking delivery of a response to a request. It can also be seen as the average time the client has to wait to get its job done. Delay sensitive applications are to be responded immediately without any delay in processing. By setting the priority for the delay sensitive data, LSI improves the response time tremendously. The observations are manifested in Fig. 2.

3) Packet loss: It is termed as the number of packet drops due to traffic or network congestion during a specified duration of time. LSI reduces the number of packet loss since it endorses the factor of horizontal scalability by deploying the provision of additional edge device whenever edge device 1 exceeds the threshold. Thus LSI is least prone to network congestion and improves the packet delivery factor for the measured duration of time. The use of mirror server aid in processing the majority of the requests at the edge server level, and not at the cloud level, therefore the packets are least faced with the network congestions. Hence reducing the probability of packet loss as shown in Fig. 3.
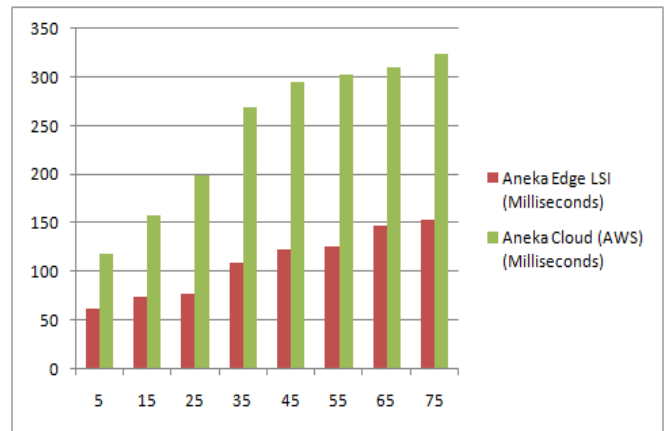

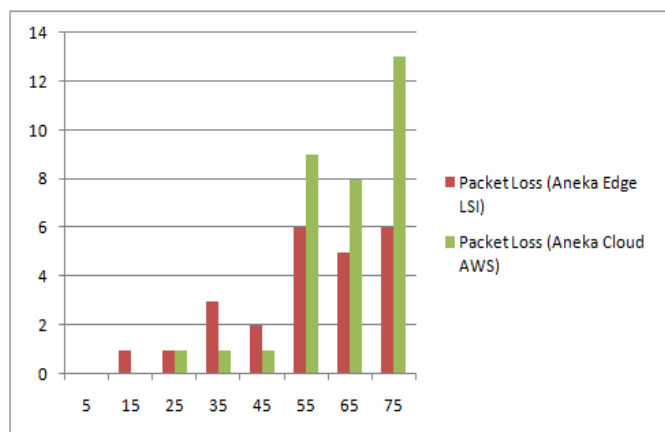Fig. 2. Comparison of Response Delays wrt Cloud and LSI.

Fig. 3. Comparison of Packet Loss wrt LSI and Cloud.

V. CONCLUSION AND FUTURE WORK

As can be clearly manifested by the current scenario, the unprecedented growth of IoT devices has generated the urgency of deploying such edge devices which are scalable and can accommodate the varying and consistently increasing number and types of IoT devices. Whether accurate or not, we must prepare for it or risk becoming the victims of our own success. In the paper an intelligent IOT framework has been proposed where in Self Organizing Map is deployed, in order to differentiate between delay sensitive data and delay insensitive data. This leads to the improvement in response time of the critical applications such as related to healthcare which to be immediately addressed, promoted scalability by utilizing the additional edge device, in turn improving the throughput. It reduced the idle time of the edge server 2 by promoting disabling phenomenon supported by Aneka Auto-scaling provision, which in turn reduced energy expenditure that may have incurred due to the idle state of the edge processor 2. As compared to the existing work, LSI not only reduces latency but also work in enhancing resource utilization, throughput and reduction in packet loss as manifested by our results. This research is done at small scale, but can be extended to incorporate various other applications. This research improves the performance to an extent as compared to the cloud because classification of delay sensitive and delay insensitive data is accomplished on the pattern of the data itself (time series), which may generate inaccurate results sometimes (i.e. may not classify data correctly) due to overlapping range of values of data of various sensors, but it shown good results specially in terms of response delay, because we are bifurcating the sensed data to the edge as well as cloud, as compared to performance when processing done at cloud only. Furthermore pre-processing can be deployed on the data to give better classification results. Also, scalability can be achieved at higher level by deploying several more additional edge servers. Further, in order to increase the scalability and to have better performance other algorithms like deep learning or machine learning can also be applied in future.

REFERENCES

[1] K. Ashton, "That Internet of Things thing," RFiD J., vol. 22, no. 7, pp. 97–114, 2009.

[2] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelfflé, "Vision and challenges for realising the Internet of things," vol. 20, no. 10, 2010.

[3] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Gener. Comput. Syst., vol. 29, no. 7, pp. 1645–1660, 2013.

[4] E. Cuervo et al., "MAUI: Making smartphones last longer with code offload," in Proc. 8th Int. Conf. Mobile Syst. Appl. Services, San Francisco, CA, USA, 2010, pp. 49–62.

[5] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: Research problems in data center networks," ACM SIGCOMM Comput. Commun. Rev., vol. 39, no. 1, pp. 68–73, 2008.

[6] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," IEEE Pervasive Comput., vol. 8, no. 4, pp. 14–23, Oct./Dec. 2009.

[7] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of things," in Proc. 1st Edition MCC Workshop Mobile Cloud Comput., Helsinki, Finland, 2012, pp. 13–16.

[8] Fatemeh Jalali ; Safieh Khodadustan, et.al, "Greening IoT with Fog: A Survey", IEEE International Conference on Edge Computing (EDGE), IEEE, 11 September 2017.

[9] https://en.wikipedia.org/wiki/Scalability.

[10] Marjan Gusev, Schahram Dustdar, "Going Back to the Roots—the Evolution of Edge Computing, an IoT Perspective" IEEE Computer Society, March/April 2018.

[11] Wei Yu, Fan Liang, Xiaofei He, et.al, "A Survey on the Edge Computing for the Internet of Things" IEEE Access, Volume 6, 9 March 2018.

[12] Fei Li, Michael Vogler, et.al, "Efficient and scalable IoT service delivery on Cloud" , 2013 IEEE Sixth International Conference on Cloud Computing.

[13] Chayan Sarkar, Akshay Uttama Nambi S. N., et.al, "A Scalable Distributed Architecture Towards Unifying IoT Applications", 2014 IEEE World Forum on Internet of Things (WF-IoT), March-2014.

[14] Chayan Sarkar, et.al, "DIAT: A Scalable Distributed Architecture for IoT" IEEE INTERNET OF THINGS JOURNAL, VOL. X, NO. X.

[15] Onoriode Uviase, et.al, "IoT Architectural Framework: Connection and Integration Framework for IoT Systems", EPTCS 264, 2018, arXiv preprint arXiv:1803.04780.

[16] Ju Ren, Hul Guo, et.al, " Serving at the edge: A Scalable IoT Architecture Based on Transparent Computing", IEEE Networks, September/October 2017.

[17] Dr. Manu Pratap Singh, Kajal Sharma, "Video Compression using Self Organizing Map and pattern storage using Hopfield Neural Network", International Conference on Industrial and Information Systems (ICIIS), IEEE, 11 March 2010.

[18] Vecchiola, Christian, Xingchen Chu, and Rajkumar Buyya. "Aneka: a software platform for .NET-based cloud computing." High Speed and Large Scale Scientific Computing 18 (2009): 267-295.

[19] Shreshth Tuli, Nipam Basumatary, Rajkumar Buyya.

[20] EdgeLens: Deep Learning based Object Detection in Integrated IoT, Fog and Cloud Computing Environments" 4th IEEE ISCON (November 21-22, 2019).

Scalable Asymmetric Security Mechanism for Internet of Things

Ayesha Siddiqa[1], Sohail Ahmed[2]
Department of Computer Science and Information
University of Lahore, Gujrat Campus
Pakistan

Abstract—**The Internet of things stances rigorous demands on excellence of quality and the vitality of security. It becomes vital to provide an extremely reliable encryption algorithm with less complexity and computational expense in IoT paradigm. Most of the protocols designed in past for communication between sender and the receiver based on asymmetric cryptography algorithms poses high computational cost. Therefore, this paper presents a less complex and more secure and fast encryption algorithm for communication between devices i.e. Asymmetric Scalable Security between sender and the receiver of the information. We present a reliable, secure, scalable and efficient communication protocol that used asymmetric algorithm for securing the exchange of information between sender and the receiver. The proposed communication protocol is lightweight encryption method that does not require complex resources to perform the computations involved for using the asymmetric cryptography. The simulation results also show that the proposed method is efficient in terms of time and space and ensures confidentiality. Therefore, the proposed scheme is beneficial for providing the secure communication for the power and resource constrained IoT devices.**

Keywords—Asymmetric cryptography; confidentiality; internet of things; security

I. INTRODUCTION

The Internet of things is one of the most demanding technology developments field in this era of digital world. The devices that comprises of sensors and actuators have sufficient functionality of supporting proficiencies of networking and other processing abilities. So these abilities make them communicate over the internet, to communicate with each other and provide services over the Internet [1]. The advent of big data analysis has brought tremendous advantage to the creation of a smart society so this realization poses many challenges such as getting authenticating on the network, encrypting the information shared between power and resource constrained devices [3,22]. These challenges require an efficient and lightweight communication mechanism, which is scalable and lightweight too for constrained devices [4, 25]. The IoT has a lot of diverse and heterogeneous devices and multiple purpose technologies that are manufactured and distributed by the different vendors so these devices may vary in their proficiencies. We can say that the Internet of things is comprised of different types of sensors and objects, which are called constrained devices; these are called constrained devices because they have constrained resources in terms of memory, power, processing power, communication and the user interfaces. They have very low power and very little memory. When these devices are used on the network; they also put constraints on the network as well; so the networks may expose to the large number of packet loses, less throughput rate and many types of advanced security facets. Hence, the first and foremost challenge in the world of IoT is to adapt to capabilities of these networks along with their integration with the traditional Internet Communication standards.

Conventionally the encryption algorithms comprises of symmetric and asymmetric key algorithms and most of the encryption algorithms which have been proposed in past for IoT devices involve symmetric key algorithms. The symmetric cryptography is being used in IoT devices as we have low powered and low memory equipment and same key is shared between sender and the receiver of the information. However, in recent years many algorithms have been proposed for secure communication, which is the target of this study [23].

Security parameters such as authentication, confidentiality, and integrity and access control are addressed by the different researchers and proposed many possible solutions to ensure security. The work of researchers includes the use of IP security i.e. IPSec, Internet Key Exchange protocols, using Transport layer security, using Datagram Transport Layer Security (DTLS) and key bootstrapping [20,28]. Although, these solutions have been proposed in an efficient way to ensure security in IoT core devices; but these solutions cannot be applied in the IoT constrained devices directly due to their limited computation capabilities and resources [32].

In this paper, we have presented an efficient and secure mechanism for communication in an IoT domain using the Asymmetric cryptography. We have used RSA as the PKC algorithm to secure the communication between sender and the receiver. The RSA is a block cipher algorithm [6] so it can be easily used with the IoT devices. The rest of the article is arranged as follows. Section II discusses the conceptual framework that includes the basics of public key cryptography and RSA algorithm. Section III describes the related work. Section IV elaborates the framework and defines the problem statement. Section V elaborates the proposed methodology and implementation details. Section VI explains the analysis and results of the proposed methodology while in Section VII we conclude the article.
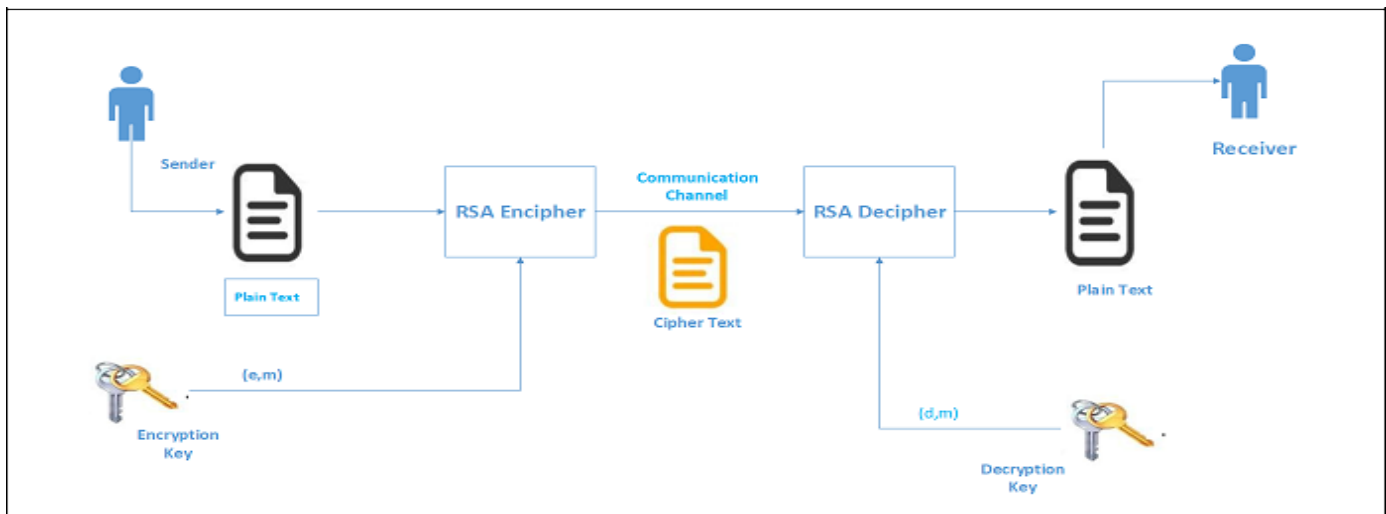
Fig. 1. Scenario of Public Key Cryptography.

II. CONCEPTUAL FRAMEWORK

A. Public Key Cryptography

Public Key Cryptography is used for encryption. Fig. 1 explains the scenario of PKC; different keys are used for encryption and decryption. Hence, PKC is considered to be more secure as compared to symmetric cryptography. PKC is used in IoT domains to transport secret keys which are encrypted using asymmetric cryptography because in this case smaller blocks are encrypted and the process of encryption is needed only once[6]. These kinds of algorithms are mainly based on the RSA, ECC and Diffie Hellman (DH) for key exchange and negotiation purposes. The RSA algorithm relies on the rigid mathematical problems of prime factorization while ECC algorithm is based on the elliptic curve discrete logarithm problem and DH (Diffie Hellman) security is based on the discrete logarithm problem [33]. The Diffie Hellman is a key exchange algorithm but it suffers from the Man-in-the-Middle attack. However, the RSA algorithm provides highest security but it is computationally extensive algorithm. The ECC algorithm is widely used for the shortest length of key as compared to the RSA algorithms but these both algorithms provide the equal strength of security. These algorithms are used widely in IoT devices for the authentication purposes.

B. Basics of RSA Algorithm

RSA algorithm was named after the names of the designers who have proposed it i.e. Rivest, Shamir and Andleman so it is named as RSA algorithm. It was initially proposed in 1978 [16]. It is one of the most powerful public key algorithms used. It has been widely used in public key cryptography and PKI (Public Key Infrastructure) [8]. The mathematical base makes it most suitable to be used for the Public key cryptography. It is also used in certificate mode of

security. Its mathematical background is the theorem of Euler and it relies on the integer prime factorization i.e. IFP. The whole algorithm bases on the selection of the prime numbers that are generated randomly and this is the strength of the algorithm.

In order to achieve the end-to-end security a protocol named MIKEY in [7,29] is used. It supports multiple modes of security such pre shared keys, public keys and key exchanges. It was specifically designed for real time and multimedia applications but this can also be used with the constrained devices such as sensors and actuators as it has attributes similar to the attributes of constrained devices. The Public Key Cryptography schemes provide high scalability and more resilience to attacks, which may raise the energy requirements and may involve complex computations [24, 26, 31]. However, they are more suitable when security is the major concern. In this paper, we present a secure way of communication using Asymmetric Cryptography.

The whole algorithm of RSA can be decomposed into three steps that are as follows:

1) Key Generation
2) Encryption
3) Decryption

In RSA algorithm the most computationally extensive part is the generation of key pairs because here the large prime numbers are selected such that they are not easily guessed and then the product of two large prime numbers is computed which will also a very large number as explained in Fig. 2. In fact two large prime numbers are selected and the minimum length of the modulus N is 1024 bits which is the minimum bit length in case of RSA algorithm for providing security [8].
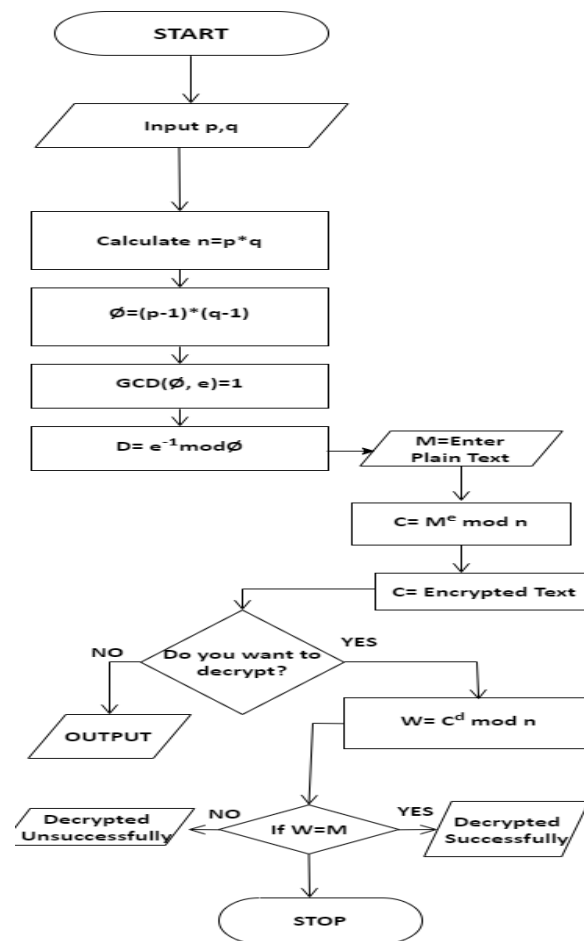
Fig. 2. The Structure of RSA Algorithm.

III. LITERATURE REVIEW

The Symmetric key algorithms incur less overhead as pre-shared keys are used while the public key based cryptographic solutions are more scalable and more robust to assure key distribution to the masses of devices.

CoAP (Constrained Application Protocol) has been introduced by the IETF's (Internet Engineering Task Force) the core working group to ensure the unified transmissions [21,27]. It has been designed specifically to address the issues of low power and low memory devices along with the support of multicast messages and abridged consumption of energy [19]. CoAP uses the UDP as the transport layer protocol so the reliable transmission of packets.

To ensure the reliable transmission DTLS has been introduced to implement the security such as the TLS (Transport Layer Security) we have in the traditional networks. The purpose of using the DTLS is to provide the end-to-end security at the transport layer. However, the security of DTLS is not incorporated with the application layer protocol such as CoAP and it also does not support the message-oriented approach of security so there is a need of object model or message oriented approach at the application layer. An alternative approach is to incorporate security in the CoAP using an added security option [9,30]. However, the current specifications of CoAP shows three different security modes which are used with DTLS.

1) Pre-Shared Mode PSK: In this mode of security, the devices are per-shared with the symmetric keys, which can be used to secure the communication between devices. The keys are pre-programmed with the keys so this mode of security is useful for devices that cannot support the public key cryptography algorithms.

2) Raw Public keys (RPK): In this mode of security, the devices can use PKC but they are not the PKI (public Key Infrastructure). The devices are pre-configured with asymmetric keys, which are authenticated using out of band validation. The devices can obtain the identity from the Public key and the devices also contain a list of IDs and a list of nodes it has to communicate with.

3) Certificates: In this mode of security, certificates are used for authentication called PKI (public key infrastructure) so a kind of security infrastructure should be available which is still a challenge. The certificates used for binding the security and are also signed by the common trusted store so the device also holds a list of trusted stores which can be used for the validation of certificates.

An ECC-based signcryption method was introduced in "Henriques and Vernekar [5]", the public values were bound with the public keys so this is also a certificate based approach and it relies on the binding of certificates from a server which is known as the trusted server. In this approach the computational overhead of using certificates is solved but the use of public key encryption is stagnant challenge.

In "Chen [10]" the author has proposed a scheme called symmetric security with symmetric cryptography particularly for constrained devices. It uses the symmetric key cryptography because it is lightweight for IoT devices and the author has made it scalable for a variety of devices. This scheme uses a trust Anchor that is responsible for providing and establishing the connection between client and server. This approach uses the Trust Anchor to establish the connection using symmetric key cryptography using PSK mode only.

In "Rescorla and N. Modadugu [11]" a key agreement approach is used to establish the authentication scheme for IoT and the cloud and authentication is carried out using HTTP cookies. This is ECC based method but it is used to achieve the symmetric key session. ECC algorithm is used to perform the registration phase followed by the login and authentication phase. This approach specifically provides the authentication between cloud servers and embedded devices.

In "Chavan and Nighot [12]", the author has proposed an approach to use and implement RSA cryptography for sensor nodes in smart cities. The author has proposed an efficient way to implement RSA in sensor nodes using Montgomery multiplication instead of using Chinese Remainder Theorem for implementing RSA. This approach address the security issues in sensor nodes but does not specify the way of communication and performing encryption and decryption using RSA. Moreover, it does not interact with the application layer protocol and hence it is also not scalable to a large number of IoT devices. Therefore, this cannot be used security measure in IoT environments; instead, it implements the RSA in hardware; it cannot be used in client/server communication that is the bases for IoT environment.

A. Hardware based Approaches

In "Granjal, Monteiro, and Silva [13]", the authors have proposed a model for establishing authentication key-scheme which is signature based for IoT applications. It is a complex approach as the system of establishing secure signature based keys consists of the eight phases and all the phases involve some kind of mathematical computations, which is not suitable for IoT environment. Moreover, it is designed specifically for the future IoT devices, which may have more power and memory as compared to the devices used today so this scheme is not applicable to the current scenario of IoT application. Furthermore, it also does not specify that how communication between clients and resource servers will take place after being authenticated by using the described scheme of securing IoT applications using signature-based scheme for key establishment because this scheme is proposed for providing the authentication purposes "Zhou, Liu, Tang, and Tinashe [18]".

Many approaches have been proposed so far to accelerate the hardware of IoT devices such as AES. Although these kinds of methods seem to ensure the security requirements of IoT environment but they do require devices with more memory and power so these solutions are not suitable for IoT devices that are used and deployed in the current era; we may be capable of producing and manufacturing such devices in future which may have more memory and power as compared to the devices used today "Malik, Dutta, and Granjal [2]".

In "Raza, Seitz, Sitenkov, and Selander [14]", the author has proposed a model to be used with the distributed applications of IoT. This method provides authentication based mechanism which involves many phase just to initiate the communication between client and the server. This approach consists of the registration phase followed by the login phase and the key agreement phase.

The method proposed in "Chang, Wu, and Sun [15]" provides the security measures which consists of a lot of complex computation which involves the use of RSA as well as ECC along with some hash functions and some kind of MAC also. Although this method has introduced intense security measures, but due to its complexity, it is not applicable to all types of devices in the IoT paradigm.

B. Public Key Infrastructure

The research on PKI is mostly engrossed on the compression of protocol headers such as [16, 17]. The author has proposed the deployments of PKI at the DTLS layer. The author has analyzed the headers of DTLS and established that the size of DTLS header is too long and cannot fit into single packet of IEEE 802.15.4 for providing the end-to-end security. Therefore, the author proposed a scheme of compressing the header of 6LoWPAN for DTLS and they have further claimed to reduce the number of bits to 75% in the DTLS handshake header. However, the DTLS handshake is providing the automatic key management at the transport layer, providing authentication to the server and the client using and claimed to ensure end-to-end security. The author has not assessed the PKI and even not elaborated that how public keys will be transported.

All of these schemes are addressing the security issues in IoT domain as well as providing the solutions. The Hardware based approaches are efficient to provide essential security measures but they can also be exposed and do not provide robust security. PKI based approaches use DTLS header to provide secure communication but the approaches do not elaborate that how keys will be transported. Hence, studying all these proposed models we draw conclusion that we need a solution of providing security through a mechanism that should secure communication using dynamic keys assignment.

IV. PROPOSED METHODOLOGY

In this section, we propose a lightweight model for communication between sender and the receiver in a constrained devices environment. The proposed model uses asymmetric cryptography to secure the communication between CoAP client and the server; it includes the Trust Manager TM for the generation of asymmetric keys dynamically. The proposed model is inspired by the S3k and

includes four phases i.e. a) Key Generation phase that is performed by the Trust Manager TM, b) request sent by the client after getting the keys from the TM, c) Resource Server RS servicing the request of client and finally d) receiving the data from RS in encrypted form and performing decryption operation at the client side. The notations used in the proposed algorithm are listed in the Table I.

TABLE I.   NOTATIONS

P	Large prime numbers of bit length
Q	Large prime number of bit length
N	Product of prime numbers p and p
E	Smallest integer in the range 1<e<n
Φ(n)	Product of p-1 and q-1
D	Exponent of private key
n,e	Pair of public key
d, n	Pair of private key
P	Plain text
C	Cipher text
Mod	Modulus

A. Terminologies used in the Proposed Methodology

1) Client: Client is a machine that wants to communicate with things i.e. sensors and actuators. The user is a client but the user will communicate using a machine that can be a laptop, smartphone or desktop computer. The user can connect to the Resource Server using one of the above-mentioned devices.

2) Trust Manager: Trust Manager is a kind of application that is obliged for generating the key pair for the client. The Trust Manager will generate keys i.e. private and public keys for client and the client will use these keys for the communication purposes. Only Client can connect to the Trust Manager and the Resource is kept apart from the Trust Manager to avoid the communication overhead from the RS.

3) Resource Server: It is a kind of Raspberry pi device, which is able to communicate with the different types of sensors. It has enough power and memory resources, which make it suitable for communicating with the different types of sensors efficiently. This device has different types of specifications according to the model of the Raspberry pi [22].

a) Phase I: Key Generation: The Client sends request to the TM for connection with the Resource Server RS; the TM generates the keys that are used for the communication purpose and sends the public and the private keys to the Client machine as shown in the Fig. 3. The pseudocode for the key generation process is as follows:

1) Generate a random number r using secure Random

2) Generate p and q two big prime numbers according to the bit length

3) Calculate n and Φ(n) and select an integer e

4) Compute d such that d= e.modInverse(Φ(n)).

Fig. 3.   Phase I. Key Generation Request.

b) Phase II- Request Sent By Client: In this phase the client machine that can be desktop computer, laptop or smart phone generates the request.

1) Public key = n, e

2) Private Key= n, d

3) n,e$\rightarrow$radix(16)…………………..(1)

4) Append (n, e)

5) Setting payload

Key Generation Algorithm is explained in the flow chart represented in Fig. 4.

c) Phase III- Request Processed By Rs

1) GetRequest (n+e)

2) Parse Request to get n and e

3) n+e $\rightarrow$ radix(10)

4) receive status s from sensor node

5) response.s^e mod n……………….(2)

6) Response back to Client.

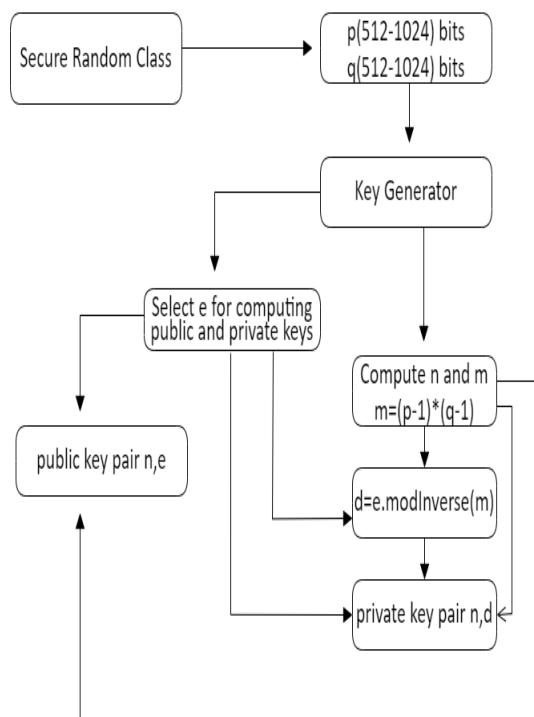Fig. 5 shows the computations performed for processing the request at the client side as well as the RS side.
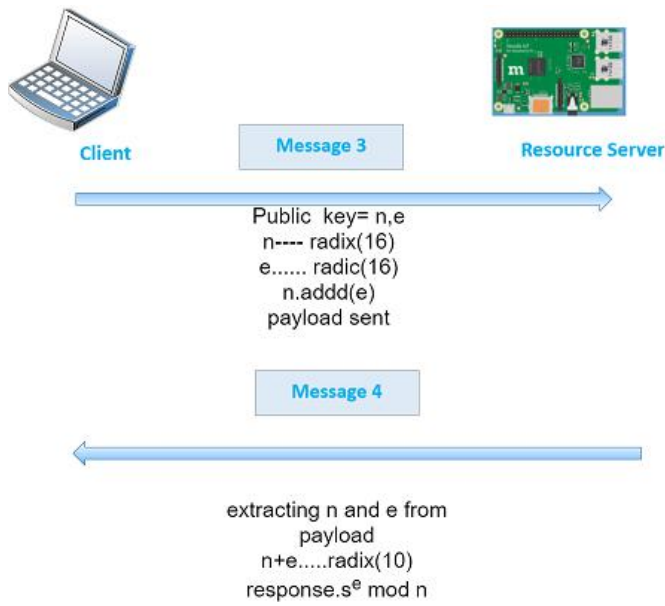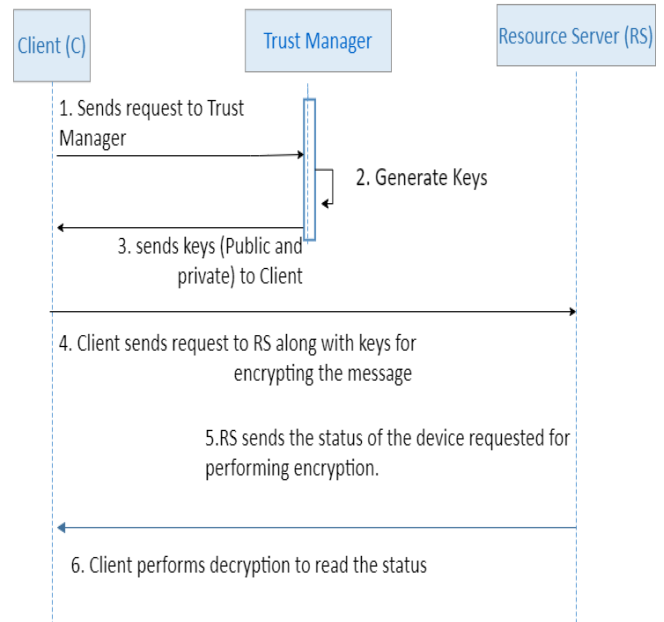

Fig. 4.   Key Generation Algorithm.

Message 3

Public key= n,e
n---- radix(16)
e...... radic(16)
n.addd(e)
payload sent

Message 4

extracting n and e from
payload
n+e.....radix(10)
response.s^e mod n

Fig. 5. Request Process Phase.

d) Phase Iv- Decryption At The Client Side: The response received is decrypted at the client side to extract the information hidden. The whole communication sent and received by the Client and the RS to compress the bits and to make it more secure to send over the channel. The pattern of flights between RS and the Client showed in the figures below.

The pattern of flights is shown between client and the RS when the client machine initiates the communication and the user from the client machine can access the devices attached to the Resource Server remotely by using the TM in a secure way. However, this solution presented here is one sided. As we know that, the RS is a kind of Raspberry Pi device and many sensors connected to the RS. The RS continuously gets the latest, updated command and takes suitable action. The above scenario applies to the critical situation where the device status matters a lot and immediately in a hospital, where intense care patient need immediate response from the doctor, so in this case, the RS can also initiate the communication and inform the client about the latest condition.

The RS would connect to the TM and request to generate the keys. The TM will generate the keys and send public and private key pair to the Resource Server. The RS sends the status of the device along with the public keys as depicted in Fig. 6.

The Client on the other end receives the status, issues the command accordingly after encrypting it with the public keys, and sends the encrypted message to the RS. The RS decrypts the message received and acts accordingly. The proposed scenario of the communication between RS and the client is explained in Fig. 7.
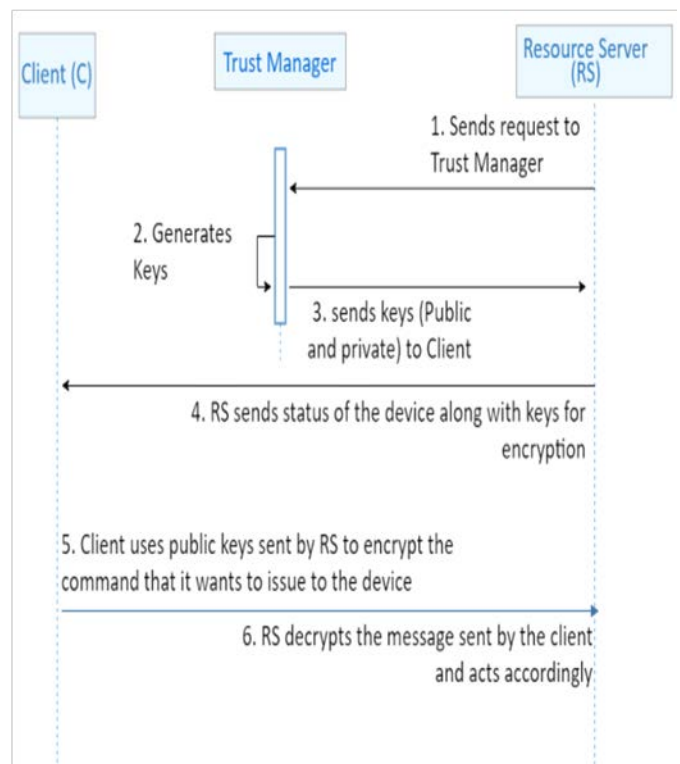

Fig. 6. Communication between Client and Resource Server.

Fig. 7. Communication between Resource Server and Client.

V. ANALYSIS OF RESULTS AND DISCUSSION

In the following section, we discuss about the implementation detail of our work, the results obtained and the analysis of results. We have simulated our proposed framework using Californium library, which is available as open source for implementing in Eclipse using Java language. We have setup the environment for the testing purpose of our methodology using the following experimental setup. We changed the data sets against different approaches, which are being used for ensuring secure communication in IoT domain. We carried out equal number of tests for each approach and obtained the results using the same machine and other specifications being same as well as described in Table II. Results are attained for time required to perform the encryption and the decryption operation as well as using the more secure to communicate.

A. Time Consumption

In this section, we have evaluated the results to compute the time utilized by our proposed framework to encrypt the data that is being sent and the decrypt the received data at the other end. We have recorded the time as when the request payload is formed and the keys for encryption are also sent along with the request; the RS receives the request, encrypt the status and sends it to the client and the client performs the decryption operation to check the status of the device. However, we assume that the RS is continuously receiving the statuses of the devices it has been attached with. We have obtained the results and then compare them with the well-known symmetric algorithms i.e. AES and the DES as they have been implemented in past. Our purpose was to reduce the time used to perform encryption and decryption along with enhancing the security. The results are shown in the following figures.

The results show that the proposed framework takes less time as compared to AES so we can say that the proposed methodology is efficient in terms of time consumption. Moreover, we recorded the time by changing the packet size and the size of the key as well, but the results still show that our proposed work is more efficient. We have also evaluated the results against DES algorithm because DES is also a well-known symmetric key algorithm. Table IV shows the observations recorded for DES and the proposed framework.

Table III shows the comparison between our proposed framework and the AES scheme.

The data set was remained constant for recording the time for AES and the proposed framework and obtained results are shown in a graph in Fig. 8.

In the previous approaches, the symmetric algorithms were used to ensure the secure communication between nodes. Moreover, these symmetric algorithms were used either by using PSK i.e. pre-shared keys or by embedding the keys in the hardware as in [8], [9], [10], [12] and [13]. However, if we want to use PKC in constrained nodes then it requires more resource for the generation of keys and keeping both the public and private keys. Our proposed methodology has used asymmetric cryptography, which is more secure and robust by

ensuring that it does not create computational overhead for the constrained devices.

The corresponding graph between the proposed framework and the DES algorithm is constructed using the observation recorded in Table IV. The graph is shown in Fig. 9.

B. Security

The most important and challenging task was to secure the constrained devices to protect them from the security threats and attacks. We have implemented RSA, which is more robust, and secure. Whenever the client or RS wants to communicate with each other, the nodes send request to the Trust Manager for the generation of keys so the client and the RS do not know the keys before initiating the communication. They are assigned keys dynamically each time i.e. every time a new key pair is generated which ensures that the same key is not repeated and moreover, the communicating parties are also kept aside for the generation of keys; this makes our proposed approach more secure which is the strength of our work.

TABLE II.   EXPERIMENTAL SETUP

Parameters	Values
Transmission Bits of Data	
Sets	1024, 2048, 3072
No of tests on data sets	25

TABLE III.   COMPARISON BETWEEN PROPOSED FRAMEWORK AND AES

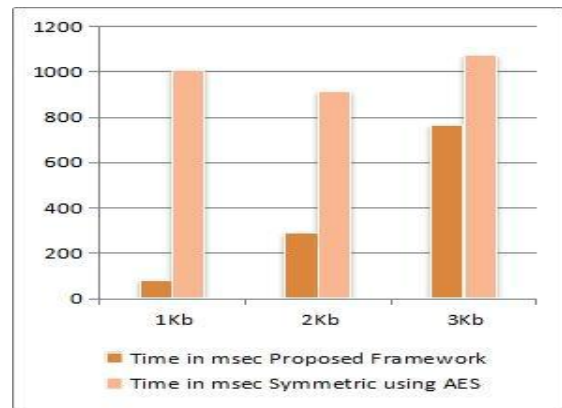|	Time in msec	
	Framework	**AES**
1Kb	79.96	1008.39
2Kb	287.76	914.03
3Kb	765.88	1075.64

Fig. 8.   Encryption and Decryption Time using Proposed Framework and AES.

TABLE IV.   COMPARISON BETWEEN PROPOSED FRAMEWORK AND DES

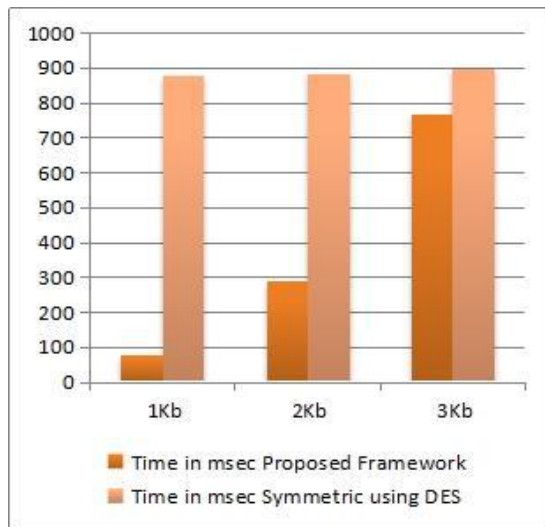|	Time in msec	
	Framework	**DES**
1Kb	79.96	877.86
2Kb	287.76	882.75
3Kb	765.88	897.92

Fig. 9.   Encryption and Decryption Time using Proposed Framework and DES.

C. Dynamicity

The proposed methodology ensures the dynamicity as well because the nodes connect to the Trust Manager before initiating the communication and the Trust Manager does not store the keys statically in memory, instead, it generates each time a new key pair by selecting the prime numbers randomly using cryptographically secure random number generator.

D. Computational Overhead

We demonstrate the computational cost as the computations needed to perform by the constrained devices to ensure the security. As we want to implement Asymmetric algorithm so it is computationally complex and our constrained devices cannot support such computationally complex algorithms. Our proposed framework also ensures that the constrained devices should not be involved in the computations of the complex algorithm of RSA. The RS performs the encryption whenever it wants to send the status of device in encrypted form so the only operation it has to perform is encrypting the status with the key it has received from client. The operation of encryption s^e mod n is simple; it completes this task in 12 millisecond, 17 millisecond and 39 millisecond on average for message length of 1Kb, 2Kb and 3Kb respectively. According to the specification of RS in [21], and this time does not create any kind of computational overhead for RS so it can perform this operation easily and efficiently.

VI. CONCLUSION

In the proposed framework, we have implemented the Asymmetric key algorithm i.e. RSA to ensure the confidentiality of the data shared between RS and the Client. The experimental results show that our proposed work is more secure as compared to the symmetric key approach used earlier. Moreover, we have used Trust Manager, which is responsible for the generation of asymmetric keys because the most computationally extensive part of RSA algorithm is to generate the keys i.e. public and private key pair. We used the Trust Manager for this purpose to reduce the computational

overhead from the constrained nodes; this increases the efficiency of our proposed algorithm. Our Proposed framework is more secure because asymmetric approach is being used which provides more security than the symmetric key algorithms. We also provided a protocol suite that can be used for the communication between client and the RS. This protocol suite is applicable to both of the scenarios when the client initiates the communication as well as when the RS initiates the communication with the Client. The present work deals with the security features at the application layer and the transport layer. We have used CoAP as the application layer protocol in our work. This work can be extended in future and security features at the other layers of protocol stack of IoT can be added. Moreover, we have ensured the secure communication between client and the server i.e. Confidentiality and in future other features of security such as integrity, authentication and non-repudiation can also be added to make it more robust.

REFERENCES

[1] J. D. de Hoz, J. Saldana, J. Fernandez-Navajas, and J. Ruiz-Mas, "IoTsafe, Decoupling Security from Applications for a Safer IoT," IEEE Access, no. 1, pp. 1–1, 2019.

[2] M. Malik, M. Dutta, and J. Granjal, "A survey of Key bootstrapping protocols based on Public Key Cryptography in the Internet of Things," IEEE Access, no. 1, pp. 1–1, 2019.

[3] M. G. Samaila, M. Neto, D. A. B. Fernandes, M. M. Freire, and P. R. M. Inácio, "Security challenges of the Internet of Things," Internet of Things, no. 9783319507569, pp. 53–82, 2017.

[4] V. Adat and B. B. Gupta, "Security in Internet of Things: issues, challenges, taxonomy, and architecture,"Telecommun. Syst., vol. 67, no. 3, pp. 423–441, 2018.

[5] M. S. Henriques and N. K. Vernekar, "Using symmetric and asymmetric cryptography to secure communication between devices in IoT," IEEE Int. Conf. IoT its Appl. ICIOT 2017, 2017.

[6] S. Sciancalepore, G. Piro, G. Boggia, and G. Bianchi, "Public Key Authentication and Key Agreement in IoT Devices with Minimal Airtime Consumption," IEEE Embed. Syst. Lett., vol. 9, no. 1, pp. 1–4, 2017.

[7] K. T. Nguyen, N. Oualha, and M. Laurent, "Novel lightweight signcryption-based key distribution mechanisms for MIKEY," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9895 LNCS, pp. 19–34, 2016.

[8] G. Singh, "A Study of Encryption Algorithms (RSA , DES , 3DES and AES) for Information Security," Int. J. Comput. Appl., vol. 67, no. 19, pp. 33–38, 2013.

[9] R. I. Emori, "Scale models of automobile collisions with breakaway obstacles - Investigation indicates that scale models can be used to show the motion of breakaway signposts and lightposts after being struck by automobiles," Exp. Mech., vol. 13, no. 2, pp. 64–69, 1973.

[10] X. Chen, "Constrained Application Protocol for Internet of Things," Wirel. Mob. Netw., vol. 857, pp. 1–12, 2014.

[11] E. Rescorla and N. Modadugu, "Datagram Transport Layer Security Version 1.2," RFC 6347, 2012.

[12] A. A. Chavan and M. K. Nighot, "Secure CoAP Using Enhanced DTLS for Internet of Things," Int. J. Innov. Res. Comput. Commun. Eng. (An ISO Certif. Organ., vol. 3297, no. 12, pp. 7601–7608, 2007.

[13] J. Granjal, E. Monteiro, and J. S. Silva, "Application-layer security for the WoT: Extending CoAP to support end-to-end message security for

internet-integrated sensing applications," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7889 LNCS, pp. 140–153, 2013.

[14] S. Raza, L. Seitz, D. Sitenkov, and G. Selander, "S3K:Scalable Security with Symmetric Keys - DTLS Key Establishment for the Internet of Things," IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1270–1280, 2016.

[15] C. C. Chang, H. L. Wu, and C. Y. Sun, "Notes on 'Secure authentication scheme for IoT and cloud servers,'"Pervasive Mob. Comput., vol. 38, pp. 275–278, 2017.

[16] L. Qiu, Z. Liu, G. C. Geovandro, and H. Seo, "Implementing RSA for sensor nodes in smart cities," Pers. Ubiquitous Comput., vol. 21, no. 5, pp. 807–813, 2017.

[17] S. Challa et al., "Secure Signature-Based Authenticated Key Establishment Scheme for Future IoT Applications," IEEE Access, vol. 2016, no. 2016, pp. 1–16, 2017.

[18] Y. Zhou, T. Liu, F. Tang, and M. Tinashe, "An Unlinkable Authentication Scheme for Distributed IoT Application," IEEE Access, vol. 7, no. c, pp. 14757–14766, 2019.

[19] S. Raza, T. Helgason, P. Papadimitratos, and T. Voigt, "SecureSense: End-to-end secure communication architecture for the cloud-connected Internet of Things,"Futur. Gener. Comput. Syst., vol. 77, pp. 40–51, 2017.

[20] S. Raza, D. Trabalza, and T. Voigt, "6LoWPAN compressed DTLS for CoAP," Proc. - IEEE Int. Conf. Distrib. Comput. Sens. Syst. DCOSS 2012, pp. 287–289, 2012.

[21] S. Raza, H. Shafagh, K. Hewage, R. Hummen, and T. Voigt "Lithe: Lightweight secure CoAP for the internet of things,"IEEE Sens. J., vol. 13, no. 10, pp. 3711–3720, 2013.

[22] H. H. Hadwan and Y. P. Reddy, "Smart Home Control by using Raspberry Pi & Arduino UNO," Int. J. Adv. Res. Comput. Commun. Eng., vol. 5, no. 4, pp. 283–288, 2016.

[23] C. Hennebert and J. Dos Santos, "Security protocols and privacy issues into 6LoWPAN stack: A synthesis," IEEE Internet Things J., vol. 1, no. 5, pp. 384–398, 2014.

[24] K. Mikhaylov, N. Plevritakis, and J. Tervonen, "Performance Analysis and Comparison of Bluetooth Low Energy with IEEE 802.15.4 and SimpliciTI," J. Sens. Actuator Networks, vol. 2, no. 3, pp. 589–613, 2013.

[25] IEEE Computer Society, IEEE Standard Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs), vol. 2011, no. September. 2011.

[26] Ahmad Mujtaba, "IPv6 over Low-power Personal Area Network (6LoWPAN)," RFC 4919, vol. 67, no. 6, pp. 14– 21, 2007.

[27] P. K. Kamma, C. R. Palla, U. R. Nelakuditi, and R. S. Yarrabothu, "Design and implementation of 6LoWPAN border router," IFIP Int. Conf. Wirel. Opt. Commun. Networks, WOCN, vol. 2016-Novem, pp. 2–6, 2016.

[28] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications,"IEEE Internet Things J., vol. 4, no. 5, pp. 1125–1142, 2017.

[29] S. L. Keoh, S. S. Kumar, and H. Tschofenig, "Securing the internet of things: A standardization perspective," IEEE Internet Things J., vol. 1, no. 3, pp. 265–275, 2014.

[30] K. K. R. Choo, S. Gritzalis, and J. H. Park, "Cryptographic solutions for industrial internet-of-things: Research challenges and opportunities," IEEE Trans. Ind. Informatics, vol. 14, no. 8, pp. 3567–3569, 2018.

[31] N. Sklavos and I. D. Zaharakis, "Cryptography and security in internet of things (IoTs): Models, schemes, and implementations," 2016 8th IFIP Int. Conf. New Technol. Mobil. Secur. NTMS 2016, pp. 3–4, 2016.

[32] M. G. Samaila, M. Neto, D. A. B. Fernandes, M. M. Freire, and P. R. M. Inácio, "Security challenges of the Internet of Things," Internet of Things, no. 9783319507569, pp. 53–82, 2017.

[33] O. P. Pinol, S. Raza, J. Eriksson, and T. Voigt, "BSD-based elliptic curve cryptography for the open Internet of Things,"2015 7th Int. Conf. New Technol. Mobil. Secur. - Proc. NTMS 2015 Conf. Work., 2015.

A Framework for Brain Tumor Segmentation and Classification using Deep Learning Algorithm

Sunita M. Kulkarni[1], Dr. G. Sundari[2]
Department of ECE, Sathyabama Institute of Science and Technology
Chennai, India

Abstract—The brain tumor is a cluster of the abnormal tissues, and it is essential to categorize brain tumors for treatment using Magnetic Resonance Imaging (MRI). The segmentation of tumors from brain MRI is understood to be complicated and also crucial tasks. It can be further use in surgery, medical preparation, and assessments. In addition to this, the brain MRI classification is also essential. The enhancement of machine learning and technology will aid radiologists in diagnosing tumors without taking invasive steps. In this paper, the method to detect a brain tumor and classification has been present. Brain tumor detection processes through pre-processing, skull stripping, and tumor segmentation. It is employing a thresholding method followed by morphological operations. The number of training image influences the feature extracted by the CNN, then CNN models overfit after some epoch. Hence, deep learning CNN with transfer learning techniques has evolved. The tumorous brain MRI is classified using CNN based AlexNet architecture. Further, the malignant brain tumor is classified using GooLeNet transfer learning architecture. The performance of this approach is evaluated by precision, recall, F-measure, and accuracy metrics.

Keywords—Brain MRI; segmentation; CNN; deep learning; transfer learning

I. INTRODUCTION

Brain tumor detection and classification is one of the active research areas. The segmentation technique extracts the brain tumor from the brain MRI, and classification algorithms classify the brain tumor into respective categories. It has an essential task in interpreting, extracting features, analyzing, and interpreting images in many applications. It has been widely used in brain imaging to classify tissues, detect tumors, assess tumor size, delineating blood cells, and operation preparation.

The brain tumor identification, size, shape, and location are carried using Magnetic Resonance Imaging (MRI). The brain tumor is caused by the irregular uncontrolled spreading out of cells in the brain called trauma. It is classified as primary and secondary tumors. Primary brain tumors are non-cancer or benign tumors developed in the brain tumor itself. The brain tumor that initiates in the rest of the body parts such as the lungs, breast, and then migrates to the brain over the blood flow is the secondary tumor. These secondary tumors are cancer-causing or malignant.

"The American Association of Neurological Surgeons (AANS)" has demonstrated the types of tumors according to their nature [1]. Fig. 1 shows the types of tumors.
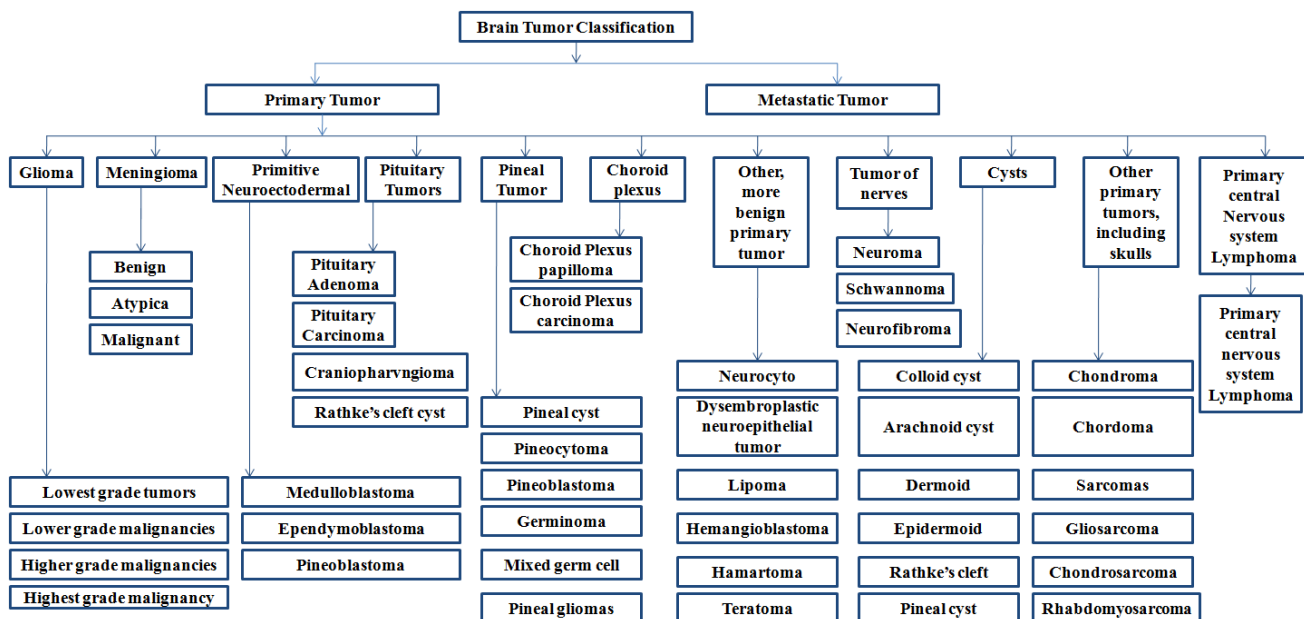

Fig. 1. Brain Tumor Classification According to AANS.

This research aims first to detect the brain tumor from MRI. The tumorous MRIs further classifies using the transfer learning architecture, i.e., AlexNet, into Malignant and benign. The cancerous malignant tumors are also classified into glioma and meningioma using the GoogLeNet architecture of CNN. The classification architectures are selected based on their accuracy by hyper tuning the training parameters. As the database is limited, the transfer learning model helps speed up training and improves the job of the classifier.

This paper is structured as Section II offers a brief outline of the recent progress in the brain tumor detection, and classification aided by Machine Learning (ML) and Deep Learning (DL) algorithms into malignant and benign as well as glioma and meningioma. Section III presents the architecture of the proposed methodology. It includes the brain tumor segmentation and classification using transfer learning. Quantitative and qualitative analysis is present in Section IV. Section V gives concluded the approach and suggests the future direction for this research work.

II. LITERATURE SURVEY

Brain tumor classification plays a substantial role in the suggested methodology. In the latest times, the classification process widely uses DL and ML algorithms. This section presents the methods and previous works in brain tumor segmentation and ML-based classification from MRI.

A. Padma et al. [2] introduced segmentation and classification of the brain tumor in the CT image. It used Dominant Gray Level Run Length Matrix (DGLRLM) as a feature extraction technique. The proposed method for finding tumors in brain CT images with the use of a Support Vector Machine (SVM) classification shows an efficient segmentation algorithm. Their work aims to combine the DGLRLM and wavelet-based feature extraction techniques. Ideal shape features are selected using a genetic algorithm. SVM uses DGLRLM and wavelet features as input. The average accuracy rate is over 97%.

A. Hamamci et al. [3] proposed a brain tumor segmentation approach for radiotherapy applications. They provide a quick real-world platform for the classification of tumors with the least user collaboration to aid researchers along with clinicians in surgical preparation and assess response to treatment. Importantly, Cellular Automata (CA) focuses on seeded tumor classification method on MRI, which offered seed selection. First, they build a relationship between CA-based segmentation for graph-theory techniques to establish that the redundant CA structure responds to the shortest path problem. The state conversion performance of the CA followed to estimates the shortest path. Besides, the segmentation problems adjusted the sensitivity factor amplitude for the segmentation problem, and a level set of tumor probability maps generated from CA states. Only adequate clinical data can be collected from the user with clinical practice to reset the algorithm by drawing on the largest diameter of the tumor.

Aneza and Rawat et al. [4] introduced the Fuzzy Clustering Mean (FCM) segmentation approach. The performance of the segmentation is evaluated based on cluster validation functions, processing time, and convergence rate. It achieves an of 0.537% misclassification error using the Intuitionistic Fuzzy C-Means (IFCM) method.

Ravindra Sonavane et al. [5] present the approach for the sorting of brain MRI classification into malignant and benign. This system used the AdaBoost algorithm. It consists of pre-processing using Anisotropic Diffusion Filtering, feature extraction by Discrete Wavelet Transform (DWT), and classification by using the Adaboost ML algorithm. Experimental results use 155 MRI images for performance evaluation.

V. Wasule and P. Sonar [6] presented a method for MRI classification of the brain in malignant versus benign. In this paper, feature extraction uses the GLCM algorithm. This system uses SVM and K Nearest Neighbors (KNN) for the classification of malignant vs. benign and low grade vs. high-grade glioma. The clinical dataset is used for malignant and benign classification, while the BRATS 2012 dataset for high-grade and low-grade glioma classification. The system shows that SVM performs better.

Saleck et al. [7] presented a robust and accurate system using the FCM segmentation technique. It extracts the tumorous mass from the MRIs. The presented method aims to avoid problematic estimation by selecting the cluster in the FCM as input data, which can provide us with the data needed to execute mass partitions by fitting only two clusters of pixels. GLCM is used to extract texture properties to obtain the optimal threshold, which divides between the selected group and the pixels of other groups, which significantly affects the precision.

M. Rashid et al. [8] examined the MRI image and a method for an even clearer vision of the position acquired by the tumor. MRI brain image is an input of the system. This method used an Anisotropic filter to remove the noise from the brain MRI, SVM used for segmentation followed by a morphological operation.

T. Ren et al. [9] proposed the method to solve brain tumor segmentation. Initially, the irrelevant information is removed from the image by histogram equalization. Then, by the study and research, three segmentation techniques were proposed by them are FCM, Kernel-based FCM (KFCOM), and Weighted Fuzzy Kernel Clustering (WKFCOM). The assessment displays that WKFCOM does better as compared to KFCOM, a 2.36% lesser false classification rate.

P. Kumar et al. [10] presented a four-step sorting method for brain tumor segmentation and classification. The Wiener filter is used to denoising the image in the initial stage, image decomposition in the second stage. Then the combined edge and texture feature is combined with the Principal Component Analysis (PCA) is performed to minimize the dimension of the features. The last step is the classification step, which classifies brain tumors from MRI using SVM classification.

S. Kebir et al. [11] provides a supervised approach to detect abnormalities of the brain, predominantly proposing MRI images in three stages. The first stage is the creation of a DL CNN model, and the next section of tumor segmentation using the K-mean clustering. They are introducing CNN models to classify the abnormality of the MRIs.

Muhammad Talo et al. [12] suggested the deep transfer learning technique classifies MRI images of the brain as normal and abnormal. The pre-trained CNN model used the ResNet34 algorithm. The database is expands using data augmentation techniques. This method is validated on the Harvard Medical School MR dataset [13] and suggests finding additional abnormalities such as autism, stroke, Parkinson's, and Alzheimer's disease.

S. Deepak et al. [14] presented a brain tumor classification technique using GoogLeNet. Brain tumors are classified according to their nature as glioma, meningioma, and pituitary. Because the classification of brain tumors is comparatively complex, the convenience of classification means that there is a reasonable degree of deviation in size and shape, which affects the classification. This problem is most surprising when it comes to use traditional ML techniques. To defeat this problem, they introduced transfer learning to achieve a higher level of learning accuracy compared to the previous model. The significant enhancement was attained even for smaller datasets. This system suggested GoogLeNet, which is prevalent at the Softmax level, with some modifications for a wide variety of tumor classifications. The CNN focused GoogLeNet method attained a better precision of 92.3% that reached 97.8% by using multiclass SVM.

III. PROPOSED METHODOLOGY

This section presents brain tumor detection and classification techniques shown in Fig. 2. The three stages of the proposed system are:

- Brain Tumor Detection.
- Benign and Malignant Brain MRI Classification.
- Glioma and Meningioma Brain MRI Classification.

A. Brain Tumor Detection

The methodology to detect the brain tumor from the brain MRI discusses in this section.

1) Tumor Vs. Non-Tumor Dataset: The online data is collected from the online source for tumorous and non-tumorous classification [15]. This dataset consists of 154 tumorous MRIs and 91 non-tumorous MRIs. Fig. 3(a) and Fig. 3(b). show the sample of tumorous and non-tumorous brain MRI.

2) Pre-processing: In the normalization process, the intensity falls within the range of pixel values converted into [0 1] range. In this process, each pixel intensity is divided by the maximum intensity values within an image. Normalization can create binary thresholding by creating a more extensive source. Such MRI images can help to prevent classifications affected by variations of grayscale value.

3) Skull stripping: Skull stripping is a necessary procedure in the biomedical image examination for the efficient analysis of brain tumors from brain MRI [16]. It eliminates the non-brain parts like skin, fat, and skull from the brain MRI.
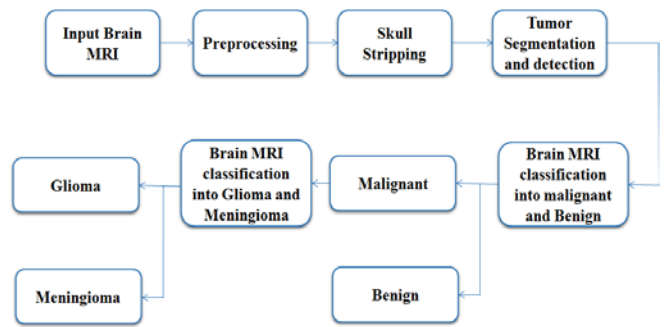

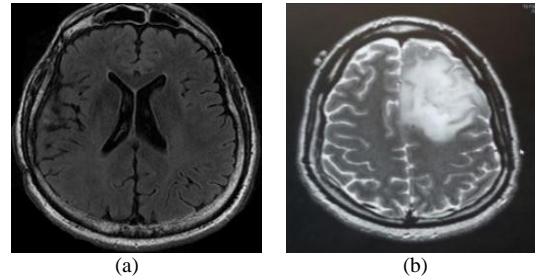Fig. 2. Block Diagram of the Brain Tumor Detection and Classification System.

Fig. 3. Brain MRI Samples (a) Tumorous (b) Non-Tumorous.

There are many different ways to segment the skull. Among them, skull stripping is the technique that focuses on automatic segmentation and morphological operation. Fig. 4 shows the process of the skull stripping algorithm.
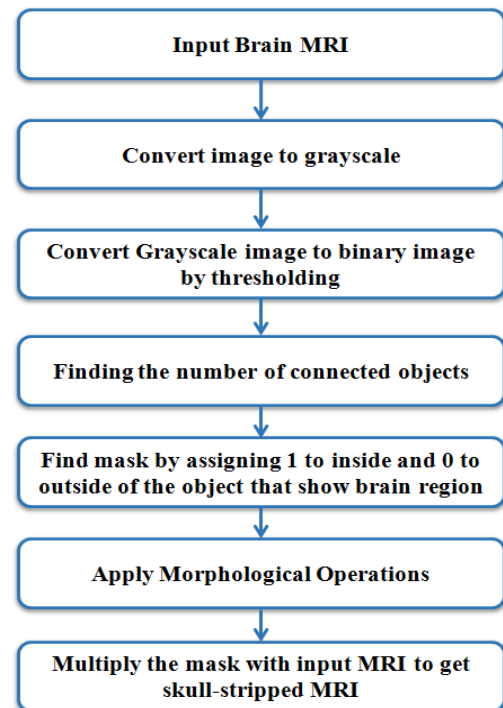

Fig. 4. The flow of the Skull Stripping Process.

4) Brain tumor detection and area calculation: The following methods achieve the brain tumor segmentation from the brain MRIs: Initially, the input brain MRIs are pre-processed and converted into a binary image by thresholding technique and morphological operation. The thresholding process chooses 128 as a threshold value. The pixel whose values are more significant than the defined threshold is subject to 1, while rest are subject to 0. It makes two areas built around the tumor region. The conversion of a grayscale image into a binary image is done by (1).

$$f_{g(x,y)} = \begin{cases} 1 & I(x,y) > T \\ 0 & else \end{cases} \tag{1}$$

Where $I(x,y)$ is the intensity value of the grayscale pixel and $f_{g(x,y)}$ is the resultant binary pixel.

In the next step, morphological processing is performed by erosion and dilation process on binary image to gain the proper boundary of the tum while dilation operation filling the gaps within the detected object using Erosion operation. The morphological process uses a small mask (3x3 or 5x5) of different sizes, and it is applied over the image known as the structuring element. This element has various shapes, such as lines, disk, diamonds, etc.

The erosion of an image is given by

$$A \ominus B = \{z \in E \mid B_{\approx} \subseteq A\} \tag{2}$$

The dilation of an image is given by

$$A \oplus B = \bigcup_{b \in B} A_b \tag{3}$$

After the detection of tumors, calculates the area of the tumor for further processing (Eq. (4)).

$$Area = \sum_{i=0}^{n} i == 1 \tag{4}$$

B. Benign and Malignant brain MRI Classification

Once the tumor is detected, the brain MRI is analyzing for the malignancy of the image. In this section, the CNN based AlexNet transfer learning architecture is utilized for training and classifying the brain MRI for classification.

1) Benign Vs. Malignant Dataset: The classification of brain tumors into benign and malignant uses clinical dataset collected from the hospital. The database comprises of benign and malignant MRI images [17]. The complete database-distribution is tabulated in Table I. The database has MRIs that are pre-processed by brain augmentation and segmentation methods after separating testing and training data. Finally, it presents the performance of the training and testing process.

MRIs of the dataset are pre-processed and remove non-brain parts by skull stripping, as illustrated in Sections III (A(2)) and III (A(3)) of this paper.

2) Data augmentation: DL architecture requires an enormous amount of data that includes variation. In transfer learning, data augmentation forms happen to be a vital component of the pre-processing. If the database is comparatively lesser, then that DL model may begin to remember the features that are too explicit to that particular database, which leads to overfitting. Thus to avoid the overfitting dilemma, the dataset should be huge with significant dissimilarity, but this is an extremely puzzling task in this scenario of clinical images. One remedy could be to augment the existing database synthetically. Such a practice is general when dealing with image-based data [18]. Data augmentation includes various methods like flipping, rotation, adding noise, scaling translation, resizing, perspective transform, etc. The parameters related to data augmentation used in the proposed method are as charted in Table II.

3) Training using AlexNet: In the presented work, the pre-trained CNN network, referred to as 'AlexNet,' is employed. AlexNet is among the renowned architecture that comprises of Convolution Layers (CL) (five in number), max-pooling layers (three in number), and Fully Connected Layer (FCL) (three in number). It was training for classifying 1000 different objects [20]. There may be few objects that don't go with the original dataset. So the network can have specific layers kept back to identify those non-belonging objects. Fig. 5. illustrates the architecture of AlexNet.

With the actual size of the dataset, training a DL model from zero is inadequate since there is a tremendous rise in the number of training images. Pre-trained AlexNet architecture employs three different steps to avoid this situation. In the first step, the classification layer of the AlexNet replaces with the softmax layer that includes two categories (benign and malignant). Then in the second step, the weights are modified and backpropagate to retrain the images. The learning rate is set to a lesser value so that the weights of the CL do not alter intensely. At the same time, the weights of the FLC set asystematically. The stochastic gradient descent algorithm is employed to appraise the weight depending on the input database of brain MRI. This procedure helps in achieving the optimum weights of the exact network system.

TABLE I.   DATABASE DISTRIBUTION FOR MALIGNANT VS. BENIGN CLASSIFICATION

Type of Tumor	Training	Testing	Total Image
Benign MRI	75	25	100
Malignant MRI	75	25	100

TABLE II.   DATA AUGMENTATION PARAMETERS FOR ALEXNET

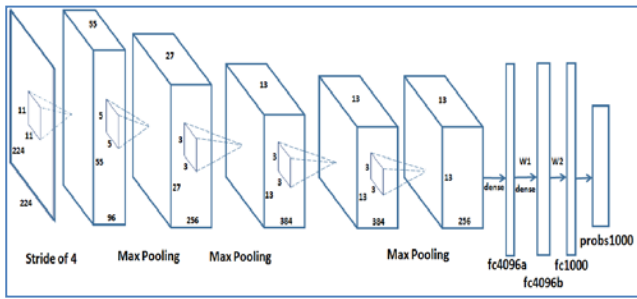| Sr. No	Parameter	Value
1	Random X Reflection	1
2	Random Y Reflection	0
3	Random Rotation	[0,0]
4	Random X Scale	[1,1]
5	Random Y Scale	[1,1]
6	Random X Shear	[0,0]
7	Random Y Shear	[0 0]
8	Random X Translation	[-10 10]
9	Random Y Translation	[-10 10]

Fig. 5.   Architecture of AlexNet.

C. Glioma and Meningioma brain MRI Classification

This section presents the GoogLeNet algorithm for the classification of brain MRI into Glioma and Meningioma.

1) Glioma Vs Meningioma Dataset: This approach uses a clinical dataset of Glioma and Meningioma brain MR images as an input. The dataset is collected from the Cancer Hospital and validated by the radiologist. The database distribution for training and testing is as shown in Table III.

The skull stripping is the initial step used to remove the non-brain part and select the region of interest. The process of skull stripping is shown in Sections 3.1.2 and 3.1.3 of this paper.

2) Data augmentation: Data augmentation includes various methods such as flipping, rotation, adding noise, scaling translation, resizing, perspective transform, etc. The parameters related to data augmentation of the proposed method are as charted in Table IV.

3) Training using GoogLeNet: VGG-16 has CL piled one above the other while GoogLeNet has pooling and CL in a parallel fashion, which helps in the feature extraction through various kernel sizes. It enhances the network depth and a higher performance level. Also, the network employs 1×1 convolution to govern the volume size passed for additional processing in the inception module. It is nothing but a collection of pooling and convolution operations executed parallelly to extract features with the aid of various scales. GoogLeNet network consists of 24 million parameters due to which it has lesser computing complexities as associated with VGG-16 and AlexNet. Instead, of FCL network employs a Global Average Pooling layer. Eventually, in ILSVRC-2014, GoogLeNet had got an error of 6.67%. The GoogLeNet architecture is as displayed in Fig. 6.

TABLE III.   Database Distribution for Glioma Vs. Meningioma Classification

Type of Tumor	Total Image	Training	Testing
Glioma MRI	271	218	53
Meningioma MRI	98	79	19

TABLE IV.   Data Augmentation Parameters for GoogLeNet

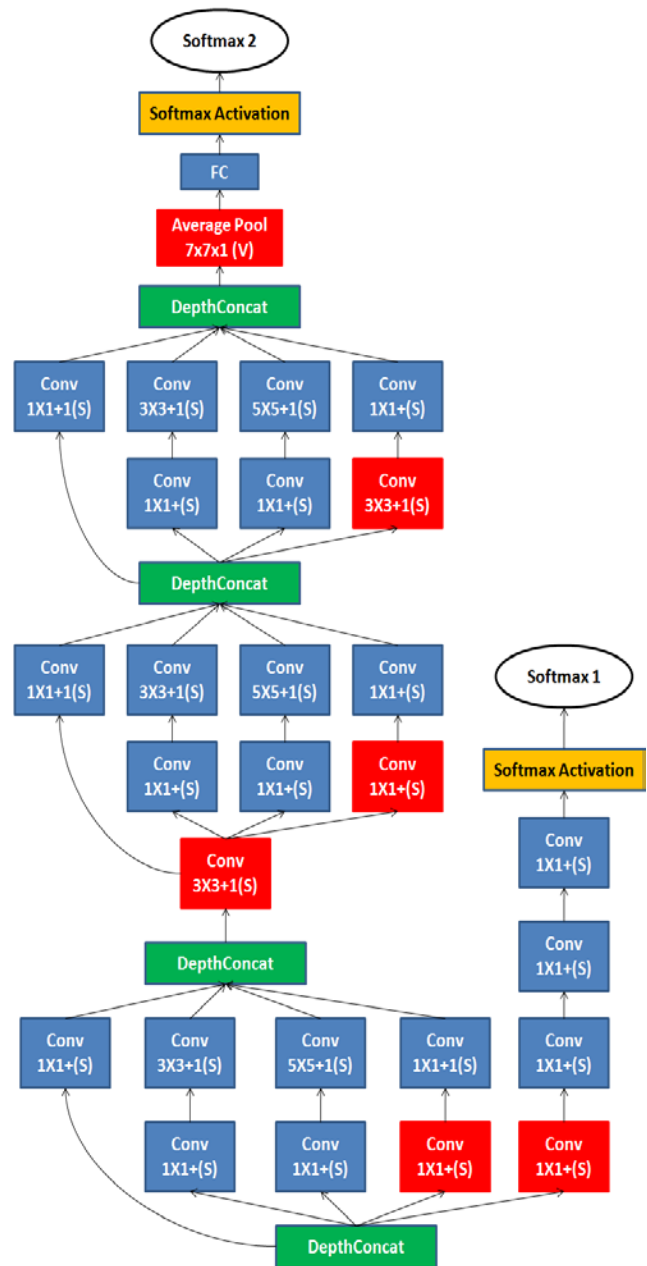| Sr. No	Parameter	Value
1	Random X Reflection	1
2	Random Y Reflection	0
3	Random Rotation	[0,0]
4	Random X Scale	[1,1]
5	Random Y Scale	[1,1]
6	Random X Shear	[0,0]
7	Random Y Shear	[0 0]
8	Random X Translation	[-10 10]
9	Random Y Translation	[-10 10]

Fig. 6.   Layer Architecture of GooLeNet.

IV. RESULT AND DISCUSSION

This section presents the Qualitative and Quantitative analysis of the proposed system. This system uses 64-bit MATLAB 2019a software. This system uses a single Central Processing Unit (CPU) to train the network. The quantitative analysis evaluates precision, recall, F-measure and accuracy evaluation metrics, and mathematically denoted as:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (6)$$

$$\text{F} - \text{measure} = 2 \text{ X} \frac{\text{PrecisionXRecall}}{\text{Precision} + \text{Recall}} \qquad (7)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (8)$$

True Positive (TP) is referring to as a benign identifies as a benign (or glioma is identifying as glioma). True Negative(TN) is referring to as a malignant identified as a malignant (or meningioma is identifying as meningioma). False Positive(FP) is referring to as a benign identified as a malignant (or glioma is identifying as meningioma). False Negative(FN) is referring to as a malignant identified as a benign (or meningioma identifying as glioma).

A. Analysis of Brain Tumor Detection method

The results of the tumor detection from brain MRI is as shown in Fig. 7(a-f).

The input brain MRI image is as shown in Fig. 7(a), is chosen for analysis. First, the image is pre-processed using a median filter and then binarized using the thresholding method. The binary mask is again processed by erosion and dilation operations and selects the most massive mask, as displayed in Fig. 7(b). This mask is multiplied with the input image to get a skull stripped image, as shown in Fig.7(c). The brain tumor segmentation uses the thresholding method to select a region of interest. Threshold output, as displayed in Fig. 7(d). The detected tumor is shown in Fig. 7(e) and declared as a tumorous image with the percentage of area covered by the tumor in the brain MRI Fig. 7(f).

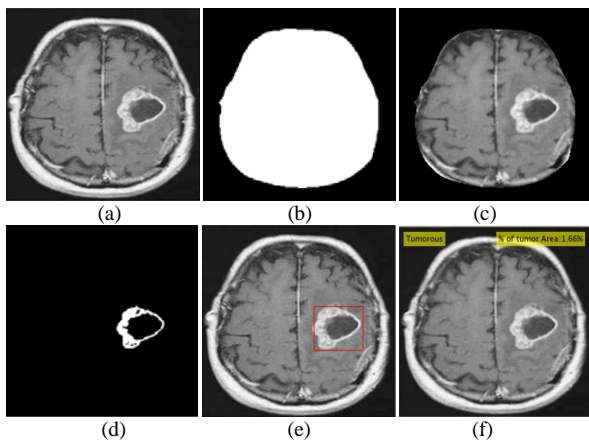| (a) | (b) | (c) |
| (d) | (e) | (f) |

Fig. 7. Tumor Detection: (a) Input Image (b) Binary Mask Skull (c) Stripped Image (d) Segmented Tumor (e) Detected Tumor (f) Output Image (Shows Tumorous and % of the Area).

Table V shows the input brain MRI in the second column. Images in the first three rows contain non-tumorous images, while the next three rows contain tumorous images. The respective segmented output is shown in the third column. The non-tumorous images contain no white pixels (1's); hence % area of the tumor is zero while the tumorous images show respective values. The area of the tumor could help the doctors in the analysis and planning of surgery.

TABLE V. ANALYSIS OF BRAIN TUMOR SEGMENTATION AND % AREA OF THE TUMOR

Sr. No	Input Image	Segmented tumor	% area of the tumor
1			0
2			0
3			0
4			1.66
5			6.8039
6			3.0685

B. Analysis of Benign and Malignant Classification Method

The qualitative analysis of malignant and benign classification using the AlexNet CNN algorithm is as displayed in Fig. 8.

Qualitative analysis of benign vs. malignant classification is performed on the testing database. The input samples of the benign and malignant brain MRI areas are shown in Fig. 8(a) and Fig. 8(c) and a resultant class of the AlexNet are displaying in Fig. 8(b) and Fig. 8(d), respectively.

The training progress of AlexNet is as displayed in Fig. 9.

The quantitative analysis of malignant and benign classification using the AlexNet, Vgg16, ResNet18, ResNet50, and GoogLeNet CNN algorithm is as displayed in Table VI, and its graphical analysis is showing in Fig. 10.

This approach is more generalized and shows better accuracy on the testing dataset. The relative examination of this system with state-of-art methods shows the dominancy of the proposed AlexNet method.

TABLE VI. COMPARATIVE ANALYSIS OF BENIGN AND MALIGNANT CLASSIFICATION METHODS

Methods	Precision	Recall	F-measure
SVM [6]	1	0.76	0.8636
KNN [6]	0.88	0.73	0.7999
Proposed Method (AlexNet)	**0.9375**	**1**	**0.9677419**
Proposed Method (Vgg16)	0.55	0.5	0.5238095
Proposed Method (ResNet18)	0.9	0.783	0.8372093
Proposed Method (ResNet50)	0.25	0.833	0.3846154
Proposed Method (GoogLeNet)	0.8	1	0.8888889

Fig. 10. Performance Analysis of Proposed Systems with State of the Art Method for Malignant and benign Classification.

C. Analysis of Glioma and Meningioma Classification Method

Qualitative analysis performs on the testing dataset for classification of Glioma vs. Meningioma. The results are displaying in Fig. 11. The input samples of the Glioma and Meningioma brain MRI display in Fig. 11(a), and Fig. 11(c) and a resultant class of the GoogLeNet are shown in Fig. 11(b) and Fig. 11(d), respectively.

Fig. 8. Qualitative Analysis of the benign and Malignant Classification using AlexNet (a) (c) Input MRI Image (b) (d) Output Classified Image.
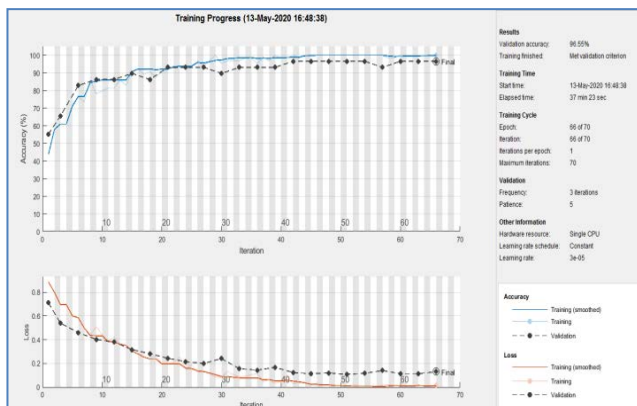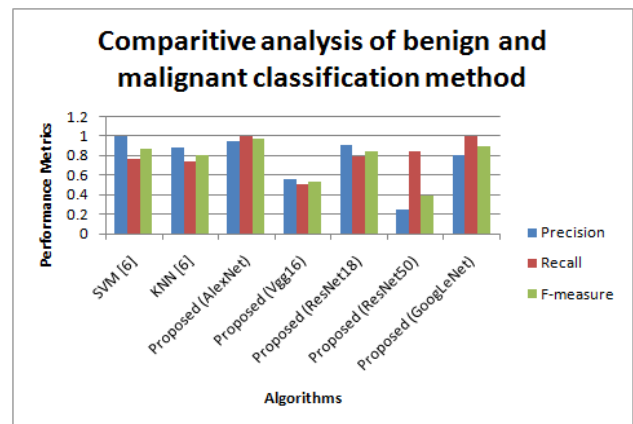

Fig. 9. Training Progress of AlexNet Transfer Learning Architecture for benign and Malignant Brain MRI.
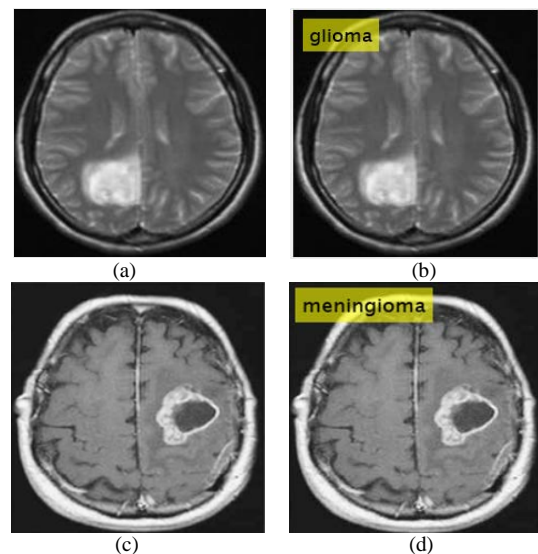

Fig. 11. Qualitative Analysis of the Glioma and Meningioma Classification using AlexNet (a) (c) Input MRI Image (b) (d) Output Classified Image.
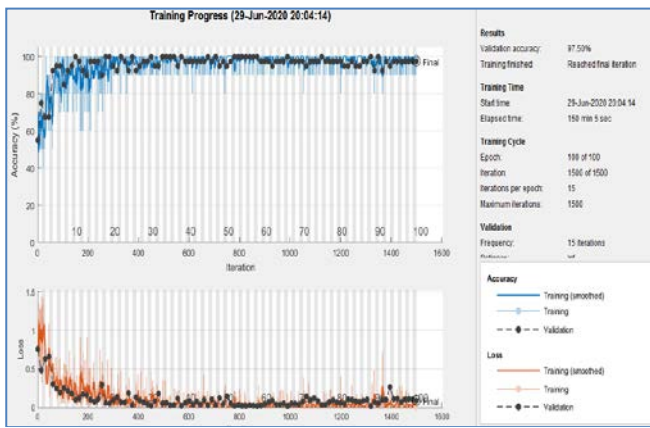
Fig. 12. Training Progress of GoogLeNet Transfer Learning Architecture for Glioma and Meningioma Brain MRI.

The training progress of GoogLeNet is as displayed in Fig. 12.

The quantitative analysis of malignant and benign classification using the AlexNet, GoogLeNet, ResNet18, and ResNet50 CNN algorithm is as displayed in Table VII.

The graphical analysis of the proposed approaches with state of the art method is given in Fig. 13.

TABLE VII. COMPARATIVE ANALYSIS OF GLIOMA AND MENINGIOMA CLASSIFICATION METHODS

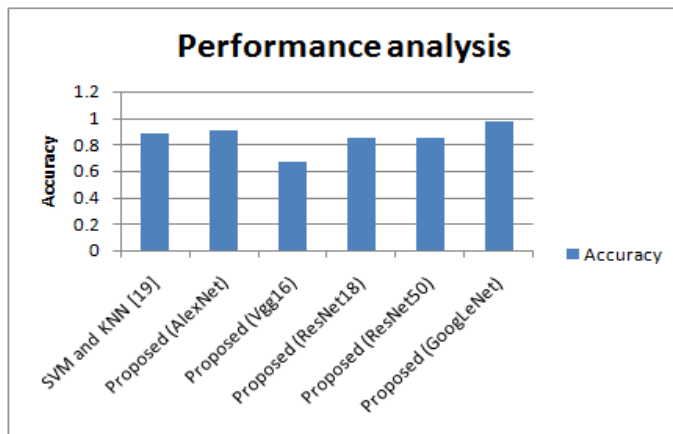| Methods	Precision	Recall	F-measure	Accuracy
SVM and KNN [19]	-	-	-	0.88
Proposed Method (AlexNet)	0.8863	0.975	0.9285	0.9047
Proposed Method (Vgg16)	0.6591	0.8529	0.8966	0.6667
Proposed Method (ResNet18)	0.9318	0.8723	0.9010	0.85
Proposed Method (ResNet50)	0.8863	0.9069	0.8965	0.85
Proposed Method (GoogLeNet)	**0.95**	**1**	**0.9743**	**0.9750**

Fig. 13. Performance Analysis of Proposed Systems with State of the Art Method.

The approach of GoogLeNet is more generalized and shows better accuracy on the testing dataset. The relative analysis of the proposed system with state-of-art approaches shows the dominancy of the proposed GoogLeNet approach.

V. CONCLUSION

In this paper, a technique for brain tumor detection and grading of tumorous MRIs into malignant and benign as well as malignant brain MRI into glioma and meningioma is proposed. The brain tumor detection is executed by pre-processing techniques followed by skull stripping and brain tumor segmentation. This approach promisingly segments the brain tumor from MRI. The tumorous images are further classified into malignant and benign using CNN based AlexNet transfer learning algorithms. The proposed method achieved a precision of 0.9375, recall of 1, and f-measure of 0.9677.

Similarly, the malignant MRI further classified into Glioma and Meningioma using CNN based GoogLeNet transfer learning algorithm. GoogLeNet model successfully got a precision of 0.95, recall of 1, f-measure of 0.9743, and accuracy of 0.9750. The proposed approaches (malignant vs. benign or glioma vs. meningioma) shows better results than existing methods.

The deep neural networks, particularly CNN, are rarely used for boundary detection problems. Therefore, the uses of deep neural networks can be a future direction for brain tumor segmentation and detection problem. In addition to this, the 3D brain boundary detection can be possible with this method. In the case of classification of dissimilar types of brain MRI (malignant vs. benign or glioma vs. meningioma), the system can be made more robust and generalized by training the deep neural networks on extensive data. Also, the system will be extended for further classification of low-grade and high-grade glioma as well as meningioma.

REFERENCES

[1] Classification of Brain Tumors: https://www.aans.org/en/Media/Classifications-of-Brain-Tumors. Access on 29 June 2020.

[2] A PADMA and R.Sukanesh, "Automatic Classification and Segmentation of Brain Tumor in CT Images using Optimal Dominant Gray level Run length Texture Features" International Journal of Advanced Computer Science and Applications(IJACSA), 2(10), pp. 53-59, 2011.

[3] A. Hamamci, N. Kucuk, K. Karaman, K. Engin, and G. Unal, "Tumor-Cut: Segmentation of Brain Tumors on Contrast Enhanced MR Images for Radiosurgery Applications," in IEEE Transactions on Medical Imaging, vol. 31, no. 3, pp. 790-804, March 2012.

[4] Aneja, Deepali, and Tarun Kumar Rawat. "Fuzzy Clustering Algorithms for Effective Medical Image Segmentation." International Journal of Intelligent Systems and Applications, Vol. 5, pp. 55-61, 2013.

[5] R. Sonavane and P. Sonar, "Classification and segmentation of brain tumor using Adaboost classifier," 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, pp. 396-403, 2016.

[6] V. Wasule and P. Sonar, "Classification of brain MRI using SVM and KNN classifier," 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), Chennai, pp. 218-223, 2017.

[7] M. M. Saleck, A. ElMoutaouakkil, and M. Mouçouf, "Tumor Detection in Mammography Images Using Fuzzy C-means and GLCM Texture Features," 2017 14th International Conference on Computer Graphics, Imaging and Visualization, Marrakesh, pp. 122-125, 2017.

[8] M. H. O. Rashid, M. A. Mamun, M. A. Hossain, and M. P. Uddin, "Brain Tumor Detection Using Anisotropic Filtering, SVM Classifier and Morphological Operation from MR Images," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, pp. 1-4, 2018.

[9] Ren, T., Wang, H., Feng, H., Xu, C., Liu, G., & Ding, P, "Study on the improved fuzzy clustering algorithm and its application in brain image segmentation," Appl. Soft Comput., Vol. 81, pp. 1-9, 2019.

[10] Preetham Kumar and Vijayakumar B, "Brain Tumour MR Image Segmentation and Classification Using by PCA and RBF Kernel Based Support Vector Machine," J Middle-East Journal of Scientific Research, Vol. 23, Issue 9, pp.2106-2116, 2015.

[11] S. T. Kebir and S. Mekaoui, "An Efficient Methodology of Brain Abnormalities Detection using CNN Deep Learning Network," 2018 International Conference on Applied Smart Systems (ICASS), Medea, Algeria, pp. 1-5, 2018.

[12] M. Talo, U. B. Baloglu, O. Yıldırım, & U. Rajendra Acharya, "Application of deep transfer learning for automated brain abnormality classification using MR images," Cognitive Systems Research, Vol. 54, pp. 176-188, 2019.

[13] Harvard Medical School MR dataset: http://www.med.harvard.edu/AANLIB/. Access on 29 June 2020.

[14] S. Deepak, P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," Journal of Computers in Biology and Medicine, Vo. 111, 2019.

[15] Brain Tumor Detection: https://github.com/MohamedAliHabib/Brain-Tumor-Detection/tree/master/yes. Access on 29 June 2020.

[16] Y. Zhang, Z. Dong, L. Wu, S. Wang, and Z. Zhou, "Feature Extraction of Brain MRI by Stationary Wavelet Transform," 2010 International Conference on Biomedical Engineering and Computer Science, Wuhan, pp. 1-4, 2010.

[17] Himaja Byale, Dr. Lingaraju G M and Shekar Sivasubramanian, "Automatic Segmentation and Classification of Brain Tumor using Machine Learning Techniques," International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 14 (2018) pp. 11686-11692.

[18] J. Wei, K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," 9th International Joint Conference on Natural Language Processing, pp. 6382-6388, 2019.

[19] Evangelia I. Zacharaki, Sumei Wang, Sanjeev Chawla, Dong Soo Yoo, Ronald Wolf, Elias R. Melhem, and Christos Davatzikos, "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme," Journals of Magn. Reson. Med., Vol. 62, Issue 6, pp.1609-1618, 2009.

[20] ZarNawab Khan Swati, Qinghua Zhao, Muhammad Kabira, Farman Ali, Zakir Ali, Saeed Ahmed, Jianfeng Lu, " Brain tumor classification for MR images using transfer learning and fine-tuning, "Journal of Computerized Medical Imaging and Graphics, 75 2019, pp. 34–46.

Movie Rating Prediction using Ensemble Learning Algorithms

Zahabiya Mhowwala[1], A. Razia Sulthana[2], Sujala D. Shetty[3]
Department of Computer Science
Birla Institute of Technology and Science
Dubai, United Arab Emirates

Abstract—Over the last few decades, social media platforms have gained a lot of popularity. People of all ages, gender, and areas of life have their presence on at least one of the social platforms. The data that is generated on these platforms has been and is being used for better recommendations, marketing activities, forecasting, and predictions. Considering predictions, the movie industry worldwide produces a large number of movies per year. The success of these movies depends on various factors like budget, director, actor, etc. However, it has become a trend to predict the rating of the movie based on the data collected from social media related to the movie. This will help a number of businesses relying on the movie industry in making promotional and marketing decisions. In this report, the aim is to collect movie data from IMDB and its social media data from YouTube and Wikipedia and compare the performance of two machine learning algorithms – Random Forest and XGBoost – best known for their high accuracy with small datasets, but large feature set. The collection of data is done from multiple sources or APIs.

Keywords—Machine learning; ensemble learning; random forest algorithm; XGBoost; movie rating prediction

I. INTRODUCTION

Living in a socially and digitally connected world, everyone leaves a digital trace of themselves in different forms on the web. People are actively sharing their emotions, feelings, opinions and views through social communications. These communications generate a huge amount of data which is being analyzed to predict, recommend, monitor or cope with varied events, from simple matters to complex problems.

The movie success prediction has a vast range of attributes that gives a holistic approach to perform predictions, movies are something that creates lots of buzz in digital space, and based on the stardom of celebrities there are lots of hailing and criticism bubbling on social media platforms. Also, movie prerelease phase is very strategic and well defined which includes releasing teasers, trailers, a series of movie promotion activities, etc. Connecting the dots, these dots are the number of views, page counts, and sentiments of people on social media. But everyone's views and comments do not hold equal values. There are influencers and well-known critics whose views and comments hold more value than rest. Many social media platforms by default work on an algorithm where they rank the comments based on popularity and relevance thus, rather than collecting all of the comments of each video, only the comments which hold greater relevance and enough values to cover almost every point of view were collected. This

strategy allowed to trace the digital footprints of buzz and move closer to predict the future of the product.

In everyday lives, when people have to make important decisions it is highly unlikely that the decision is based on only one factor. Rather, the decision is always based on the collective opinions coming from a lot of different sources. Ensemble learners operate on a very similar idea, where decisions from multiple learning models are combined to improve the performance of the output. The idea is to combine the results of weak individual models which are prone to overfitting and generate a model which would reduce the error. The simple techniques used for combining the results include – Maximum Voting, Averaging, and Weighted Averaging. Some advanced techniques include – Model Stacking, Blending, Bagging and Boosting.

Random forests are bagging ensemble method that use decision trees as weak learners. It performs feature bagging along with randomly selecting training data for each of its learner models. The random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of the observations, splitting nodes in each tree considering a limited number of features. The final predictions of the random forest are made by averaging the predictions of each individual tree. This reduces the correlation between the different predictions and makes the process more robust to missing values.

Gradient boosting is a boosting ensemble that trains many models in a gradual, additive, and sequential manner. It does so by minimizing the loss function, by repetitively leveraging the residuals (or error) of the previous model and strengthening it. XGBoost is the optimized gradient boosting algorithm using parallelization for tree building, tree pruning and regularization to avoid overfitting, efficient handling of missing data and hardware optimization. There are many hyperparameters that needs to be set for efficient use of this algorithm, it includes – n_estimaters, max_depth, eta or learning rate, reg_alpha & reg_lamba (regularization terms).

A. Problem Statement

Prior knowledge about the success or failure of a particular movie and what factors affect the success will benefit the production houses in how to go about with promoting and other expensive business decisions. For example, if a movie does well in test screenings or if they anticipate good reviews [1] from social media due to promotional activities, they can decide to release it on an opening weekend in more theaters in

hopes of bringing in more revenue. One of the ways the success of a movie is identified is through its ratings, predicting the ratings of a movie before it is released based on the data available at that moment will help in decision making.

The objective is to predict the ratings of the movies using two ensemble learning algorithms - Random Forest and XGBoost, that can be used to evaluate the success or failure of a movie before its release and then to compare them on their performance. The data will be collected from IMDb, people's reviews on the movies will be collected from YouTube comments, and other information to strengthen the model will be collected from Wikipedia page of the movie. The rest of the paper is organized in following fashion. Section 2 discuss the literature work related to the proposed work. Section 3 details the data collection procedure. Section 4 describes the proposed model and the implementation procedure. Section 5 explains the evaluation results. Section 6 concludes the research work.

II. Literature Review

There have been many different models proposed for prediction of movie ratings. Different papers included different types of data related to the movie to evaluate its outcome on ratings. Yasseri et al. [2] proposes a linear regression model for predicting the box office revenue. The model works well for successful movies, but for not so successful movies the model fails to provide accurate predictions. The movie data set used for this model is very small. The data used for the model [3] is from IMDb movie data and social media data from Twitter and Wikipedia. These features are then used in factorization machines to predict the ratings of the movie. The model gave 0.88 R^2 score. In this study [4], the features are classified into Who, What, and When. A lot of different binary and multi-class classification algorithms are used and all of them are analyzed based on ROI, accuracy, precision, recall and the best is selected.

Dhir and Raj [5], compares several machine learning algorithms - SVM, Random Forest, Ada Boost, Gradient Boost and KNN on data from IMDb. Random Forest gave the best accuracy (83%) in terms of success prediction. Predicting if the movie is a hit or flop [6] using 3 features– IMDb ratings, MusicRatings, and No of Screens. Logistic Regression algorithm is used. The model has achieved an accuracy of 80%.

Magdum and Megha uses sentiment analysis [7] using S-PLSA, and also used ARSA (Autoregressive Sentiment Analysis) model for predicting the sales performance, on the reviews and tweets. The paper [8] proposes to integrate movie data from IMDb with the user response from social media on the movie. This will give more features that can help in making better predictions. This paper [9] tries to find the relation between blogs and movie and music success. The paper could predict the monetary success with 79.84% precision value using Decision Tables. It is vague about how the features were selected from the blogs and how it's used in the algorithms. Santosh et al. [10] aimed in analyzing the effect of hype of the movie on the success of the movie. The hype factor was calculated from number of tweets per second related to the movie, on Twitter. The tweets were taken seven days before the release date of the movie. Real-time sentiment and Twitter data analysis [17], [18] is used in various applications.

This paper [11] has illustrated the use of different regression models, neural networks and classification algorithms to predict the success of Bollywood movies using data from Wikipedia, RadioMirchi and BoxOfficeIndia. The data is normalized before feeding into models. This study [12] not only predicts movie grosses before the release, but also for newly released movies. It says that some of the movies do not do well in the first week but later on it does become successful due to publicity, Oscar nominations and other factors.

Some other studies included prediction of housing prices [13], flight delays [14], crowdsourcing quality [15], etc. using ensemble learning models. For housing price prediction, the paper collected the data from Pune region, extracted the relevant features and used CART Decision Tree and Random Forest Regressor models to perform the analysis. In all of the tests, the RF performed better. Prediction of delay in flights was performed on US flights data after the analysis of features, using Gradient Boosted Decision Trees. Another paper predicted and controlled the quality of crowdsourcing data using Random Forests algorithm, thereby improving the ability and efficiency of the contractor in identifying deceptive participants and take necessary actions.

All of these earlier studies, as mentioned, have tried to either identify the factors that affect the ratings or concentrated on finding a better model for prediction. The studies before on rating prediction had not taken the sentiments of the audience into consideration, before the release of the movie. This study aims at collecting and generating a dataset from different sources that includes the movie's data and social media related data (from YouTube and Wikipedia) before the movie's release and then using this dataset in ensemble learning algorithms to perform predictions of the movie's rating and to compare the performance of the algorithms. Also, this study aims at analyzing the effect of social media data on the predictions. All of these accumulated together is handled by a separate stream of machine learning called Recommendation System (RS) [16].

III. Data Collection Procedure

The most important and basic step for training a model is using an enriched dataset. In this case, it does not exist as such and had to be collected; Our primary objective is to get the data from different platforms. For this, we have used data from IMDB, Wikipedia, and YouTube.

- Movie related metadata: Sources like TMDb and OMDb which expose their APIs to retrieve IMDB movie data were used. To start with these APIs, IMDb ids of the movies had to collected. To get the list of the ids of the movies released in late 2018 or later, an IMDb list of movies was scraped using a script. These ids were then used to get movie related data such as directors, writers, cast, runtime, votes, ratings, etc.

- Wikipedia data: The popularity of movies was collected through Wikimedia REST API for Wikipedia page views of the movie – 30 days prior to the release date of the movie. To retrieve page views from the API, it

required Wikipedia links of the form "Parasite (2019_film)" for the movie Parasite. So these links were first scraped using Google search and then used with the API.

- YouTube data: To understand the buzz created, data from YouTube's API was collected to analyze likes, shares, and comments of the official trailer and teaser. Again, the API needed the video id to get any sort of information on the videos. The YouTube video ids of the trailers and teasers of the movies, which are uploaded before the movie is released, were scraped using Python scripts. And then the comments and other video statistics were collected through API.

Data to be analyzed is collected in the range of – movies released between late 2018 and 2019. The focus is on the movies that are relatively new as the trend and popularity of movies from older times is irrelevant from the trend today.

A. Data Pre-Processing and Transformation/Feature Selection

All the data retrieved from different sources and platforms can be termed as raw data as it lacks the uniformity and contains missing value, hence a wrapper script was prepared that would make a uniform data to include relevant data and discard the rest. Drilling down more into data obtained from each source and their relevance, scripts returning values from TMDb and OMDb are combined to obtain movie metadata which includes attributes such as genre, runtime, directors, actors, budget, etc. And social media data from Wikipedia and YouTube gives us attributes such as page view count, likes, dislikes, comment count, comments, etc. The movie metadata features are given in Table I and social media data is given in Table II.

The comments that were collected were of the trailers and teasers of the movie. Comments from social media cannot be used as such; hence we calculated the sentiment of comments to analyze its impact on the rating of the movie in the future, using Cognitive Services from the Microsoft Azure platform. The sentiment is calculated and categorized into positive, negative and neutral, where each category is given a score from 0 to 1 depending on the sentiment of the comment. We will be dealing with data of 4045 movies with following features.

Additional extracted features (feature selection), mentioned in Table III, were calculated from the data – categorical data was converted to numerical columns using one hot encoding, the popularity of the directors, writers, cast, and production companies among the dataset was calculated by taking the mean of the ratings of the movies they were involved in (using target/mean encoding), and since the release date had large number of values when compared to the number of records, only the month was taken into consideration.

Fig. 1 and Fig. 2 shows the interaction of movie ratings (imdbRating) with the extracted features – mean rating of the director and the ratings of production companies. The correlation matrix of the dataset which gives how each feature is related with other features is given in Fig. 3. Positive values denote higher correlation.

TABLE I.   MOVIE METADATA

Feature Name
Genre
Runtime
Rated
Production Companies
Vote Count
Rating
Adult

TABLE II.   SOCIAL MEDIA DATA

Feature Name
Wikipedia View Count
YouTube video comment count
Video Likes
Video dislikes
Video Comment's sentiment
Video view count

TABLE III.   EXTRACTED FEATURES

| Feature Name	Description
Directors Movie Sum	Average of total movie number made by the directors
Directors Rating Mean	Average of arithmetic means of the ratings of the all movies made by the director
Cast Movie Sum	Average of the total movie numbers in which the first three leading cast played
Cast Rating Mean	Average of the arithmetic means of the ratings of the all movies in which the cast played
Writers Movie Sum	Average of total movie number
Writers Rating Mean	Average of arithmetic means of the ratings of all the movies
Production Company Sum	Average of total number of movies produced by the companies
Production Company Rating Mean	Average of the arithmetic means of the ratings of all the movies produced by each company
Release Month	Release month of the movie extracted from release date
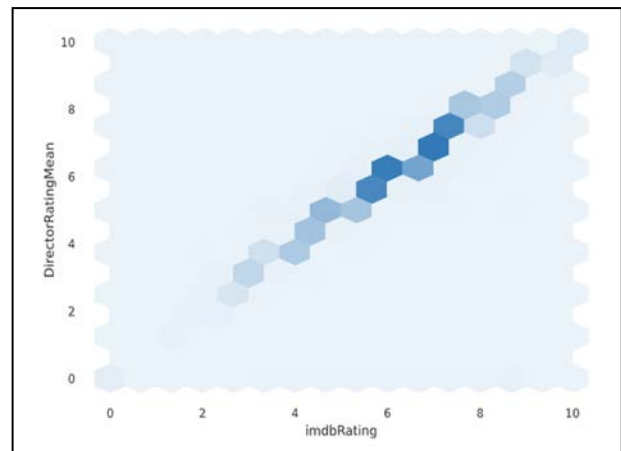
Fig. 1.   Interaction of Director Mean Rating with imdbRating.
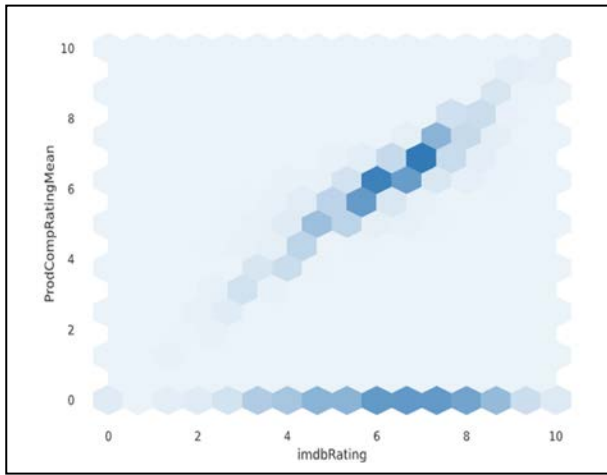
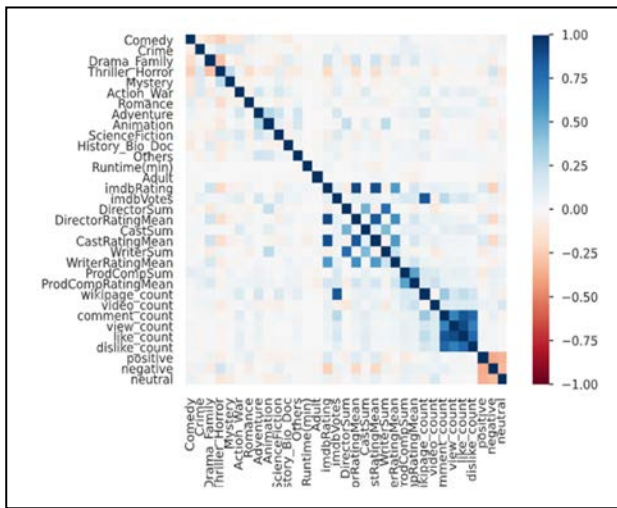Fig. 2.   Interaction of Production Company Ratings with imdb Ratings.

Fig. 3.   Correlation Matrix.

IV. MODEL IMPLEMENTATION USING SCIKIT'S LEARN LIBRARY

In order to explore how social media data affects the prediction accuracy, two datasets were made based on their sources. One was with movie metadata and social media data and the other with only movie metadata without social media data from Table II. Then, random forest and xgboost were implemented using both these sets of data.

After getting the data ready, implementing the algorithm in Python is done using an efficient and widely used package called Scikit-learn. It has provided methods to perform classification and regression tasks. We will be using RandomForestRegressor and XGBRegressor from xgboost class.

A parameter is a configuration variable that is internal to the model and it is very important for finetuning the model. It is also responsible for overall performance of the model and its value can be estimated from the given data. Both of these algorithms have hyperparameters that when set properly, avoids overfitting. Some of these parameters that we have used are given below.

- RandomForestRegressor – n_estimators, max_depth, min_samples_split, min_samples_leaf, random_state.

- XGBRegressor – n_estimators, learning rate, objective, reg_lambda, reg_alpha.

These parameters define how many trees are built for training the model and different attributes for the tree. Also, it defines how many samples are to be picked during the building of a tree (bootstrapping). The learning rate and regularization terms to control overfitting are some of the other parameters.

V. TESTING AND EVALUATION RESULTS

To calculate the performance metrics, 10-fold cross validation was performed using both the algorithms on the training data that was randomly selected from datasets. The performance score was averaged over all the scores for each model and datasets. Different metrics are displayed in the Table IV and Table V.

In a situation where dataset is small, and there is a possibility that the model might overfit the training subset of the data, cross validation uses the whole dataset for training and testing and hence avoiding overfitting to some extent. It also reassures that the model will work well for unseen data.

We also calculated the mean squared error (1); root mean squared error (2) and mean absolute error (3) for each of the models (shown in the Table IV). These functions consider equally - the negative and positive deviation from the target value. But, mean squared error (MSE) penalizes the predicted errors which are further away from the target value more. Whereas, the mean absolute error (MAE) does not penalize based on this deviation. That is, MSE is sensitive to outliers but MAE is robust to outliers.

$$\frac{1}{n}\sum_{t=1}^{n} e_t^2 \tag{1}$$

$$\sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{2}$$

$$\frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{3}$$

TABLE IV.   PERFORMANCE METRICS (MOVIE METADATA + EXTRACTED FEATURES)

Performance Metric	Random Forest	XGBoost
Cross Validation Score	0.894	0.931
MSE	0.28	0.18
RMSE	0.53	0.42
MAE	0.37	0.22

TABLE V.   PERFORMANCE METRICS (MOVIE METADATA + EXTRACTED FEATURES + SOCIAL MEDIA DATA)

Performance Metric	Random Forest	XGBoost
Cross Validation Score	0.907	0.953
MSE	0.16	0.09
RMSE	0.40	0.30
MAE	0.25	0.13

Fig. 4 and Fig. 5 are the plots of predicted values to target values for dataset with social media data. For comprehensibility, only 200 points are displayed in the plots. The red line shows the predicted values and the blue dots display the original or target values. From the plots, it is easily understood that xgboost predicts more accurately than random forest.
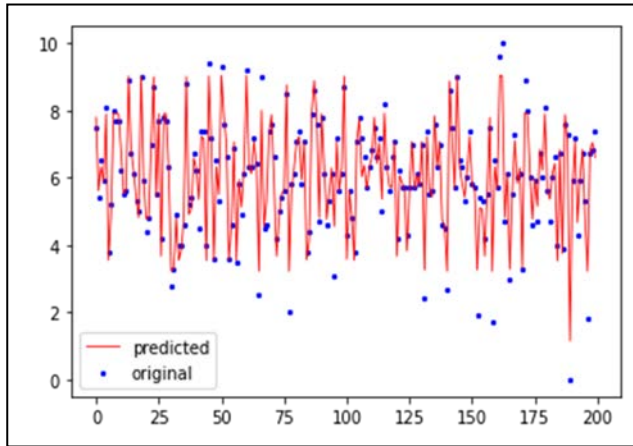

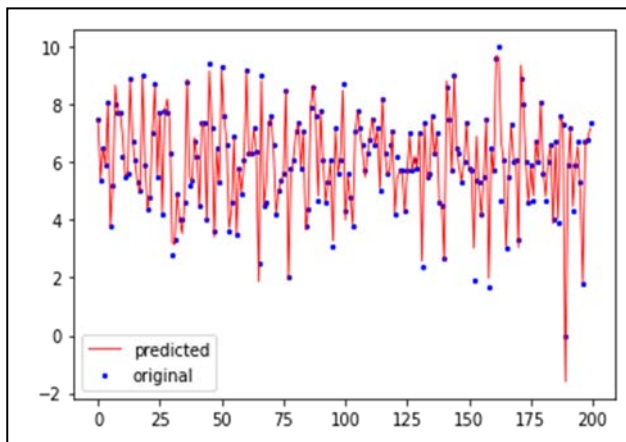Fig. 4.   Target (Original) vs Predicted Plot for Random Forest.

Fig. 5.   Target (Original) vs Predicted Plot for XGBoost.

The implementation results show that XGBoost is the better performer than the Random Forest algorithm for both datasets. And the data that contains both metadata and social media data performs better in predicting the ratings. Though, Random Forest in itself gives a good performance measure, it works well on training and test data.

VI.   Conclusion and Future Scope

In this paper, comparison of two ensemble learners to predict a continuous output was carried out. It was observed that xgboost being a more efficient algorithm performed better when compared to random forest. But the cross-validation score which is 0.95 for xgboost, could lead to overfitting if not properly regularized. There are several hyperparameters that needs to be set to achieve the same. Used movie's metadata with and without its social media data to see how it affected the predictions. From the correlation matrix, it is seen that the popularity of directors, actors and writers affect the ratings most.

For the future work, this can be extended by including a greater number of movies. Since the dataset needed the Wikipedia page counts and YouTube comment sentiments before the movie was released, the dataset in this paper included the movies from only 2019-2020. The data from before this is scarce and would need more preprocessing, if considered.

Further study can also include data from other social media platforms such as Twitter, or blog related features of the movies. This will give the hype being created around the movie and the effect of this can be analyzed on the ratings. Twitter platform gives the popularity of people connected with the movie and how their popularity and social media activity might lead to more people going out to watch their movies and eventually rating it. For example, in this paper the director rating means and cast rating means were extracted from the dataset, these ratings give the popularity of the people amongst the audience based on the movie ratings. The popularity of the cast and director affects the ratings of the movie, as seen from the interaction diagrams discussed. The popularity calculated through twitter followings and retweets can be considered and analyzed for their effect on prediction of movie ratings. This study can also be extended to predict the revenue that would be generated, before a movie is released. The factors that affect the revenue generation will have to be considered for prediction.

References

[1]   A. R. Sulthana and S. Ramasamy, "Context based classification of Reviews using association rule mining, fuzzy logics and ontology," Bulletin of Electrical Engineering and Informatics 6, no. 3, 250-255, 2017.

[2]   T. Y.,. J. K. Márton Mestyán, "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data," PLoS ONE, vol. 8, no. 0071226, 2013.

[3]   S. G. O. Beyza Çizmeci, "Predicting IMDb Ratings of Pre-release Movies with Factorization Machines Using Social Media," in (UBMK '18) 3rd International Conference on Computer Science and Engineering, Turkey, 2018.

[4]   K. Z. Michael Lash, "Early Predictions of Movie Success: the Who, What, and When of Profitability," Journal of Management Information Systems, vol. 33, 2017.

[5]   A. R. Rijul Dhir, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," in International Conference on Secure Cyber Computing and Communication (ICSCCC), India, 2018.

[6]   G. Verma and H. Verma, "Predicting Bollywood Movies Success Using Machine Learning Technique," in 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 2019.

[7]   S. S. Magdum and J. V. Megha, "Mining Online Reviews and Tweets for Predicting Sales Performance and Success of Movies," in International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2017.

[8]   A. Bhave, H. Kulkarni, V. Biramane and P. Kosamkar, "Role of different factors in predicting movie success," in International Conference on Pervasive Computing (ICPC), Pune, India, 2015.

[9]   F. Abel, E. Diaz-Aviles, N. Henze, D. Krause and P. Siehndel, "Analyzing the Blogosphere for Predicting the Success of Music and Movie Products," in International Conference on Advances in Social Networks Analysis and Mining, Odense, Denmark, 2010.

[10] P. K. Ajay Siva Santosh Reddy, "Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining," International Journal of Computer Applications, vol. 56, no. 1, 2012.

[11] A. Kanitkar, "Bollywood Movie Success Prediction using Machine Learning Algorithms," in 3rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018.

[12] I. R. S. Jeffrey S. Simonoff, "Predicting movie grosses: Winners and losers, blockbusters and sleepers," CHANCE, vol. 13, no. 3, 2012.

[13] R. Sawant, Y. Jangid, T. Tiwari, S. Jain and A. Gupta, "Comprehensive Analysis of Housing Price Prediction in Pune Using Multi-Featured Random Forest Approach," in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018.

[14] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta and S. Barman, "A statistical approach to predict flight delay using gradient boosted decision tree," in 2017 International Conference on Computational Intelligence in Data Science(ICCIDS), Chennai, India, 2017.

[15] H. Lan and Y. Pan, "A Crowdsourcing Quality Prediction Model Based on Random Forests," in 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 2019.

[16] A. R. Sulthana and S. Ramasamy, "Ontology and context based recommendation system using Neuro-Fuzzy Classification," in Computers & Electrical Engineering, 74, 498-510, 2019.

[17] Lekha R. Nair, Sujala D. Shetty, Siddhanth D. Shetty, "Streaming Big Data Analysis for Real-Time Sentiment Based Targeted Advertising", International Journal for Electrical and Computer Engineering, Institute of Advanced Engineering and Science , 7.1, pp. 402-407, 2017.

[18] Lekha R. Nair, Dr. Sujala D. Shetty, "Streaming Twitter Data Analysis using SPARK for effective Job Search", Journal of Theoretical and Applied Information Technology (E_ISSN 1817-3195/ISSN 1992-8645).

A Comparative Study of Microservices-based IoT Platforms

Badr El Khalyly[1], Abdessamad Belangour[2], Mouad Banane[3], Allae Erraissi[4]

Laboratory of Information Technology and Modeling
Hassan II University, Faculty of sciences Ben M'sik
Casablanca, Morocco

Abstract—**Internet of Things (IoT) is a set of technologies that aim at fitting together smart devices and applications to build an IoT ecosystem. The target of this kind of ecosystem is to enhance interaction between machines and humans through hardware to software binding while reducing cost and resource consumption. On the application level, IoT ecosystems were implemented by various technologies that all seek better interconnection, monitoring, and controlling of IoT smart devices. Among recent technologies, Microservices, which are a variant of the service-oriented architecture, are subject to great excitement. In fact, Microservices are an emerging technology built around Microservice paradigm which goal is to offer services with a small granularity, which exactly meets the distributed nature of IoT devices while maintaining a loosely coupled architecture between IoT components among other advantages. Efforts to build Microservice-based IoT platforms sooner emerged to take advantage of the numerous benefits of the Microservice paradigm to build scalable, interoperable, and dynamic ecosystems. The goal of this paper is to list these approaches, classify them and compare them using a Weighted Scoring Model (WSM) method. This involves, studying these platforms, establishing relevant criteria for comparison, assigning weights for each criterion, and finally calculating scores. The obtained results reveal the weaknesses and strengths of each of the studied platforms.**

Keywords—IoT platforms; microservices; WSM method; IoT architecture; smart devices

I. INTRODUCTION

Internet of things is gaining widespread interest and popularity across almost everyday life domains ranging from smart homes, through smart cars or smart factories to smart cities. On the application level, which is our focus here, propositions and solutions are varying according to the interest point of research. Some have been dealing with domain-centric concerns while others are of general architecture purpose.

Among general-purpose architectures, there are Microservices based IoT platforms. Microservices are the next generation of service-oriented architecture designed with the idea of avoiding the hell of monolithic services through diminishing the size of exposed services and characterized with [1] (Componentization via Services, Organization around Business Capabilities, Providing of products, not Projects, Providing of smart endpoints and dumb pipes, Decentralization of Governance, Decentralization of Data Management, Automation of Infrastructure, Design for failure, Evolutionary Design).

The nature of Microservices matches exactly the needs and the requirements of IoT. Hence, many IoT platforms adopt Microservices for their implementations. While some knew how to maximize profiting from Microservices, others did less. Thereby, this paper proposes to make a comparative study of these approaches based on the Weighted Scoring Model (WSM) method. This paper starts our comparative study by extracting relevant criteria for comparison and provides a short description of each criterion (Section 4). These criteria are gathered from related work comparative studies (Section 2) together with other criteria they missed. Then, there is an application of the WSM method, which requires weight attribution to get final scores for each Microservices based IoT platform (Section 3). The obtained scores are represented using a spider chart and are well discussed (Section 5).

II. RELATED WORK

Many scientific works have as purpose to build an Internet of Things platforms and applications that are based on Microservices. Many researchers have made scientific efforts in this optic to implement the architecture style of Microservices. The authors propose solutions in different domains such as Smart City, Industry, Smart Vehicle, logistics, agriculture. The designs that are adopted in these approaches have been made considering functional and non-functional requirements. Functional requirements means a set of features that related to the domain to which the application is built. Non-functional requirements are a set of technical features that are mandatories in the application like (database management, authentication), and other options that aim to increase quality applications like (historical data, Data classifier).

Several works compare Internet of Thing platforms [2] [3] [4] [4] [5] [6] [7] [8]. All those comparative studies compared monolithic approaches of Internet of Things platforms. The advent of Microservice architecture style pushed the designer to propose many Microservice based Internet of Things platforms and solutions. However, it is relevant to have a study that assembles feature applications and demonstrates their different options and the domain of the function. In this work, there is a comparison of Microservice based Internet of Things efforts. This work is beginning by listing a set of criteria, classify the above listed platforms, and assign to each platform a value that corresponds to a specific criterion. Those criteria represent the features that characterize the platforms. Then, there is a proposition to evaluate other criteria like Interoperability–Customizability–Devops–Domain–Virtual

Objects. This work resides as well in the comparison between these platforms following the criteria. The Weighted Scoring Model is adopted to carry out this comparison, and get the work that is more favored for the most criteria and discuss the result of this comparison, and then this paper provides an opinion to increase the quality of each platform in future works.

Currently, the majority of IoT applications are deployed in the Cloud [9]. This method satisfies the need for IoT in terms of ubiquitous access, high availability, and scalability of computing performance and storage capacity. Also, the Cloud frees users from the specification of many details (eg platforms, dependencies, etc.) and simplifies the deployment and integration of services [10,11]. However, the cloud paradigm is a centralized IT model. This means that all data and requests must be transmitted to cloud data centers which represent centralized points in the heart of the network [12]. With the increase in the number of objects and the amount of data produced, this solution suffers from many problems. Transferring data from network edges to data centers can cause bottlenecks and high and unpredictable latencies. Thus, the QoS will be degraded [13]. Besides, many IoT applications are sensitive to latency, such as telemedicine, the Internet of vehicles, and security (e.g. intrusion detection) [9]. In addition, certain decisions can be made locally (i.e. at the edge of the network), without having to be transmitted to the Cloud [12]. On the other hand, several IoT applications require mobility support, geo-distributed deployment, and localization knowledge that the Cloud cannot efficiently provide [14].

In Chafle et al. [15], the proposed approach aims to adapt defined service processes for composition and during execution by selecting web services as well as by adapting the abstract process to a concrete service process. The approach dissociates the functional and non-functional aspects in different prerequisites used. Thus, the approach makes it possible to react to changes at runtime based on abstract process templates and a directory of services. In Mosincat et al. [16] the proposed approach works on the selection of services for the maintenance of execution plans formalized in BPEL. To do this, the approach uses light execution monitoring systems to detect performance declines and to look for other services to be executed in the event of a fault in order to repair the execution. In Yu and [17], the approach considered works on matching services in terms of QoS needs in order to find services corresponding to the need formulated in particular in terms of QoS parameters with several parameters considered. In Guo et al. [18] an approach is proposed in order to carry out service selection on a larger scale over large areas of candidate services. The service selection process is thus divided into different stages involving a covering algorithm as well as a swarm particle algorithm using Spark. The spanning algorithm is used to reduce the initial search space and the particle algorithm is used to optimize the service selection in relation to the QoS parameters considered. In Guidara et al. [19,20], the proposed approach allows a selection of dynamic services for composition. The goal is to carry out selection actions dynamically to avoid major changes and mistakes in the execution of a pre-established plan. In addition, heuristics are proposed in order to improve the service selection process in

the case of QoS constraints and also global and local time constraints. This process allows a cleaning of the space of the candidate services according to the global constraints and a better selection of services to obtain a quasi-optimal solution In Chattopadhyay et al. [21], the proposed approach works to realize an automatic selection of services web. First, an optimal solution search algorithm is proposed to build a solution satisfying the constraints.

QoS of the multi-objective optimization problem posed. Two heuristics are also proposed based on a beam search strategy [22] and a non-deterministic genetic sorting strategy. In Alsaryrah et al. [23], the proposed approach makes it possible to carry out an optimized path search among a set of candidate services for each task to be performed. The advantage of this approach is that it takes into account the energy consumed by the services which can be linked to connected objects. The approach also makes it possible to seek to optimize several parameters by normalizing these and by using a k-shortest path search algorithm. In Gronvall et al. [24], the proposed approach is based on a lightweight framework (SECREST) for composing REST services via the use of user guidance in a care network scenario. The composition is based on service templates that are instantiated into services according to the need. In Guinard et al. [25], the proposed approach aims to integrate REST services provided by connected objects into existing service architectures by integrating these services into existing service directories and by enabling the discovery of services. Other approaches that are based on the semantic Web such as: In [26], the proposed approach aims to achieve being able to determine if services correspond, in particular in terms of inputs/outputs based on semantic modeling of web services. In Kim et al. [27], the proposed platform makes it possible to perform service orchestration based on a user-centered approach as well as an ontology model in order to be able to orchestrate web services and connected objects. However, the management of QoS parameters is not integrated. In [28,29], the authors propose a system for managing semantic data in a Cloud / Big Data context, which can be easily used with a Spark query tool [30,31].

In Fähndrich et al. [32], the proposed approach makes it possible to find a correspondence in a specified process (BPMN) by finding the appropriate services using a semantic engine (based on OWL-S). A system has been designed to perform various stages from the specification of the process to the execution of the services. QoS parameters are not considered, however.

III. BACKGROUND

A. IoT

Internet of Things environments (i.e., IoT: Internet of Things) are becoming more and more present in our contemporary societies. This paradigm was born from the idea of interconnecting different types of objects with each other through the Internet in order to improve the quality of life for humans. Several organizations have taken an interest in the IoT in order to define its different components according to various layered architectures but also the concepts and technologies which make it possible to deploy it.

The IoT is no longer a science fiction subject but evidence in our everyday life. Currently, a human has on average 2 connected digital or electronic devices enabling him to offer intelligent services [33]. The integration of IoT into the daily life of humans makes it possible to offer new applications improving the quality of life through the interconnection of objects. IoT is causing a revolution in the relationship of humans with technology. This revolution concerns various sectors of daily life inducing a change in habits and behaviors. Thus, the IoT has enabled the creation of connected refrigerators, connected shoes, connected electric meters, connected cameras, etc. In this context, the management of our main residence can be carried out remotely and even autonomously and intelligently. This remote and intelligent management can make human life more comfortable, easier, and even safer. Indeed, even critical areas, such as the medical field, are impacted by the integration of IoT technologies by connecting all types of objects. From now on, doctors will be able to diagnose their patients remotely, using sensors worn or implanted on patients. This technological advance improves vital services by optimizing performance and improving the quality of these services.

The IoT is responsible for the collection and/or creation of a large volume of data. This enormous volume of data, known by the term "Big Data" [34,35], allows, on the one hand, to have an incredible wealth in terms of information allowing the offer of advanced services. On the other hand, this volume of data creates new challenges to be considered such as securing, processing, and real-time accessibility of this data [36,37].

The IEEE, through the Webinar "Integrating the IoT and Cultural Heritage in the Smart City", proposed architecture for IoT based on four layers (Interface layer, Service Logic layer, Data layer, and Resources layer) [38], Fig. 1. Each layer includes specific entities and elements to provide various functionality.
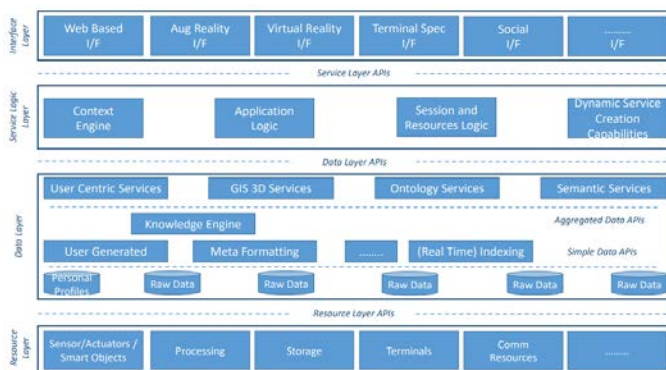

Fig. 1.   Internet of Things Architecture of IEEE [38].

B. Microservices

The first to describe architectures based on microservices are J. Lewis and M. Fowler [39]. This software architecture is a way of dividing a monolithic application into a set of small programs having only one precise task to execute (the service) and also being able to communicate with the other services using a protocol; common to all (usually an HTTP REST API). Microservice-based architecture can be seen as an extension of

Service Oriented Architecture (SOA) [40], but with the addition of additional concepts, such as direct communication between services. Microservices are currently mainly used for Cloud applications and Internet of Things (IoT). Much of the literature, such as [41], recognizes many advantages to microservices, such as simpler development and maintenance compared to its monolithic equivalent, with for example a dedicated development team per microservice. This is because, over time, a monolithic application will grow and become more and more complex with the addition of new functionalities to the point of being complex and difficult to understand for a newcomer. In addition, this complexity and the historical code can also block the technologies used because of the changes that are too deep (a good example is banking applications, some of which are still in COBOL). Microservices, while remaining simple, do not have this problem and thus allow them to be more reactive to new technologies. Also, due to the relative simplicity of each service, an application based on microservices can be more easily scaled because only the limiting elements will need to be replicated instead of the complete application. But microservices do not only offer advantages, among the most significant drawbacks, for example, more complex application deployment and the need to implement a communication protocol between microservices to compose them.

C. Microservice-based IoT Platforms

Many scientific research efforts proposed Microservice based internet of things implementation. These efforts can be classified into seven categories or domains that are:

- Smart Cities: A Smart city is an ecosystem that is rolled out in an urban emplacement, and uses different technologies to collect and process data. The gathering data is analyzed to control and manage traffic and transportation systems, energies, administration services, waste management, and fraud detection. In the literature, there is InterSCity [42], and DIMMER [43], InterSCity doesn't suggest functional services like DIMMER.

- Smart Commerce: It is a restricted ecosystem that is implemented on the fog level. It is a local system that is rolled out in a set of local devices and servers, and meet the personal need in the stores. Smart commerce provides automation services for customers in the stores. In [51] researchers proposed a solution that is based on Microservices and blockchains and is deployed on AWS. The solution developers adopt shadows that are abstractions of physical objects and provided by AWS.

- Smart Farms: It is a system that is rolled out in the farms to help farmers to gain time and improve the quality of farm services. Author in [50] described a solution that is destined to the agriculture domain and is based on Microservices. They deployed their solution on three-level: Fog, Edge, Cloud.

- Smart Cars: It is a set of services that aim to automate car components. Author in [49] talked about a Testbed

Architecture using Web of Thing. Their architecture is based on Microservices.

- Smart Factories: It is an ecosystem that is implemented in the factories that automate the production chain. Kontou [44] proposed a solution for Evolvable assembly systems that are based on Microservices. They described a solution on 3 levels: Edge, Fog, and Cloud. They used an ontology-based approach to organize the assembly process as Microservice.

- Smart Logistics: It is a set of services that are rolled out to process data to improve the supply chain quality. Author in [45] described an approach for the transportation planning that is based on Microservices and deployed on the cloud. They described its life cycle as well in experimental validation. Author in [47] proposed a solution for the safe transportation of dangerous products. This solution is based on Microservices that are organized following functional and non-functional requirements.

- Cross-domain or general-purpose: It is a generic system that is designed to any kind of Internet of Things ecosystem. Author in [46] proposed a four-layer model for deploying IoT Microservices in fog computing. In [48] authors adopted the concept of virtual object and web of the object that is an abstraction of a physical object, then they proposed Web of Objects Architecture for IoT Service Provisioning, and implemented a Functional Model of Microservices in Web of Objects Platform. In [52] designed a framework that is composed of four Microservices that contribute to providing Fault tolerance support.

IV. COMPARISON OF MICROSERVICE-BASED IoT PLATFORM

Weighted Scoring Model is a methodology that helps choosing between compared objects based on a set of criteria [20].

To compare the works that are cited below in the Section 3, the WSM Method is applied [53]. The following steps can perform the application of this method:

- Step 1: Selection of criteria that constitutes platform features.

- Step 2: Assigning weights to group of criteria based on their importance.

- Step 3: Elaboration of a table that contains nominal values of each distribution for each criterion.

- Step 4: Elaboration of a table that contains weights for each criterion. The weight is a percentage. The weight total equals 100%.

- Step 5: For each platform, a calculation of the product score of weights and nominal values is elaborated.

A. Comparison Criteria

The choice of comparison criteria is based on the set of Internet of Things studies. Some criteria from the works that are cited below are extracted. Those criteria are mostly the common characteristic that is shared between Internet of Things platforms:

- Data Management: This criterion shows if the platform can manage the gathering data obtained by things. This criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion for such a distribution.

- Data Processing: This criterion evaluates if the platform can process data gathering data obtained by things. This criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion for such a distribution.

- Internet of Things Integration: This criterion shows if the platform provides a hub or gateway that allows to bind things to the platform. This criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion on this distribution.

- Context Awareness: This criterion shows if the platform provides intelligence to the bound things. This criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion for such a distribution.

- Specific-domain Services: This criterion if the platform offers services that help facilitates life in the smart city. This criterion is evaluated by calculating the number of services that are specific to the domain to which the platform is built.

- Cloud: This criterion shows if the platform uses the services and features of the cloud. This criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion on the distribution.

- Scalability: This criterion shows if the platform is scalable. This criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion on the distribution.

- Security: This criterion shows if the platform is safe for users, this criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion on the distribution.

- Privacy: This criterion shows if the platform protected user data, this criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion on the distribution.

- Adaptation: This criterion shows if the platforms can be adapted to updates and changes. This criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion on the distribution.

- Experimental Evaluation: This criterion shows if the platform is well evaluated on the real world; this criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion on the distribution.

- Free/Open Source: This criterion shows if other users can access and use the solution freely. This criterion is evaluated in Boolean value that demonstrates the availability of this feature/criterion on the distribution.

- Customizability: This criterion shows if other users can extend the existing code of the platform and develop new features.

- Agile Method [58]: This criterion shows if the designers respect agile methods.

- Virtual Object [57]: This criterion shows if the platforms provides an abstraction of things called Virtual Object.

- Technologies: This criterion shows the technologies that are used to build the platform.

- Functional Features: This criterion shows if the platform has features that do not belong to the technical scope.

- Syntactic Interoperability [54]: This criterion shows if the platform supports heterogeneous actors that communicate using different languages.

- Semantic Interoperability [55]: This criterion shows if the platform supports can exchange data with shared meaning.

- Cross-Platform Interoperability [56]: This criterion shows if the platform can exchange data with external platforms.

- Devops [59]: This criterion demonstrates if the platform respects the technologies regarding continuous development and integration.

- Domain: This criterion shows the number of domains that are covered by the platform.

B. Comparison Study

Table I of nominal values is carried out. For each criterion, the value that corresponds to each platform is assigned. These values are extracted from studies that are cited in the next section.

C. Application of Weighted Scoring Model

Following the application of WSM approach, the table below determines the score of each Microservice based Internet of Things platform. The assignment of weight percentages is realized following the importance of the criterion. Because of their mandatory requirement, the priority is given to this set of criteria: Data management – Internet of Things Integration – Context-awareness – Adaptation – Domain – Security - Scalability - Data Processing. To each criterion of them a weight of 6% is assigned. The second category of importance is given to the criteria: Cloud – Privacy. To each criterion of them a weight of 5% is assigned. The third category of importance is given to the criteria: - Experimental Evaluation - Virtual Object – Technologies - Syntactic Interoperability - Semantic Interoperability - Cross-Platform Interoperability. To each criterion of them a weight of 4% is assigned. This set of criteria is not such a big importance: Specific-domain Services - Functional Features - Free/Open Source – Customizability - Agile Method – Devops. To each criterion of them a weight of 3% is assigned. The total weight equals 100%. Table II shows the result.

TABLE I.   Nominal Values of Criteria Corresponding to the Platform

	InterScity [42]	Dimmer [43]	A Cyber-Physical Microservices An IoT-based Framework for Manufacturing Systems [44]	Smart ITS sensor for the transportation [45]	Leveraging Microservice Architecture for Next-Generation Iot Applications [46]	Real-Time HazMat Environmental Information System [47]	Microservices in web objects enabled IoT environment for enhancing reusability [48]	IoT and Microservices Based Testbed for Connected Car Services [49]	SmartHerd management [50]	Block chain-based IoT Platform [51]	Microservices Architecture for Reactive and Proactive Fault Tolerance [52]
Data Management	1	1	1	1	n/a	1	1	n/a	1	1	n/a
Data Processing	1	1	1	1	1	1	1	1	1	n/a	1
IoT Integration	1	1	1	1	1	1	1	1	1	1	1
Context Awareness	1	1	2	n/a	n/a	n/a	1	1	n/a	n/a	n/a
Specific-domain Services	n/a	BIM Services, GIS Services, SIM Services	n/a	n/a	n/a	GIS Services– Regulation– Alert – Transport Document- Atmosphere dispertion	n/a	Vehicle data processing	Heat detection, SmartAgri Page, Weather Data, Livestock mobility Monitorin	Shop Service,	n/a

									g, Animal welfare		
Cloud	1	0	n/a	n/a	n/a	1	1	1	1	1	1
Scalability	1	1	n/a	n/a		n/a	1	1	n/a	1	1
Security	0	0	n/a	n/a		n/a	n/a	1	n/a	1	n/a
Privacy	0	0	n/a	n/a		n/a	n/a	n/a	n/a	1	n/a
Adaptation	1	1	1	n/a	n/a	n/a	n/a	n/a	1	n/a	n/a
Evolvability	1	1	1	1	1	1	n/a	n/a	1	n/a	n/a
Experimental Evaluation	1	1	n/a	1	1	1	1	1	1	1	n/a
Free/Open Source	1	1	n/a	n/a	n/a	n/a	n/a	1	n/a	n/a	n/a
Customizability	1	0	1	n/a	1	n/a	1	n/a	n/a	n/a	n/a
Agile Method	1	1	n/a	n/a	1	1	1	n/a	n/a	1	n/a
Virtual Object	1	0	n/a	n/a	n/a	1	1	n/a	n/a	n/a	n/a
Technologies	PostGres - RabbitMq - Ocean cloud - Docker - REST- Redis- Ruby [6]	LinkSmart – SPARQL- MQTT	CPMS - UML4IoT– MDE – RDF - OWL	REST Python C++ - Shell	Devops – HTTP – JAVASCRIPT – SSL – Azure	MQTT, Node, MongoDB, PostgreSQ, Qgis server, Angular,	JSON, XML, RDF, HTTP, REST	JWT, ITS-G5, NodeJs, Mongo DB, W3C	NoSQL, MQTT, Java, Docker, Kubernetes, IBM, Fog Computing, Edge Computing, 4G, WIFI.	Block Chain, EVM, JSON, Web3.js, AWS, MQTT, RPC, NAT	CEP, MQTT, REST
Functional Features	Resource Viewer - Data Collector - Resource Catalog - Actuator Controller- Resource Adaptor	Energy simulator - Energy efficiency - Context awareness- Historical storage - Semantic storage - Resource catalog - Message broker - Service catalog	n/a	n/a	n/a	Data Collection, Hazmat	Microservice Lookup & Discovery Task, Microservice management task, Microservice creation Task, Object Reusing Microservices, Microservice Listener, Authentication, Template Database, Registry Database	Experiment Submission, Report generation, New Service Addition, Discovery, TD Registration, Protocol Binding	Data Acquisition, Data classification, MQTT Publisher, Service Registry, Herd Classification, Alert Service, Local Visualization Webpage, Availability Monitoring, Save Data, NoSQL	Registration, Encryption, Rule Engines,	Predictive Fault Tolerance, Real-time Fault Tolerance
Syntactic Interoperability	1	1	1	n/a	n/a	n/a	1	1	n/a	n/a	1
Semantic interoperability	1	1	1	n/a	n/a	n/a	1	1	n/a	n/a	n/a
Cross-Platform Interoperability	1	1	1	n/a	n/a	n/a	1	n/a	n/a	n/a	n/a
Devops	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	1	n/a	
Domains	Smart city	Smart city	Industry	Logistic		Logistic		Vehicle	Agriculture	Commerce	

TABLE III.   TABLE OF WSM RESULTS

Criterion	Weight	Requirement Score										
		[42]	[43]	[44]	[45]	[46]	[47]	[48]	[49]	[50]	[51]	[52]
Data Management	6	6	6	6	6	0	6	6	0	6	6	0
Data Processing	5	6	6	6	6	6	6	6	6	6	0	6
Internet of Things Integration	6	6	6	6	6	6	6	6	6	6	6	6
Context Awareness	6	6	6	6	0	0	0	6	6	0	0	0
Specific-domain Services	3	0	9	0	0	0	15	0	3	18	3	0
Cloud	5	5	0	0	0	0	5	5	5	5	5	5
Scalability	4	6	6	0	0	0	0	6	6	0	6	6
Security	5	0	0	0	0	0	0	0	6	0	6	0
Privacy	5	0	0	0	0	0	0	0	0	0	5	0
Adaptation	6	6	6	6	0	0	0	0	0	6	0	0
Experimental Evaluation	4	4	4	0	4	4	4	4	4	4	4	0
Free/Open Source	3	4	4	0	0	0	0	0	4	0	0	0
Customizability	3	4	0	4	0	4	0	4	0	0	0	0
Agile Method	3	4	4	0	0	4	4	4	0	0	4	0
Virtual Object	4	4	0	0	0	0	4	4	0	0	0	0
Technologies	4	28	12	20	16	20	24	20	20	40	32	12
Functional Features	3	15	24	0	0	0	6	24	18	30	9	6
Syntaxic Interoperability	4	3	3	3	0	0	0	3	3	0	0	3
Semantic Interoperability	4	3	3	3	0	0	0	3	3	0	0	0
Cross-Platform Interoperability	4	3	3	3	0	0	0	3	0	0	0	0
Devops	3	0	0	0	0	0	0	0	0	3	0	0
Domain	6	6	6	30	6	30	6	30	6	6	6	30
Score	100	119	108	93	44	74	86	134	96	130	92	74

V. DISCUSSION

According to the previous results, the work [48] is the most favored Microservice based platform. Most of the works concentrate their efforts in Data Management, Data Processing, and Agile Methods, which is relevant to a client who needs a completed product without the need to modify it. There is a considerable richness in most of the efforts. All the efforts implement IoT Integration Microservices. This result is reflected in the Multicriteria Radar Graph shown in Fig. 2.

Only [42] [43] [49] are Open Source. A considerable effort must be made in Evolvability, Scalability, and Customizability. The customizability is important to allow other stakeholders to customize the Microservices following the needs. Context-awareness and Interoperability still a feature that encounters shortcomings. Many works do not take into consideration the Interoperability factor to connect and bind devices that work with heterogeneous protocols. There is a big lack of security, privacy, and Devops method that must be taken into consideration by researchers.
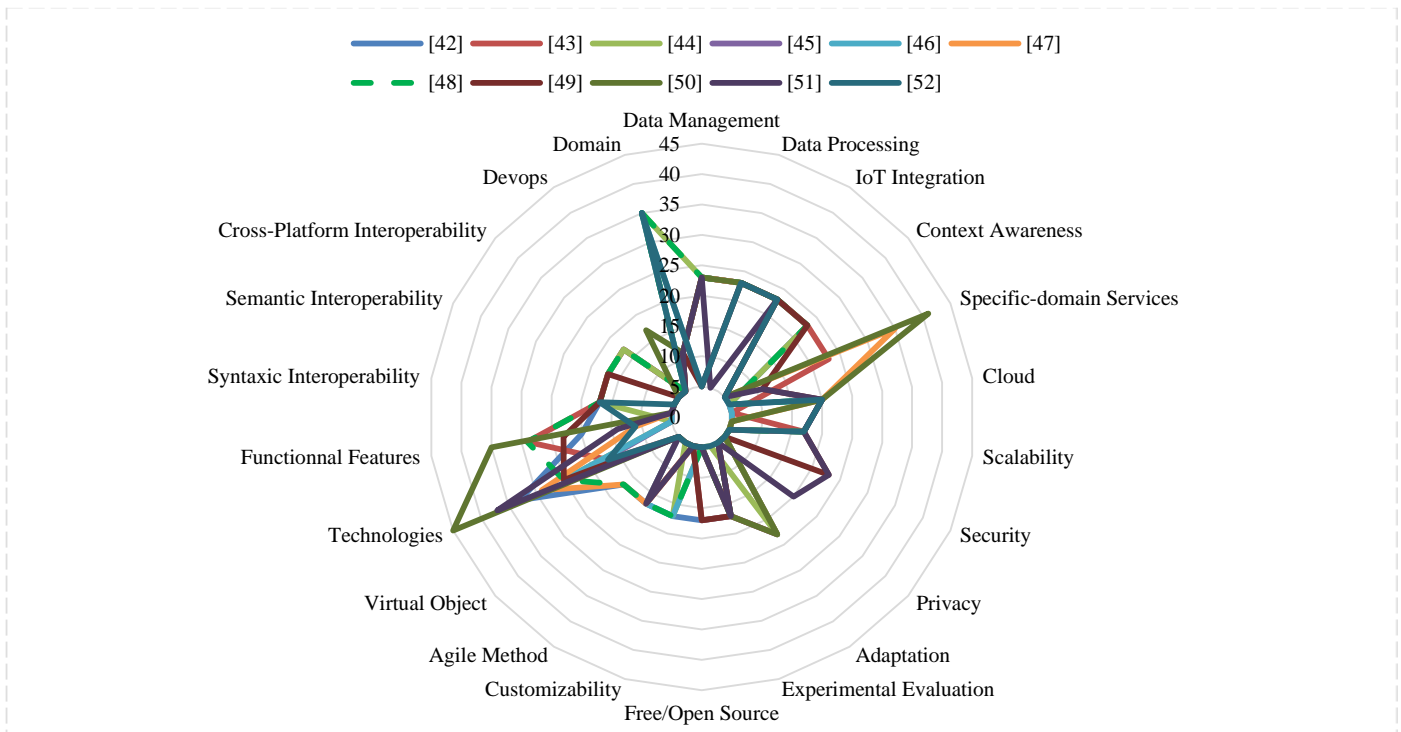
Fig. 2.   Multicriteria Spider Graph.

VI. Conclusion

In this paper, a comparative study of the Microservice based IoT platforms is presented. This work is started by identifying a set of relevant works that adopt Microservices and propose a platform that meets the need of different domains. Then a set of domains that are concerned by these works is enumerated. A set of criteria is identified as well upon which this comparison is based. Based on Weight Score Model method the scores for each of the studied platforms is obtained. These scores or results helped us establishing a general ranking between these platforms but also showed their inner strengths and weaknesses regarding each studied criterion. This study allows recognizing a set of fails that the Microservice based platform suffers from. The main problems that platforms face are contexted awareness and interoperability. The companies have to develop a platform that takes into consideration context awareness, so that they allow the ecosystem to have an intelligence aspect, such as voice recognition, biometrical identification, and another kind of acknowledgment. The developers have to develop an interoperable system that connects heterogeneous devices that use different communication protocols and work in different ecosystem domains. Scalability is a primordial factor that must be the object of scientific research to improve Microservice based Internet of Things platform. Lastly, the researchers and the designers have to adopt Devops culture, which is a set of tools that help improving product quality and reducing development costs, with providing continuous delivery and integration for IoT Designers and Developers.

References

[1] Martin Fowler and James Lewis. Microservices, 2014. http://martinfowler.com/articles/microservices.html.

[2] Anthopoulos LG, Janssen M, Weerakkody V (2015) Comparing Smart Cities with different modeling approaches. Association for Computing Machinery (ACM), pp 525–528.

[3] Fernández-Vázquez A, López-Forniés I Analysis and comparison of smart cities initiatives.

[4] Nakhuva B, Champaneria PT (2015) STUDY OF VARIOUS INTERNET OF THINGS PLATFORMS. Int J Comput Sci Eng Surv 6:. https://doi.org/10.5121/ijcses.2015.6605.

[5] Solapure SS, Kenchannavar H (2016) Internet of Things: A survey related to various recent architectures and platforms available. In: 2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016. Institute of Electrical and Electronics Engineers Inc., pp 2296–2301.

[6] Guth J, Breitenbücher U, Falkenthal M, et al (2016) Comparison of IoT Platform Architectures: A Field Study based on a Reference Architecture. https://doi.org/10.1109/CIOT.2016.7872918.

[7] Torán D Comparison of different Internet of Things platforms.

[8] Salami A, Yari A (2018) A framework for comparing quantitative and qualitative criteria of IoT platforms. In: 2018 4th International Conference on Web Research, ICWR 2018. Institute of Electrical and Electronics Engineers Inc., pp 34–39.

[9] H. Gupta, A. Vahid Dastjerdi, S.K. Ghosh, and R. Buyya. iFogSim : A Toolkit for Modeling and Simulation of Resource Management Techniques in Internet of Things, Edge and Fog Computing Environments. ArXiv eprints, June 2016.

[10] H. Atlam, R. Walters, and G. Wills. Fog computing and the internet of things : A review. Big Data and Cognitive Computing, 2(2), 2018. ISSN 2504- 2289. doi : 10.3390/bdcc2020010. URL http://www.mdpi.com/2504-2289/ 2/2/10.

[11] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. Fog computing and its role in the internet of things. In Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, MCC '12, pages 13–16, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1519-7. doi : 10.1145/2342509.2342513. URL http://doi.acm.org/10.1145/2342509. 2342513.

[12] P. Hu, S. Dhelim, H. Ning, and T. Qiu. Survey on fog computing : architecture, key technologies, applications and open issues. Journal of Network and Computer Applications, 98 :27 – 42, 2017. ISSN 1084-

8045. doi : https:// doi.org/10.1016/j.jnca.2017.09.002. URL http://www.sciencedirect.com/ science/article/pii/S1084804517302953.

[13] S. Sarkar, S. Chatterjee, and S. Misra. Assessment of the suitability of fog computing in the context of internet of things. IEEE Transactions on Cloud Computing, PP(99) :46–59, 2015. ISSN 2168-7161. doi : 10.1109/TCC.2015. 2485206.

[14] S. Yi, C. Li, and Q. Li. A survey of fog computing : concepts, applications and issues. In Proceedings of the 2015 Workshop on Mobile Big Data, pages 37–42. ACM, 2015.

[15] Girish Chafle, Koustuv Dasgupta, Arun Kumar, Sumit Mittal et Biplav Srivastava. Adaptation in Web Service composition and execution. Proceedings - ICWS 2006 : 2006 IEEE International Conference on Web Services, pages 549–557, 2006.

[16] Adina Mosincat et Walter Binder. Automated maintenance of service compositions with SLA violation detection and dynamic binding. International Journal on Software Tools for Technology Transfer, vol. 13, no. 2, pages 167–179, 4 2011.

[17] Shih Yuan Yu, Chi Sheng Shih, Jane Yung Jen Hsu, Zhenqiu Huang et Kwei Jay Lin. QoS oriented sensor selection in IoT system. Proceedings - 2014 IEEE International Conference on Internet of Things, iThings 2014, 2014 IEEE International Conference on Green Computing and Communications, GreenCom 2014 and 2014 IEEE International Conference on CyberPhysical-Social Computing, CPS 20, no. iThings, pages 201–206, 2014.

[18] Xing Guo, Shanshan Chen, Yiwen Zhang et Wei Li. Service Composition Optimization Method Based on Parallel Particle Swarm Algorithm on Spark. Security and Communication Networks, vol. 2017, no. 1, pages 1–9, 2017.

[19] Ikbel Guidara, Imane Al Jaouhari et Nawal Guermouche. Dynamic Selection for Service Composition Based on Temporal and QoS Constraints. Dans 2016 IEEE International Conference on Services Computing (SCC), pages 267–274. IEEE, 6 2016.

[20] Ikbel Guidara, Nawal Guermouche, Tarak Chaari, Ikbel Guidara et Nawal Guermouche. Heuristic based Time-aware Service Selection Approach. 2017.

[21] Soumi Chattopadhyay et Ansuman Banerjee. QoS aware Automatic Web Service Composition with Multiple objectives. pages 1–15, 2018.

[22] Christoph Tillmann et Hermann Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. Computational linguistics, vol. 29, no. 1, pages 97–133, 2003.

[23] Osama Alsaryrah, Ibrahim Mashal et Tein Yaw Chung. Energyaware services composition for Internet of Things. IEEE World Forum on Internet of Things, WF-IoT 2018 - Proceedings, vol. 2018-Janua, pages 604– 608, 2018.

[24] E Gronvall, M Ingstrup, M Ploger et M Rasmussen. REST based service composition : Exemplified in a care network scenario. Visual Languages and Human-Centric Computing (VL/HCC), 2011 IEEE Symposium on, pages 251–252, 2011.

[25] Dominique Guinard, Vlad Trifa, Stamatis Karnouskos, Patrik Spiess et Domnic Savio. Interacting with the SOA-Based Internet of Things : Discovery, Query, Selection, and On-Demand Provisioning of Web Services. IEEE Transactions on Services Computing, vol. 3, no. 3, pages 223–235, 7. 2010.

[26] Z Wu, AH Ranabahu et K Gomadam. Automatic composition of semantic web services using process and data mediation. 2007.

[27] Youngjun Kim, Sanghum Lee et Ilyoung Chong. Orchestration in Distributed Web-of-Objects for Creation of User-Centered IoT Service Capability. Wireless Personal Communications, vol. 78, no. 4, pages 1965–1980, 2014.

[28] BANANE, Mouad, ERRAISSI, Allae, et BELANGOUR, Abdessamad. SPARQL2Hive: An approach to processing SPARQL queries on Hive based on meta-models. In : 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO). IEEE, 2019. p. 1-5.

[29] BANANE, Mouad et BELANGOUR, Abdessamad. Towards a New Scalable Big Data System Semantic Web Applied on Mobile Learning. International Journal of Interactive Mobile Technologies (iJIM), 2020, vol. 14, no 01, p. 126-140.

[30] BANANE, Mouad et BELANGOUR: A Big Data Solution To Process Semantic Web Data Using The Model Driven Engineering Approach, International Journal of Scientific & Technology Research. vol. 9, no 02, 2020.

[31] Banane, Mouad, and Abdessamad Belangour. "A new system for massive RDF data management using Big Data query languages Pig, Hive, and Spark." International Journal of Computing and Digital Systems 9.2 (2020): 259-270.

[32] Johannes Fähndrich, Tobias Küster, F Johannes, K Tobias et Nils Masuch. Semantic Service Management and Orchestration for Adaptive and Evolving Processes. International Journal on Advances in Internet Technology, no. January, 2016.

[33] A. Nordrum, « Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated », IEEE Spectrum, 2016. Disponible sur : https://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-thingsforecast-of-50-billion-devices-by-2020-is-outdated.

[34] A. Erraissi, B. Mouad and A. Belangour, "A Big Data visualization layer meta-model proposition," 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain, 2019, pp. 1-5. doi: 10.1109/ICMSAO.2019.8880276.

[35] Erraissi Allae, et Abdessamad Belangour. « A Big Data Security Layer Meta-Model Proposition ». Advances in Science, Technology and Engineering Systems Journal 4, nᵒ 5 (2019). https://doi.org/10.25046/aj040553.

[36] Erraissi, Allae, and Abdessamad Belangour. "An Approach Based On Model Driven Engineering For Big Data Visualization In Different Visual Modes." International Journal of Scientific & Technology Research.

[37] Fatima Kalna, Allae Erraissi, Mouad Banane, Belangour "A Scalable Business Intelligence Decision-Making System in The Era of Big Data" International Journal of Innovative Technology and Exploring Engineering 2019. https://doi.org/10.35940/ijitee.L3251.1081219

[38] « Webinars - IEEE Internet of Things ». Disponible sur : https://iot.ieee.org/education/webinars.html.

[39] Martin Fowler. Microservices a definition of this new architectural term, 2014.

[40] Kevin Khanda, Dilshat Salikhov, Kamill Gusmanov, Manuel Mazzara, and Nikolaos Mavridis. Microservice-based iot for smart buildings. In Advanced Information Networking and Applications Workshops (WAINA), 2017 31st International Conference on, pages 302–308. IEEE, 2017.

[41] Chris Richardson. Microservice architecture, 2017.

[42] Del Esposte ADM, Kon F, Costa FM, Lago N InterSCity: A Scalable Microservice-based Open Source Platform for Smart Cities.

[43] Krylovskiy A, Jahn M, Patti E (2015) Designing a Smart City Internet of Things Platform with Microservice Architecture. In: Proceedings - 2015 International Conference on Future Internet of Things and Cloud, FiCloud 2015 and 2015 International Conference on Open and Big Data, OBD 2015. Institute of Electrical and Electronics Engineers Inc., pp 25–30.

[44] Thramboulidis K, Vachtsevanou DC, Solanos A Cyber-Physical Microservices An IoT-based Framework for Manufacturing Systems.

[45] Herrera-Quintero LF, Banse K, Vega-Alfonso J, Venegas-Sanchez A (2016) Smart ITS sensor for the transportation planning using the IoT and Bigdata approaches to produce ITS cloud services. In: 2016 8th Euro American Conference on Telematics and Information Systems, EATIS 2016. Institute of Electrical and Electronics Engineers Inc.

[46] BANICA L, STEFAN C, HAGIU A (2017) Leveraging The Microservice Architecture For Next-Generation Iot Applications. Sci Bull - Econ Sci 16:26–32.

[47] Cherradi G, El Bouziri A, Boulmakoul A, Zeitouni K (2017) Real-Time HazMat Environmental Information System: A micro-service based architecture. In: Procedia Computer Science. Elsevier B.V., pp 982–987.

[48] Jarwar MA, Kibria MG, Ali S, Chong I (2018) Microservices in web objects enabled IoT environment for enhancing reusability. Sensors (Switzerland) 18:. https://doi.org/10.3390/s18020352.

[49] Kanti Datta S, Khan MI, Codeca L, et al IoT and Microservices Based Testbed for Connected Car Services.

[50] Taneja M, Jalodia N, Byabazaire J, et al (2019) SmartHerd management: A microservices-based fog computing–assisted IoT platform towards data-driven smart dairy farming. Softw - Pract Exp 49:1055–1078. https://doi.org/10.1002/spe.2704.

[51] A Blockchain-based IoT Platform Integrated with Cloud Services. https://www.researchgate.net/publication/334506644_A_Blockchain-based_IoT_Platform_Integrated_with_Cloud_Services. Accessed 25 Nov 2019.

[52] Power A, Kotonya G (2018) A Microservices Architecture for Reactive and Proactive Fault Tolerance in IoT Systems. In: 19th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, WoWMoM 2018. Institute of Electrical and Electronics Engineers Inc.

[53] Weighted-Scoring Model. http://www.appliedmanagement.com/downloads/. Accessed 4 Dec 2019.

[54] Jovellanos C (2003) Semantic and syntactic interoperability. In: Proceedings of the 4th ACM conference on Electronic commerce - EC '03. ACM Press, New York, New York, USA, p 266.

[55] Hosseini M, Dixon BE (2016) Syntactic Interoperability and the Role of Standards. In: Health Information Exchange: Navigating and Managing a Network of Health Information Systems. Elsevier Inc., pp 123–136.

[56] Pérez-Berenguer D, García-Molina J (2016) Cross-platform interoperable component for course analytics. In: ACM International Conference Proceeding Series. Association for Computing Machinery, pp 887–892.

[57] Pascual Espada J, Sanjuán Martínez O, Pelayo G-Bustelo BC, Cueva Lovelle JM (2011) Virtual Objects on the Internet of Things. Int J Interact Multimed Artif Intell 1:23. https://doi.org/10.9781/ijimai.2011.144.

[58] Impact of Agile Methodology on Software Development Process. https://www.researchgate.net/publication/255707851_Impact_of_Agile_Methodology_on_Software_Development_Process. Accessed 18 Nov 2019.

[59] Jabbari R, Ali N Bin, Petersen K, Tanveer B (2016) What is DevOps? A systematic mapping study on definitions and practices. In: ACM International Conference Proceeding Series. Association for Computing Machinery.

Weight Prediction System for Nile Tilapia using Image Processing and Predictive Analysis

Lean Karlo S. Tolentino[1], Celline P. De Pedro[2], Jatt D. Icamina[3], John Benjamin E. Navarro[4]
Luigi James D. Salvacion[5], Gian Carlo D. Sobrevilla[6], Apolo A. Villanueva[7], Timothy M. Amado[8]
Maria Victoria C. Padilla[9], Gilfred Allen M. Madrigal[10]

Department of Electronics Engineering, Technological University of the Philippines, Manila, Philippines[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
University Extension Services, Technological University of the Philippines, Manila, Philippines[1]

Abstract—Fish farmers are likely to cultivate poor quality fish to accommodate the rising demands for food due to the ever-increasing population. Fish growth monitoring greatly helps on producing higher quality fish products which leads to a better impact in the aquatic animal food production industry. However, monitoring through manual weighing and measuring stresses them that affects their health resulting to poorer quality or even fish kills. This paper presents a low-cost monitoring and Hough gradient method-based weight prediction system for Nile Tilapia (Oreochromis niloticus) using Raspberry Pi microcontroller and two low-cost USB cameras. This study aims to improve fish growth rate through monitoring the growth of the fishes with image processing eliminating the traditional way of obtaining fish measurements. By using paired t-test, the acquired values imply that the weight algorithm used to measure the weight of the fishes is accurate and acceptable to use. Growth performance of 10 Nile Tilapia was obtained in two intensive aquaculture setups – one for automated fish weighing through image processing and predictive analysis and the other setup for manual weighing. In response to weight prediction application, the growth of the fishes increased by 47.88%.

Keywords—*Fish; growth; Tilapia; image processing; predictive analysis; weight prediction*

I. INTRODUCTION

Due to the world's ever-growing population, humans need alternative food sources. With that, aquaculture is an essential contributor when it comes to food safety and daily living. It is the fastest growing sector in the food industry worldwide having its economic significance greatly increasing at the same time [1]. However, aquaculture is having difficulties in terms of production resulting on problems settling in the market [2]. In order to accommodate the growing demands for food, fish farmers tend to harvest poor quality fish. Thus, monitoring the growth of the fish helps to a great extent on producing a higher quality product.

Automated monitoring of the growth of the fish is a more convenient way of measuring and weighing it. According to the Florida and Wildlife Conservation Commission, manual measurement takes much labor and is time consuming on fish farmers as fishes are being captured and handled. Capture and handling are stressors that affect fish growth resulting on lower quality fish [3]. The quality of fish being lower is one of the issues aquaculture industry faces these days. Hence, with the advancement of technology, science is being used in various fields including the factory-farm and is applied to aquaculture especially in water quality monitoring [4]-[13] and fish growth & detection [14]-[15]. This makes the monitoring of the system automated, reduces time consumed by the farmers and less stressful to the fishes.

The aim of this paper is to develop an aquaculture system that monitors the growth of the fishes using image processing to improve fish growth rate. The study specifically aims (1) to develop a fish weight prediction system for an intensive aquaculture setup that automatically predict the fish weight without manually weighing it, (2) to implement an Internet of Things framework for accessing the weight prediction system through an Internet-based application that displays the average length and weight to determine the growth of the fishes, and (3) to assess the efficiency and reliability of the system, and the difference in the rate of fish growth between the aquaculture setup with weight prediction system and the conventional setup which involves manual weighing.

The system focuses on monitoring the growth of the fishes through image processing and predictive analysis. It utilizes two low-cost USB camera and micro-computer. Pre-programmed in the micro-computer are the tested calibration settings for the implementation of stereo-vision technology to the cameras used and the weight prediction algorithm. This method lessens the effort of fish farmers on determining the growth of the fishes while minimizing the possibilities to have a lower quality fish. Moreover, as the system utilized low cost materials, the implementation of the study also benefits fish farmers cultivating fishes on small-scale aquaculture system. The data acquisition is only once per week and is set to occur at a specific time in the day. This study is restricted to cultivating one species in an Intensive Aquaculture Setup, namely Nile Tilapia (Oreochromis niloticus). Tilapia is a fast-growing fish and tolerant of various conditions in the aquaculture environment. Due to its growth rate, low cost of production, and inexpensive on the market, Nile Tilapia is cultivated from extensive to intensive aquaculture systems in any form.

The paper is presented as follows: Section II includes the works related to the study and their gaps. The methodology and system architecture are explained in Section III. The results and discussion are presented in Section IV. Lastly, conclusion and future work are stated in Section V.

II. Related Works

Aquaculture, also known as aquafarming, is the cultivation of aquatic plants and animals under controlled or semi-controlled system [16]. It is the farming of aquatic creature such as fish, mollusks, algae and other organisms. It provides a living habitat for various aquatic creature with environment conditions optimal for their growth [17]. With the growing fish demand and declining production from fisheries, aquaculture is acquiring further significance [18].

Tilapia as shown in Fig. 1 is very versatile and is tolerant to different aquaculture environments. Tilapia are cultivated in every method from extensive to intensive aquaculture system [19]. Because of its growth rate, low production cost and affordable in the market, Tilapia is being cultured in aquaculture [20]. Moreover, they only have 5 basic needs: (1) food (2) light (3) room to Swim (4) oxygen and (5) clean water.

A study in [21] stated that approximation of fish size and weight is important in fish farming. This provides data that are essential for fish feeding and harvesting. However, manual weighing and measurement are stressful to the fish.

According to a study [3], capturing and handling causes stress. Provided by experiments, evidences are documented that when handling the fish, its blood cortisol and/or glucose levels increases which are intrinsically stressful.

A study in [22] stated that Tilapia shows isometric growth patterns whenever they are culture in an environment condition optimal for their growth. This study obtained the length-weight relationship of Nile Tilapia.

Most previous works relating to aquaculture and fisheries systems only considered to assess the effects of stress to fishes due to manual weighing and determining the length and weight relationship of Nile Tilapia. The researches and information pertaining weight prediction specifically on the genus Tilapia, species Oreochromis niloticus (Nile Tilapia) generally, have not been applied or put to practical use. To further enhance the study, it is also important to consider some of the other factors that greatly affect the growth, as well as quality, of the fishes thus, the proponents considered developing a weight prediction system for Nile Tilapia. The incorporation of weight prediction system suggests for a lesser possibility of the fishes to have a poorer quality. It also reduces the effort and time consumed by the fish farmers on determining the fish growth rate. Profound research on fish growth issues show stress as one of the major factors influencing fish growth, which is why the lack of direct interaction with fish leads to a healthier fish.

Fig. 1. Tilapia (Oreochromis Niloticus).

III. Methodology

A. Conceptualization and Design

The design and development of this study is mainly concerned with the weight prediction system for Nile Tilapia. Weight of the fish is the vital parameter that was considered in this study.

Fig. 2 shows the process flowchart. Two low-cost USB Camera were used. These cameras are programmed into a Stereo-Vision camera for distance measurement. The camera captured images which are processed in the Arduino Mega. The system detects fishes and acquires different parameters needed for predictive analysis. The then captured images are processed through various filters to accurately calculate the actual length of the fishes for the weight prediction through length and weight correlation. From here, the data are sent to the database via Long Range Wide Area Network (LoRaWAN) IoT Protocol and are displayed to the Web Application.

B. Fish Weight Prediction System

Two aquaculture systems as shown in Fig. 3 were made to culture 10 Nile Tilapia. The first system employs conventional method of weighing Tilapia wherein the fishes are being captured, handled and placed on a digital weighing scale while the other system employs automated fish weighing where a camera is used to predict the weight of the fish.

Fig. 4 shows the system flowchart of the proposed system. The input parameter in the system includes the image of fish. For the process it includes image processing that is composed of various sections namely: cropping which is the division of image into two regions of interest [23], grayscale used for converting image into shades of gray with no apparent color, blurring which is the relative motion between a camera that is out of focus and a specific object [24], binary thresholding to create binary images, erosion for removal of pixels on object boundaries, dilation for adding pixels on object boundaries and morphological closing for filling small holes from an image without affecting the size and shape of the objects in the image.

Two low-cost USB camera were utilized to produce stereo-camera, implementing stereo-vision technology. Multiple images from both cameras were collected for calibration and depth map tuning. Tested calibration settings were loaded to attain accurate distance measurements.
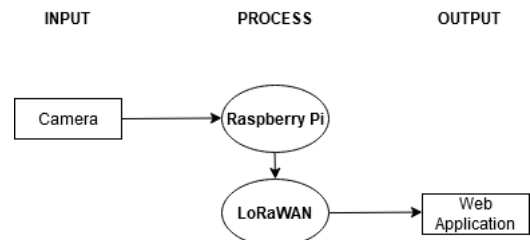

Fig. 2. Process Flowchart.

Fig. 3.   Conventional and Automated Aquaculture System.
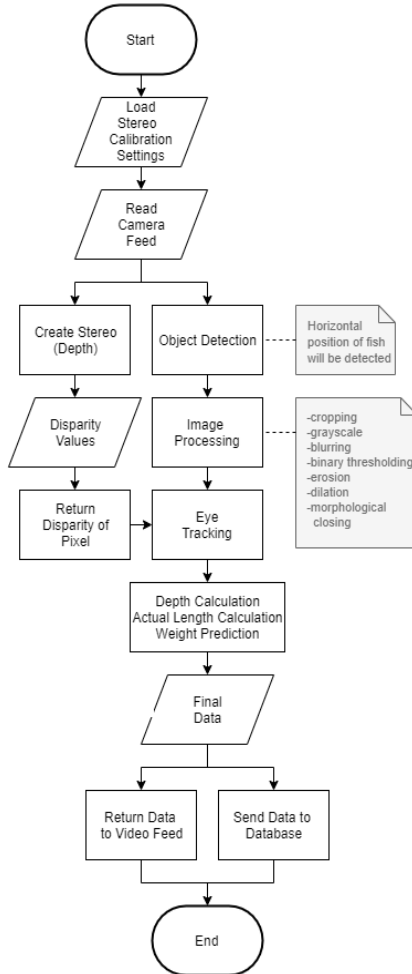

Fig. 4.   System Flowchart.

After reading the camera feed, object detection and depth map creation were processed simultaneously. Frames, with detected images of Nile Tilapia, undergone series of image processing techniques to filter only the contour of fish, as well as removal of noises and unwanted data for more accurate measurements.

Once the contour was drawn, eye-tracking technique through circle detection was performed to pin-point consistent part of fish as basis of its distance from the camera. Hough Gradient Method of the OpenCV is implemented to detect circles. This method utilizes the gradient information of edges.

If a circle is parameterized by its radius r and its center coordinates (a, b), then these are correlated to the position of edge points (x, y), which create the circle by means of the limitation:

$$(x - a)^2 + (y - b)^2 = r^2 \tag{1}$$

Equation 1 also specifies that "every certain edge point (x, y) could be a point on any circles" where "their limitations lie on the plane of a right circular cone in the (a, b, r) parameter space." All image points which are on the circle interpreted by those three parameters if the cones corresponding to several edge points cross at an only point [24].

Actual length was measured using the contour of fish, converting its pixel length to an actual measurement following pixel/metric conversion. Lastly, actual length was used for weight prediction through predictive analysis.

The weight prediction which is derived using polynomial regression is represented as:

$$w = 0.1017x^3 - 4.8944x^2 + 93.44x - 583.06 \tag{2}$$

Where: w = weight of the fish predicted

x = length measured

Fig. 5 shows the image processing once the fish is detected. Once the contour was drawn, circle detection takes place to detect the eye of the fish for basis of distance measurement.

Fig. 6 shows the actual fish detection with length measurement and weight prediction.

C. Distance Measurement

A stereo-vision camera has two or more lenses with a distinct image sensor for every lens and have the ability to perceive depth. Instead of using costly depth camera, two low-cost USB cameras are calibrated to act as a stereo-vision camera.

The cameras are calibrated in order to calculate the disparity value between frames captured from left and right camera as shown in Fig. 7.
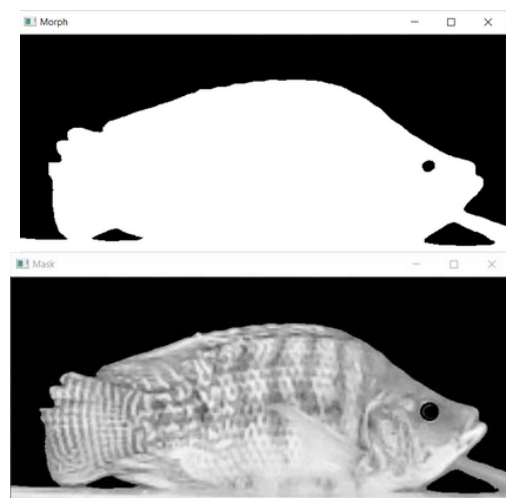
Fig. 5.   Image Processing.
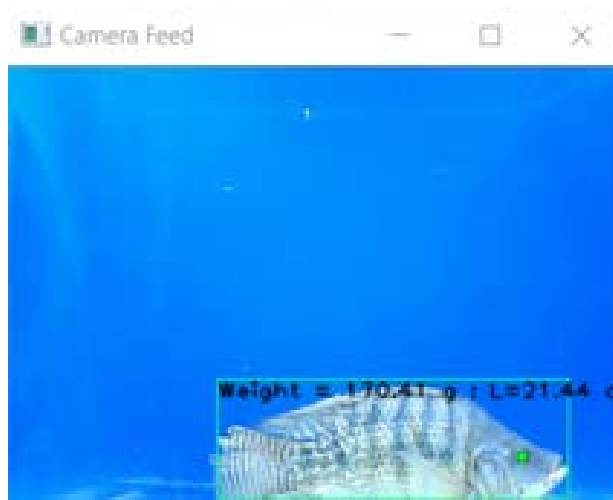
Fig. 6.   Weight Prediction.
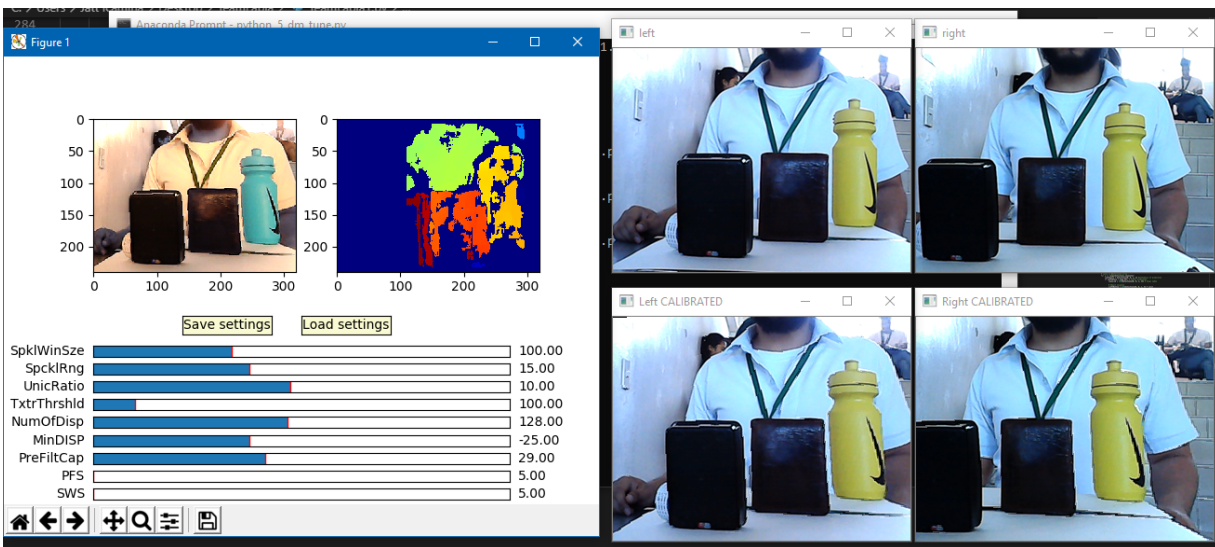
Fig. 7.   Calibration Settings.

Fig. 8 shows the Regression analysis between the actual distance and disparity value. The acquired equation for the distance of the fish in terms of pixel is calculated as:

$$D = 0.1017z^3 - 4.8944z^2 + 93.44z - 583.06 \qquad (3)$$

Where: D = distance in terms of pixel

z = disparity value

To identify the actual distance of the fish from the camera, the obtained virtual length (in terms of pixel) is calculated following pixel/cm conversion using regression analysis as shown in Fig. 9. The acquired actual distance is used to calculate the length of the fish that is needed for weight prediction.

The actual distance of the fish following pixel/cm conversion is represented as:

$$ppcm = 0.0017z^2 - 0.1601z - 10.18 \qquad (4)$$

D. Examined Trendlines

Fig. 10 shows the relationship between length and weight in a polynomial trendline. The acquired R2 value is 0.9324 while the equation for weight computation is $0.1017x^3 - 4.8944x^2 + 93.44x - 583.06$.
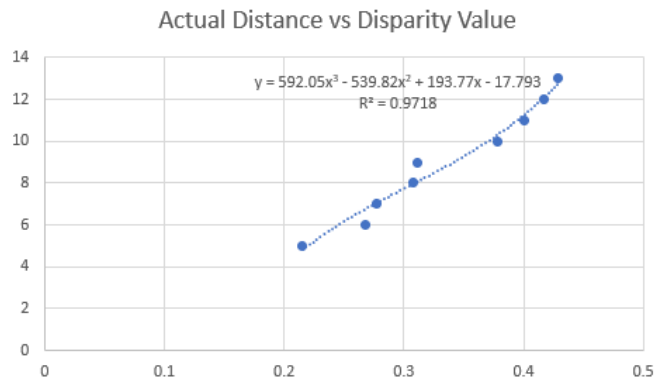
Fig. 8.   Actual Distance vs Disparity Value
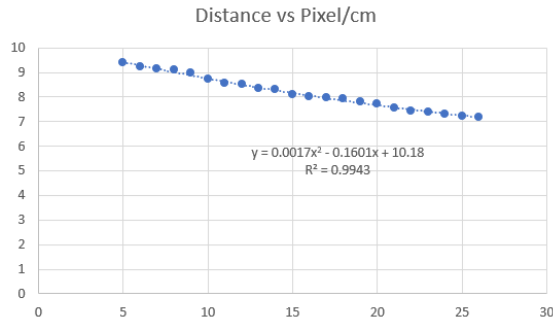
Fig. 9. Distance vs. Pixel/cm.

Fig. 11. Length in Cm vs. Weight in grams (Exponential).

Fig. 10. Length in Cm vs. Weight in Grams (Polynomial).
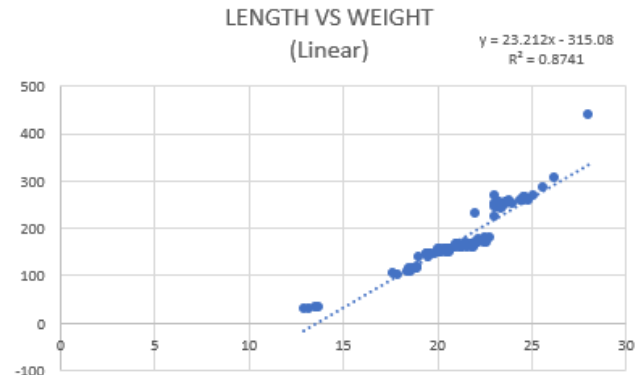
Fig. 12. Length vs Weight (Linear).

Fig. 11 shows the correlation of length and weight in a Exponential trendline. The value of the R2 acquired is 0.9286 while the equation for weight computation is $4.7915e^{0.1661x}$.

Fig. 12 shows the correlation of length and weight in a Linear trendline. The value of the R2 acquired is 0.8741 while the equation for weight computation is 23.212x - 315.08.

Fig. 13 shows the correlation of length and weight in a Logarithmic trendline. The value of the R2 acquired is 0.8125 while the equation for weight computation is 435.58ln(x) - 1149.1.

Various trendlines were examined to determine the most suitable to the data gathered such as Polynomial, Exponential, Linear and Logarithmic. The value of the R2 was used to determine the fitness of the line. The closer to the value of 1, the best fit trendline to the dataset.

E. Web Application

The TeamLapia web application as shown in Fig. 14 is made to organize all the data gathered and display the most recent status of the fishes. The web application exhibits the numerical values of the essential parameters including the average length and weight to determine the growth of the fishes. Through the use of various JavaScript and php codes, data are transmitted from the database to the UI. On the other hand, using html, CSS and JavaScript scripts, the graphs and dynamic design of the application are made to make it more suitable for end-users.

Fig. 13. Length in cm vs Weight in grams (Logarithmic).

F. Statistical Analysis

The results obtained between the measured and predicted weight of the samples were evaluated statistically using t-test: Paired Two Sample for Means.

The test statistic is calculated as:

$$t = \frac{\bar{d}}{\sqrt{s^2/n}} \tag{5}$$

Where: d' = mean difference

s^2 = sample variance

n = sample size

t = t quantile with n-1 degrees of freedom

IV. RESULTS AND DISCUSSION

To test the accuracy of the Weight Prediction Algorithm, a comparison was made between the actual measured weight of the fishes and the predicted weight. Table I shows the recorded weights of ten fish samples and their corresponding percentage error. With the data assimilated, it was inferred that the measured and predicted weight were almost the same as it only has a mean percentage error of 2.82%. With that in mind, to further test the accuracy of the device, a hypothesis test was also conducted.

The p-value for two-tailed test is greater than the significance level of 0.01 which means that we have to accept the null hypothesis that there is no significant difference between the measured and predicted weight of the samples, Table II. This implies that the weight prediction algorithm used to measure the weight of the fishes is accurate and is acceptable to use.

Fig. 14. Web Application Interface.

TABLE I. EXPERIMENTAL DATA FOR PREDICTED AND MEASURE WEIGHT OF SAMPLE FISHES

Fish Sample	Weight (g)		Difference	Error (%)
	Measured	*Predicted*		
1	105.5	110	-4.5	4.09%
2	97.5	105	-7.5	7.14%
3	148.4	150	-1.6	1.06%
4	141.6	150	-8.4	5.6%
5	152.4	150	2.4	1.6%
6	104.7	105	-0.3	0.28%
7	147.5	150	-2.5	1.67%
8	134.8	135	-0.2	0.15%
9	108.4	110	-1.6	1.45%
10	104.3	110	-5.7	5.18%
Mean Percentage Error				**2.82%**

TABLE II. TWO SAMPLE FOR MEANS PAIRED T-TEST OF EXPERIMENTAL DATA FOR PREDICTED AND MEASURED WEIGHT OF SAMPLE FISHES

	Measured	Predicted
Mean	124.51	127.5
Variance	491.8232222	445.8333333
Observations	10	10
Pearson Correlation	0.988461845	
Hypothesized Mean Difference	0	
df	9	
t Stat	2.736983041	
P(T<=t) one-tail	0.011480323	
t critical one-tail	2.821437925	
P(T<=t) two-tail	0.022960645	
t critical two-tail	3.249835542	

The weight of Nile Tilapia was monitored in both conventional and automated aquaculture setup. Initially, the weight of the fishes on both setups are equal and the growth of the fishes is observed for two weeks. The fish growth rate based on weight in the automated setup which uses camera for weight prediction was 17g (from 24g to 41g) while the conventional system which employ manual weighing of fish only obtained an average of 11g (34g to 35g) as shown in Fig. 15.

Table III shows the measured weight of the fish cultured in the automated setup which employs camera for weight prediction.

Table IV shows the measured weight of the fish cultured in the conventional setup which uses the manual way of weighing fish that involves, catching, handling and measuring on a digital weighing scale.
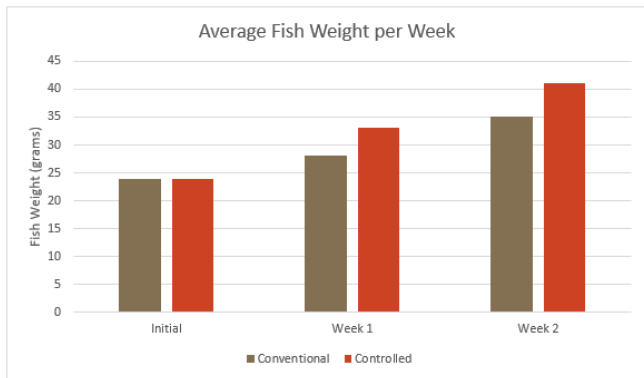
Fig. 15. Conventional vs Automated Aquaculture Setup Fish Growth measured Every Week.

TABLE III. FISH WEIGHT IN AUTOMATED SETUP MEASURED USING CAMERA

Fish No.	Initial	Week 1	Week 2
1	20	25	35
2	20	30	35
3	25	35	40
4	25	35	45
5	30	40	50
Average	24	33	41

TABLE IV. FISH WEIGHT IN CONVENTIONAL SETUP MEASURED MANUALLY

Fish No.	Initial	Week 1	Week 2
1	20	25	30
2	20	25	30
3	25	30	35
4	25	30	40
5	30	35	40
Average	24	28	35

Based on the data shown, the growth rate in the automated aquaculture setup is 30.70% each week and is greater compared to the conventional setup which has 20.76% growth rate per week. The automated aquaculture setup which employs weight prediction through camera improves the growth of the fishes in terms of weight by 47.88%.

V. CONCLUSION

The authors developed an automated aquaculture setup with weight prediction technology that makes use of image processing techniques and predictive analysis which yields to a higher growth and survival rate for Nile Tilapia. By avoiding stressors to fishes, such as capturing and handling to manually measure its weight and monitor its growth, the data gathered shows beneficent results to the rate of fish growth. The growth rate of the fishes in terms of weight improves by 47.88% with only 2.82% mean percentage error on the weight prediction. For future work, this study can be applied for growth monitoring of other animals on various environment.

For future work, this study can also be applied in predicting the weight and length of other aquatic animals in different aquaculture systems to determine its growth. This lessens the effort of the fish farmers and minimizes the possibilities of the fish to have a lower quality. Moreover, the acquired length and weight measurements may also be used to determine the number of fishes to cultivate in a given area to prevent space and food deprivation.

REFERENCES

[1] FAO, 2009. The state of world fisheries and aquaculture 2008. Rome, Italy.

[2] S. Han, Y. Kang, K. Park, and M. Jang, "Design of environment monitoring system for aquaculture farms," in 2007 Frontiers in the Convergence of Bioscience and Information Technologies, pp. 889-893, 2007.

[3] C. Harper and J. C. Wolf, "Morphologic effects of the stress response in fish," ILAR Journal, vol. 50, no. 4, pp. 387-396, 2009.

[4] L. Wenmei and L. Yuzhen, "Aquaculture monitoring system," in 2010 International Forum on Information Technology and Applications, pp. 138-141, 2010.

[5] L. K. S. Tolentino et al., "Development of an IoT-based Intensive Aquaculture Monitoring System with Automatic Water Correction," International Journal of Computing and Digital Systems, in press.

[6] N. Gallah, O. B. Bahri, N. Lazreg, A. Chaouch, and K. Besbes, "Water quality monitoring based on small satellite technology," International Journal of Advanced Computer Science and Applications, vol. 8, no. 3, pp. 357-362, 2017.

[7] L. K. Tolentino et al., "IoT-Based Automated Water Monitoring and Correcting Modular Device Via LoRaWAN for Aquaculture," International Journal of Computing and Digital Systems, in press.

[8] M. A. Shareef, A. Toumi, and A. Khenchaf, "Estimation of water quality parameters using the regression model with fuzzy k-means clustering," International Journal of Advanced Computer Science and Applications, vol. 5, no. 6, pp. 151-157, 2014.

[9] A. Maestre, Eman El-Sheikh, Derek Williamson and Amelia Ward, "A Machine Learning Tool for Weighted Regressions in Time, Discharge, and Season" International Journal of Advanced Computer Science and Applications, vol. 5, no. 3, pp. 99-106, 2014.

[10] L. K. Tolentino et al., "AquaDroid: An App for Aquaponics Control and Monitoring," in 6th International Conference on Civil Engineering (6th ICCE 2017), pp. 1-8, 2017.

[11] E. Galido et al., "Development of a Solar-powered Smart Aquaponics System through Internet of Things (IoT)," in Lecture Notes on Research and Innovation in Computer Engineering and Computer Sciences, pp. 31-39, 2019.

[12] L. K. S. Tolentino et al., "Development of an IoT-based Aquaponics Monitoring and Correction System with Temperature-Controlled Greenhouse," 2019 International SoC Design Conference (ISOCC), 2019, pp. 261-262.

[13] L. K. S. Tolentino et al., "Yield evaluation of Brassica rapa, Lactuca sativa, and Brassica integrifolia using image processing in an IoT-based aquaponics with temperature-controlled greenhouse," Agrivita Journal of Agricultural Science, in press.

[14] R. R. Herrera and F. G. Funes, "Affective Educational Application of Fish Tank Hydroponics System," International Journal of Advanced Computer Science and Applications, vol. 10, no. 12, pp. 126-131, 2019.

[15] K. Raza and S. Hong, "Fast and Accurate Fish Detection Design with Improved YOLO-v3 Model and Transfer Learning," International Journal of Advanced Computer Science and Applications, vol. 11, no. 2, pp. 7-16, 2020.

[16] J. H. Tidwell and L. A. Bright, "Freshwater Aquaculture," Encyclopedia of Ecology, pp. 91–96, 2018. doi: 10.1016/b978-0-12-409548-9.10618-9

[17] K. K. Kishore, P. V. Krishna, and D. Srikanth, "Automatic Feeding system for Aquaculture," in 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), pp. 426-429, 2017.

[18] S. N. Ojha and S. C. Babu, "Why convergence of Fisheries Co-management with Agricultural Technology Management Agency is significant," in Agricultural Extension Reforms in South Asia, pp. 329-347, 2019.

[19] K. Fitzsimmons, "Tilapia: the most important aquaculture species of the 21st century," in Proceedings from the fifth International Symposium on tilapia Aquaculture, pp. 3-8, 2000.

[20] Q. Y. Guo, X. C. Wang, X. S. Yang, Z. Xu, M. Gu, and X. Y. Li, "Freshness changes and shelf life prediction of cultured Tilapia (Oreochromis niloticus) during chilled storage," in 2011 Fourth International Conference on Intelligent Computation Technology and Automation, pp. 86-92, 2011.

[21] R. Tillett, N. McFarlane, and J. Lines, "Estimating dimensions of free-swimming fish using 3D point distribution models," Computer Vision and Image Understanding, vol. 79, no. 1, pp. 123-141, 2000.

[22] A. A. B. Adam and A. M. Khalid, "Length weight relationship and condition factor of Nile Tilapia Oreochromis niloticus (Trewavas, 1983) in the southern part of Jebel Aulia Dam, White Nile, Sudan," Direct Res. J. Agric. Food Sci, vol. 4, no. 10, pp. 286-289, 2016.

[23] L. K. S. Tolentino, J. W. F. Orillo, P. D. Aguacito, E. J. M. Colango, J. R. H. Malit, J. T. G. Marcelino, A. C. Nadora, and A. J. D. Odeza, "Fish freshness determination through support vector machine," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 9, no. 2-5, pp. 139-143, 2017.

[24] Y-. L. You and M. Kaveh, "A regularization approach to joint blur identification and image restoration," IEEE Transactions on Image Processing, vol. 5, no. 3, pp. 416-428, 1996.

Detection of Plant Disease on Leaves using Blobs Detection and Statistical Analysis

N. S. A. M Taujuddin[1], A.I.A Mazlan[2], R. Ibrahim[3], S. Sari[4], A. R. A Ghani[5], N. Senan[6], W.H.N.W Muda[7]

Faculty of Electrical and Electronic Engineering[1, 2, 4]
Faculty of Computer Science and Information Technology[3, 6]
Faculty of Civil and Environmental Engineering[5]
Faculty of Technical & Vocational Education[7]
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia[1, 2, 3, 4, 5, 6, 7]

Abstract—**Plant is exposed to many attacks from various micro-organism, bacterial disease and pests. The symptoms of the attacks are usually distinguished through the leaves, stem or fruit inspection. Disease that are commonly attack plants are Powdery Mildew and Leaf Blight and it may cause severe damaged if not controlled in early stages. Image processing has widely being used for identification, detection, grading and quality inspection in the agriculture field. Detection and identification disease of a plant is very important especially, in producing a high-quality fruit. Leaves of a plant can be used to determine the health status of that plant. The objective of this work is to develop a system that capable to detect and identify the type of disease based on Blobs Detection and Statistical Analysis. A total 45 sample leaves images from different colour and type were used and the accuracy is analysed. The Blobs Detection technique are used to detect the healthiness of plant leaves. While Statistical Analysis is used by calculating the Standard Deviation and Mean value to identify the type disease. Result is compared with manual inspection and it is found that the system has 86% in accuracy compared to manual detection process.**

Keywords—Image processing; blob detection; edge detection; statistical analysis; disease detection

I. INTRODUCTION

Nowadays, image processing is widely used in agriculture, security system and medical field. Various researchers and advancements are going on vigorously in plant disease detection in the agricultural field. One of the promising applications in image processing is agriculture. Image processing can be used to identify the plant disease, grading the fruits and ripeness of the fruit [1].

Plant that are constantly attacked by disease are papaya, chili and tomato. These plants usually attacked by disease namely Powdery Mildew, Leaf Blight and Black Spot [2]. Black Spot disease is caused by Asperisporium caricae bacteria, where both leaves and fruit of the plant can be affected. This disease can affect the plant at any stages of their growth [3]. Leaf Blight is another disease appear on leaves as small and irregular scatter. It is brown to grey in colour.

The Powdery Mildew and Leaf Blight are persistent and common threat to plant. Moreover, plant is also host to a dozen species of fungal. This fungal disease of plant leaf and fruit may cause premature defoliation. Besides that, when the plant

leaves and fruit affected by these fungal it may decrease the production [2].

To produce a high-quality fruit, there is a need to have a healthy tree, where the plant disease needs to be detected at early stages. Thus, in this research, a method is propose to detect the disease of a plant by using image processing technique. Consequently, it will help to increase the volume of production thus cater the demand of market.

II. LITERATURE REVIEW

In agriculture, the important thing to measure from the farmer's points of view are canopy, yield and quality of product. Hence, application of image processing can be used to improve the decision making for vegetarian measurement, irrigation and fruit sorting. There are some example application of image processing in agricultural field such as tree detection, counting of apple, disease detection, fruit ripeness detection and fruit quality detection.

The challenges in continual fruit cultivations is calculate the quantity of fruits on a tree because it need to get yield estimation form different farm. Circular Hough transform (CHT) and color thresholding method are used to detect and count each apple fruit. By using these method, the identification process is successfully done with fruits are clearly visible [4].

Various micro-organism and bacterial disease and pest attack on chili plant. The symptoms can be seen through the leaves, stems or fruit inspection. Leaf features inspection need to done for early detection of chili disease. Image of leaves are captured and processed to examine the healthy status of each plant. Moreover, these techniques can ensure the chemicals are only use when the plants are affected by disease. Hundreds of chili disease image are used in image processing techniques and its very useful and inexpensive system [5].

The author in [6] proposed a hybrid method to detect the fruit ripeness by using histogram matching, clustering algorithms-based image segmentation and relative value of parameter-based segmentation. The input data are colored images of fruits and vegetables. The threshold levels are set to find the maturity level of fruits and vegetables by comparing the input data images with threshold levels [6].

Mango is one of the commercial fruits produced and consumed throughout the world. Two method for classification of ripening stage based on changes in their visual features are proposed in [7]. The method are RGB and HSV where the ripening stage is detect based on the red, green and blue components of the image pixels while analysed for detection using HSV method the hue- saturation-value map [7].

The analysis that shows the fruits are quality are not based on their size by consumer performance, i.e. consumer prefer fruits of equal weight uniform size for example people like yellow bananas, dark red apple, light green or dark black grapes and yellow mango etc. Furthermore, the estimation mean of fruit size is important in meeting quality standard increasing marketing value monitoring growth. So that, the method was proposed to determine the quality (i.e. ripe, partial ripe, ripe or over ripe (bad fruit)) is based on appearance of color level estimation further which is image analysis, visual examination and inspections of color are included [8].

The most important things in agriculture are to have are high yield and quality product. However, disease is a huge problem that can reduce the production. Diseases that are usually occurred at early stages, are Powdery Mildew, Leaf Blight, Black Spot, Leaf Curl and Ring Spot. Detection of these disease is very necessary because if the plants are affected by the disease, the quality of fruit will be decreased, and the maturity of the fruit is interrupted. Hence, it will interrupt the production yield.

Powdery Mildew (see Fig. 1) growth in humid areas with warm days and cool nights, such as at north Kona on the island of Hawai'i. The symptoms for early stages of the disease is the undersides of leaves are speckled with small, water-soaked spots that becoming powdery patches of mycelium and spores. Hence, the infections usually concentrated near the leaf veins. Then the diameter of white patches is from 1-6 cm, with yellowish-green spots on the upper leaf surface.

Next, mildewed areas grow and will serve yellowing between the veins increasing. Furthermore, the spore-forming mycelium will wrap around the leaf edge and grow on the upper leaf surface and petioles. For the disease progress, infected leaves become necrotic and appear scorched. They curl and fall from plants prematurely. Mildew on immature fruit begins as circular patches of white mycelium and spores that can coalesce and cover the entire fruit. The deformed fruit is edible but has little or no value in the market [2].

The Leaf Blight is also known as, Leaf Spot or Black Rust (see Fig. 2). This disease is attacked to the leaves under certain conditions. Maturing fruit may affected too. Moreover, the fungus disease of Leaf Blight is caused by Puccniopsis caricae bacteria. Next, the symptoms of plant leaf blight are recognized through black pustular spots that appear on the under surfaces of infected leaves. Above the leaf tissue, the infected area is small, generally circular in outline and slightly raised above the leaf tissue.

The primary infections are usually scattered and the size one to eight of an inch to mere dots. The infections appear as a small brown spot of dead tissue on the upper side of affected leaves [9].

Fig. 1. Powdery Mildew on Leaves.

Fig. 2. Leaf Blight on Leaves.

III. METHODOLOGY

The process begins with image acquisition which is obtaining the image sample from various resource such as Plantix.Net website. Here, 15 sample image of healthy leaves, 15 sample images of Powdery Mildew and 15 sample images of Leaf Blight are collected. Then, pre-processing task involve some procedure to prepare the image to be ready for image processing by converting it to grey scale image. Then, the images undergoes the Blobs Detection process to identified either it is healthy or unhealthy leaves. Next, the Statistical Analysis is done, where the Standard Deviation and Mean value is calculated to classify the plant into Powdery Mildew or Leaf Blight.

Healthy leaves mean, the surfaces on leaves is clear with healthy green of color. While the unhealthy leaves have some dots coloured with white, brown or black. The input image is converted into grayscale image to detect the healthiness of leaves. This conversion is applied to transform the coloured image into black and white coloured ranged between 0 and 255. After that the value of threshold is estimated.

Here threshold value 0.45 is used. In order to perform 2-D Median filtering, the binary image need to be inverted. The grey scale image is filtered using the function of *"medfilt2"* which stand for 2-D median filtering. The 2-D median filtering is used to remove the noise. This is because each output pixel contains the median value in the m-by-n neighbourhood (m for row and n for column) around the corresponding pixel in the input image.

Next, edge detection using Canny filter technique is performed to create a boundary of contours on filtered plant leaves image. Then, contour with defective part is filled with the numerous blobs and perfect part filled with no blobs or minimum number of blobs.

Contour is used to control the image features for appearance and behaviour of objects. By using the image processing, leaves extracts the contour line information from the intensity image and the result of image was determined from contour information [10]. This statement is based on equation of contour.

$$E_{snake} = \int_0^1 [E_{int}(v(s)) + E_{ext}(v(s))]ds \qquad (1)$$

where *v(s)* is the parameter equation of the contour, and *s* is the arc length.

The Blobs Detection is a process to obtain regions of interest for further processing and these regions can detect the presence of object or parts [11]. The higher the number of blobs shows the unhealthy leaves, while the lower the number of blobs, shows the healthy leaves [10].

The statistical analysis is done by measuring the Standard Deviation and Mean value of the histogram on plant leaves [12]. Firstly, standard deviation is applied on histogram because the histogram can show the colour intensity values of colour images. Also, Standard Deviation can calculate the contrast of the colour image based on histogram. The equation to calculate the Standard Deviation is

$$std(i,j) = (\frac{1}{8}\sum_{k=1}^{8} x_k - l_m(i,j)^2)^{\frac{1}{2}} \qquad (2)$$

where x_k are eight surrounding pixel around processing pixel $l_m(i,j)$ and j is varies from 2 to number of rows and 2 to number of columns [13].

Then, Mean function are used to measure the average intensity value of the pixel distribution. The equation for the above statement are [12][14]:

$$\mu_m = \sum_{n=0}^{L-1}((x_n - y)^m \, t \, (x_n)) \qquad (3)$$

and

$$y = \sum_{n=0}^{L-1}(x_n \, t \, (x_n)) \qquad (4)$$

The range of Standard Deviation and Mean will then be used to classify the disease either Powdery Mildew or Leaf Blight [15-16]. These measurement deals with the variation of intensity of the image surface.

IV. RESULT AND ANALYSIS

There are several steps needed in blobs detection to identify either the leaves is healthy or not. First of all, the original image is read. Then, the original image is converted into grey scale image. This conversion is applied to transform the colour image into black and white colour. Next, the grey scale image are filtered using 2-D median filtering. Next, edge detection using Canny filter is performed to create a boundary of contours on filtered leaves image. Then, it goes to analysis where contour with defective part will be filled with the numerous blobs while perfect part is filled with no blobs or minimum number of blobs.

Contour is used to control the image features for appearance and behaviour of objects [17]. Using the image processing, leaves extracts the contour line information from the intensity image and the result of image was determined from contour information [10] [18].

The blobs detection is a process of detecting the image with different properties of brightness or color, compared to surrounding region and these regions can detect the presence of object or parts [11][19-21].

Based on the step used, the unhealthy leaves can be identified by extracting the contour of damaged part. The healthy and unhealthy leaves are categorized by the number of blobs. For image with number of blobs is equal or less than 90,

it is considered are healthy leaves. While image with more than 90 blobs are considered are unhealthy leaves (refer Table I).

In this analysis, 45 leaves images are used to test the blobs appearance. From 45 images, there has 13 healthy leaves and 26 unhealthy leaves are detected.

Table II shows two example of healthy leaves. These two sample shows less than 90 blobs occurrence reflecting that the surface of the leaves are perfect and smooth with no Powdery Mildew and Leaf Blight. At the contour step, at defective part the image is filled with the numerous blobs and perfect part filled with no blobs or minimum number of blobs. Hence, the outline of image is look very clear at contour's step and it show the image has a perfect surface. Then, from the contour result, calculation number of blobs is done.

Table III shows two example of Leaf Blight image sample. The detection of the brown spot and scar has undergoes process of edge detection. There is no normal shape of the plant leaves present because the outline of the leaves is not really clear and high number of blobs are present. At contour step, the shape of the leaves is not in a good surface; it has brown spot and contains a lot of blobs. The numbers of blobs were detected more than 90 in the images and it is classified as unhealthy leaves.

TABLE I.   TABLE TYPE STYLES RANGE OF LEAVES HEALTINESS BASED ON NUMBER OF BLOBS

Healthiness	Number of Blobs
Healthy	x ≤ 90
Unhealthy	x > 90

TABLE II.   BLOBS DETECTION ANALYSIS TO DETECT THE HEALTY LEAVES
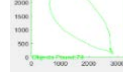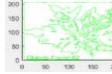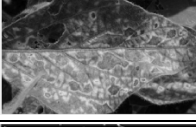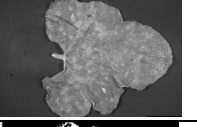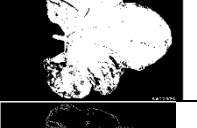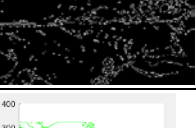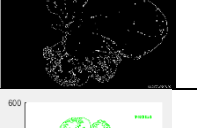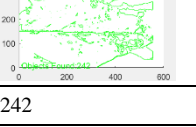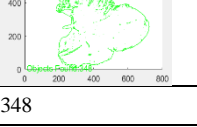
Sample	Image 1	Image 2
Original image		
Grey Scale image		
Binary image		
Filtering		
Edge detection		
Contour		
Number of blobs	78	82
Decision	Healthy	Healthy

TABLE III.   BLOBS DETECTION ANALYSIS TO DETECT THE UNHEALTHY LEAVES

Sample	Image 3	Image 4
Original image		
Grey Scale image		
Binary image		
Filtering		
Edge detection		
Contour		
Number of blobs	242	348
Decision	Unhealthy	Unhealthy

Shape of leaves and different color spotted on leaves will play an important role in identifying it as healthy or unhealthy leaves.

There are several steps in Statistical Analysis done to classify the disease either Powdery Mildew or Leaf Blight. The result from blobs detection are determined either the leaves is healthy or not. If the sample detected as unhealthy it will be proceeded to the next step which Statistical Analysis.

The Statistical Analysis is done by performing Standard Deviation and Mean value analysis. The Standard Deviation value will give a representation of the amount of variations in average while Mean value will give average intensity of all the pixels in the image representation [12]. 15 leaves with Powdery Mildew and 15 leaves with Leaf Blight are tested its Standard Deviation and Mean value. From the testing, it can be conclude that, the Powdery Mildew has the Standard Deviation ranging from 39 to 73, while Leaf Blight from 31 to 57 (see Table IV). While for Mean value, the Powdery Mildew ranging from 107 to 139, and Leaf Blight ranging from 71 to 150 (see Table V).

Tables VI and VII shows four different leaves images samples. Each of image has shown a different value of Standard Deviation and Mean. By mapping these statistic value to Tables IV and V, it lead to determination of either the leaves are effected with Powdery Mildew or Leaf Blight.

TABLE IV.   STANDARD DEVIATION VALUE RANGE FOR DISEASE CLASSIFICATION

Disease	Standard Deviation
Powdery Mildew	$39 \leq x \leq 73$
Leaf Blight	$31 \leq x \leq 57$

TABLE V.   MEAN VALUE RANGE FOR DISEASE CLASSIFICATION

Disease	Mean
Powdery Mildew	$107 \leq x \leq 139$
Leaf Blight	$71 \leq x \leq 150$

TABLE VI.   ANALYSIS ON LEAVES AFFECTED WITH POWDERY MILDEW DISEASE
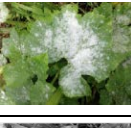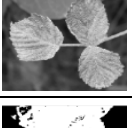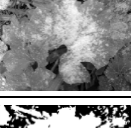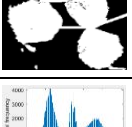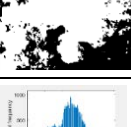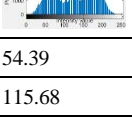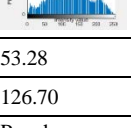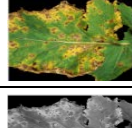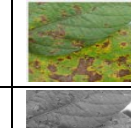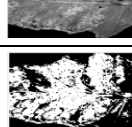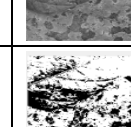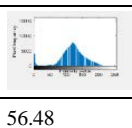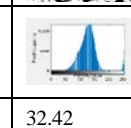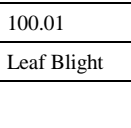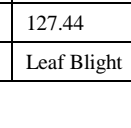
Sample	Image 16	Image 18
Original image		
Grey Scale image		
Filtering		
Histogram graph		
Standard deviation	54.39	53.28
Mean	115.68	126.70
Decision	Powdery Mildew	Powdery Mildew

TABLE VII.   ANALYSIS ON LEAVES AFFECTED WITH LEAF BLIGHT DISEASE

Sample	Image 40	Image 42
Original image		
Grey Scale image		
Filtering		
Histogram graph		
Standard deviation	56.48	32.42
Mean	100.01	127.44
Decision	Leaf Blight	Leaf Blight

V. CONCLUSION

In this research, a system to overcome the problems of manual inspection and classifying the disease of leaves is completely done. This system used blobs detection and calculation of statistical analysis to classify the disease either Powdery Mildew or Leaf Blight. Forty-five images from different categories were used. Result are compared with manual results and it is found that it was 86% accurate in this system.

ACKNOWLEDGMENT

REFERENCES

[1] P. Soni, "Identification of Plant Disease Using Image Processing Techniques," Int. J. Res. Med. &Applied Sci., vol. 1, no. 5, pp. 1–4, 2015.

[2] B. Cunningham and S. Nelson, "Powdery Mildew of Papaya in Hawaii," Plant Dis., no. July, pp. 1–4, 2012.

[3] A. Caricae, S. Maubl, and S. Pal, "Studies on Black Spot of Papaya (Carica papaya L.), 2012.

[4] K.Behera, Santi & Mishra, Namrata & Sethy, Prabira & Rath, Amiya. (2018). On-Tree Detection and Counting of Apple Using Color Thresholding and CHT. 0224-0228. 10.1109/ICCSP.2018.8524363.

[5] H.Zulkifli, A.Hallis, A.Yeon and Rohani, "Feasibility Study on Plant Chili Disease Detection Using Image Processing Techniques," 2012 International Conference on Intelligent Systems Modelling and Simulation ., no. c, pp. 291–296, 2012.

[6] M.Dadwal and V. K. Banga, "Color Image Segmentation for Fruit Ripeness Detection: A Review," 2nd International Conference on Electrical, Electronic and Civil Engineering (ICEECE'2012)., no. c, pp. 190–193, 2012.

[7] R.P.Salunkhel and A.A.Patil, "Image Processing for Mango Ripening Stage Detection: RGB and HSV method," 2015 Third International Conference on Image Infonnation Processing., no. c, pp. 362–365, 2015.

[8] R.P.Salunkhel and A.A.Patil, "Image Processing for Mango Ripening Stage Detection: RGB and HSV method," 2015 Third International Conference on Image Infonnation Processing., no. c, pp. 362–365, 2015.

[9] H.E.Stevens, "Papaya Diseases," no. c, pp. 57–63,1939.

[10] A. K. W. Law, H. Zhu, F. K. Lam, and F. H. Y. Chan, "Tumor Boundary Extraction in Multislice MR Brain Images using Region and Contour Deformation," pp. 1–5, 2001.

[11] A.Ming and Huadong Ma "A blob detector in color images,"pp. 364-370, 2007.

[12] M.Shaban Al-Ani and K.M.Ali Altheeti " Precision Statistical Analysis of Images Based on Brightness Distribution, " Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 4, 99-104 (2017).

[13] C. N. Dsouza, "EDGE Preserving Image Enhancement for Color and Gray Scale Images using Local Mean and Local Standard," no. July, 2017.

[14] D. Surya Prabha and J. Satheesh Kumar, "Assessment of banana fruit maturity by image processing technique," J. Food Sci. Technol., vol. 52, no. 3, pp. 1316–1327, 2015.

[15] P. B. Padol, "Fusion Classification Technique Used to Detect Downy and Powdery Mildew Grape Leaf Diseases," 2016 Int. Conf. Glob. Trends Signal Process. Inf. Comput. Commun., pp. 298–301, 2016.

[16] M. V Latte and S. Shidnal, "Multiple Nutrient Deficiency Detection in Paddy Leaf Images using Color and Pattern Analysis," 2016 Int. Conf. Commun. Signal Process., pp. 1247–1250, 2016.

[17] S. Ramesh and B. Rajaram, "IoT based Crop Disease Identification System using Optimization Techniques," ARPN J. Eng. Appl. Sci., vol. 13, no. 4, pp. 1392–1395, 2018.

[18] S. Puttemans, Y. Vanbrabant, L. Tits, and T. Goedem, "Automated visual fruit detection for harvest estimation and robotic harvesting," pp. 1–6, 2016.

[19] R. Jain and P. K. Johari, "An Improved Approach of CBIR using Color Based HSV Quantization and Shape Based Edge Detection Algorithm," in IEEE International Conference on Recent Trends in ELectronics Information Communication Technology, 2016, pp. 1970–1975.

[20] S. Ramesh and D. Vydeki, "Application of machine learning in detection of blast disease in South Indian rice crops," J. Phytol., vol. 11, pp. 31–37, 2019.

[21] Z. H. Husin, N. S. A. M. Taujuddin, R. Ibrahim, and S. Sari, "Detection and Classification of Malaria Parasite using Morphological Threshold Area Estimation Technique," J. Adv. Res. Dyn. Control Syst., vol. 12, no. Special, pp. 840–845, 2020.

IoT based Automatic Damaged Street Light Fault Detection Management System

Ashok Kumar Nanduri[1], Siva Kumar Kotamraju[2], G L Sravanthi[3], Sadhu Ratna Babu[4], K V K V L Pavan Kumar[5]

Department of Computer Science and Engineering, Vignan's Nirula Institute of Technology and Science for Women
Pedapalakaluru, Guntur-522009, Andhra Pradesh, India[1, 2, 3]
Department of Information Technology, Bapatla Engineering College, GBC Rd, Mahatmajipuram
Bapatla-522102, Andhra Pradesh, India[4]
Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur-522502, Andhra Pradesh, India[5]

Abstract—**The IoT (Internet of Things) is a blooming technology that mainly concentrates on the interconnection of devices or components to one another and the people. As the time being, many of these connections are changing as "Device – Device" from "Human to Device". Finding the faulty street light automatically is become a vital milestone by using this technology. The primary goal of the project is to provide control and identification of the damaged street light automatically. The lighting system which targets the energy and automatic operation on economical affordable for the streets and immediate information response about the street light fault. In general, the damage of the street light is observed by getting the complaints from the colony (street) people. Whereas in this proposed work using sensors these lights working status is easily captured without any manual interaction. So that it reduces manual efforts and the delay to fix problems. So, to reduce such problem we come with the solution wherein automatic detection of street light issues i.e.; whether the street light is working or not will be found at night time and it should send the notification to the authorized person if there is a problem in particular streetlight and also the location of the place where the streetlight is damaged. The street lights are automatically ON/OFF using IoT. In this system, it checks whether the street light is ON/ OFF. The LDR sensor will ON/OFF the street lights automatically, based on the condition of the weather.**

Keywords—*IoT (Internet of Things); GSM (Global System for Mobile); LDR (Light Dependent Resistor); LED (Light-Emitting Diode); GPS (Global Positioning System); Raspberry; Twilio*

I. Introduction

An important and essential role is being played by the Internet of Things in every one's regular life. There is a clear increase in the changes being made among the traditional systems, & other general household components and traditional systems for making a better life [1 to 6]. The major problem of the available system of electricity is the issue of connectivity because of the major connections is manually handled by several contractors. Manually, the settings of the timer are done. Additionally, the timer needs 12 hours' of power supply continuously, setting the timer may interrupt in the loss of power supply continuously [7]. IoT is characterized through largely shared, original world smaller things, and the capability of processing and with the storage of restriction that contains the presentation, confidentiality, consistency, and safety [8].

Pervasive devices are linked using IoT and many networks are convenient for delivering protected and effectual provisions for entire applications anywhere and anytime [9]. The Internet of Technology is combined with IT (Information Technology) and OT (Operational Technology) in which the unstructured data generated by machines are studied for making improvements. According to the research work of Li Da Zu et al. [10], D. Giusto et al. [11] and Saifuzzaman et al. [12] IoT is combining of physical devices with sensors, electronics, and software using the internet, which allows the objects to fetch and interchange the data through the ecosystem. It is determined as IoT because all the components described here are utilized in making a better ecosystem. As it is already known that one of the major assets of the city's is street light which provides road safety and increases the security in city centers and houses too. There are many automation applications of IoT such as smart roads, smart parking, smart lighting, smart home, and many more. In the existing system that is the manual system of street lighting contains many issues like connectivity problems, maintenance problems, and the timing issue. These issues will be solved using the technology of IoT [13]. This system depends on the smart and climatic flexibility of street lighting, management automatically [14]. Several issues are simplified by automation in the economy of the world and regular life [15]. Currently, the system of streetlight flexibility is majorly challenged. Controlling the distant area location is a major dilemma. Human mistakes lead to wastage of energy and the system's low performance [16]. Based on the survey conducted by S. K. Cho et.al. [17], it is determined that 30 percent of electricity is consumed by street lights in each city. Currently, the street lights are will be in ON state before the sunset and they get OFF after getting light in the atmosphere, sometimes the light will be in ON state whole day. As it is our responsibility to save energy, we need to take initiative in saving energy. Our project yields the best solution for electricity wastage. Automatic ON/OFF of the street light is an effective solution that will decrease the consumption of the lights of the streets up to 20 percent when the environment contains light.

II. Related Work

PLC usage in Automated street lighting, usage of PLC in street light handling is a wonderful concept by the XD26 PLC controller [17]. Manual work is not necessary for this system.

LDR plays an important role in changing the state as ON/OFF automatically, as a result of the sunlight. The benefits of this method are seasonal variations effect, growth in the efficiency of energy, less cost for operating and maintenance. Crouzet Millennium software involvement in this project helps in testing and analytics with the exact function of streetlights.

In [18], Street light observing and controlling system is an automated system based on GSM, helps in amplifying the effectiveness and precision of an organization by providing the change in the street lights automatically in schedules, the 2 fundamental modules are client-side and the server-side. The microcontroller is the additional connection to GSM modern of client-side whereas the server-side is made of Java-based web server. Building and implementing the future evolution through implanted systems to save the power of the street light system is the main aim of the street light controlling system automatically using the help of a microcontroller. Nowadays, human life became so busy that there's no time to switch off unnecessary lights.

The best remedy for this expenditure of electric power is described in this paper. Moreover, manual work is abolished completely. LDR (Light Dependent Resistor) and photoelectric sensors are the 2 sensors utilized to identify either light or dark time and to find out the movement on the street respectively. The street lighting system can be controlled using the microcontroller PIC16F877A acts as if the brain, where usage of programming language to execute the software of the microcontroller using the language of C [19].

A new path of diminished power, usage was proposed using the approach of RFID based on GSM for the automatic street light system. The reduction in recovering power failure time can be done by this system [20]. Through this GSM street light, load maintenance and also any complaints regarding power could be warned. Power could be saved by the electric department by adopting this system. This system also moves ahead to minimize the time during the processing of all new connections of power suggestions with the help of RFID.

Automatic street lights project motto is just to control the utility of power on the streets, abolish manual work. In this aspect, the street light circuits consisted of particular sensors, LDR, and microcontrollers the whole day. LDR, sensors, and microcontrollers are the 3 basic components required LDR helps in consistently maintaining the streetlight off during daytime as there is no requirement until there is a low light point or low frequency of streetlights thereby the LDR consists of high resistance [20-22]. This stops the power supply to the transistor's base. Finally, the lights of the street will not be glowing.

During most of the night, Intelligent street lighting systems utilizing GSM, street light systems of conventional in regions with less incidence of passersby are online without any cause. The result of this is the bulk of power is wasted unnecessarily. Street lighting systems became reality because of the vast accessibility of flexible technology of lighting such as the wireless connection of the internet, light-emitting diode lamp, well-grounded functionality, power conservation, and quick reacting. This main theme of this study is to describe that the Intelligent street lighting ISL system was the initial to achieve

the command for the public systems of lighting in a flexible way. Building and implementing the future evolution through implanted systems to save the power of the street light system is the main aim of the street light controlling system automatically using the help of a microcontroller. Nowadays, human life became so busy that there's no time to switch off unnecessary lights. The best remedy for this expenditure of electric power is described in this paper. Moreover, manual work is abolished completely. LDR (Light Dependent Resistor) and photoelectric sensors are the 2 sensors utilized to identify either light or dark time and to find out the movement on the street respectively. The street lighting system can be controlled using the microcontroller PIC16F877A acts as if the brain, where usage of programming language to execute the software of the microcontroller using the language of C.

III. PROPOSED WORK

In the research of the damaged Street Light identification System, we will identify the reduction in power consumption. According to our modern lifestyle energy is the basic requirement for both developing and developed countries. For better understanding, we build a prototype for street lighting to find the effectiveness of our system.

This research work covers all reviews from various sources, which targets different concepts of street light systems and risks of human effort. In this work, we discuss two important concepts and get unique results effectively. The proper operation of the system of street lighting by the addition of other systems will help us to develop a research work which is easy and also consumes less power and cost, which can be accessed by the internet. The driver architecture of LED is everywhere to design a street light system in a smart way, which is reliable and reduces energy waste. They undoubtedly describe the automatic ON/OFF of the lights automatically. Later, Morais et al. and Silva et.al studied about the dimming concepts of light according to IEEE.

2017 R10-HTC (Region 10 Humanitarian Technology Conference), 2017 21-23 December, Dhaka which is the capital of Bangladesh 122 large amount of energy is saved by making the ON/OFF of the street light depended on the object movement by the help of modules (IR modules have been used in this). They increased the intensity of the lamp when the object moves under that light region, which helped in the reduction of power consumption. In the same way, in this project we used the LDR concept to detect the light condition, GPS is used to fetch the exact location, and LED is used to indicate the damaged light. The GPS module contains the data in the raw format; it contains several data in different forms. In this paper, the latitude, and longitude of the region where the light is damaged need to be found. Now the series of data is modified in terms of latitude and longitude; using these easily access the location. Each street light connects with an LDR and a LED. In this, the red-colored LED has been used, which specifies danger in general. Relays are used for connecting with lights. The variation of the lights OFF because of the power loss and damage is known by setting threshold value. Whenever the threshold value coincides, treat it as power loss. If the threshold value does not match it specifies the fault of light in large volume. The cloud storage system monitors

continuously anywhere and anytime. Internet is used to monitor the entire system and all information will be saved in the database for future purposes. The whole system is designed using the concept of IoT as all kinds of sensors, software, cloud, and hardware are combined to make the system effectively. A smart city is one of the booming concepts. Therefore, this work will take one of the initiatives in developing a smart city.

To describe the utility, effectiveness of the entire system, we need to design an architecture that can be used to view the whole functionality in a look as shown in Fig. 1. The architecture, simply says that the entire system can be classified into two parts. The initial part is the operation of street lights and the other is sending information to the authorized person.

Here we use RaspberryPi program for controlling the street light and to detect the fault using the storage system of cloud. The lights of the street are operated manually in general.

Raspberry pi program is utilized for the street lights controlling and the detection of a fault in them through the storage system of the cloud. Present, manually the lights of streets are operated. The control of the street lights and the detection of a fault in them is done through the storage system of the cloud. It checks the ON/OFF of the street lights, and the faulty in the lamps identified. The ON/OFF of the street lights done after checking the weather condition. For checking the weather condition the LDR is used. The LDR (Light Dependent Resistor) checks the condition of weather is light/dark, dark indicates no light, and light indicates the presence of light. If the LDR results in light from the lamp, then consider it as the time of day with light. It is considered as the time of night while it receives results as dark from the LDR. Then the system permits the Switch ON the lamps of the street.

Now the lamps of the street are in ON state. And few lights are in OFF state due to the fault in them. And the system identifies that there is a fault in the lamps, they are not glowing. The lights of the street are not glowing because of some failure; this is known through the LDR. The LDR is inversely proportional to the intensity of light. Now the red LED attached to that street lamp will start glowing. And a message will send to the respective ward member or authorized person. The message includes the message of damage along with the URL.

The URL specifies the exact location of the faulty light. The location of the damaged light is known by GPS. This message will send from the account of Twilio. In the Twilio account, we will specify the contact details of the ward member.

At any point in time, the status of the system will be known anytime, anywhere as we are using the storage system of the cloud. As the lights of the streets are connected to the account of Twilio through the storage system of the cloud. The data of the sensor is stored in the system of the cloud using the account of Twilio. Therefore, the data of the street lights can be accessed easily from the storage of the cloud.
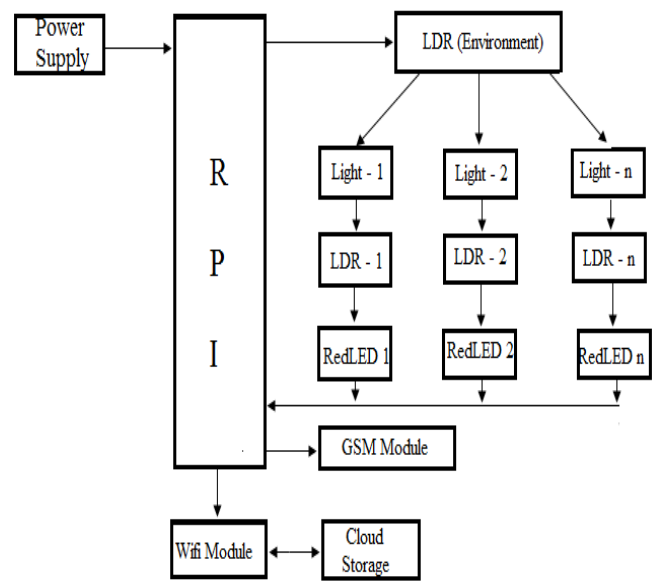
Fig. 1.   Architecture for Automatic Street Light Damage Detection.

A. System Design

- To design an environment that uses IoT in real-time to determine the controlling of the system of street light and testing the working performance of the application.

- The sensor of ambient lights is used to find the intensity experimentally and using that intensity appropriate action of control is taken on the lights.

- To build a library of many modules that replicates the conditional parameters of the network.

- To prepare a GUI (graphical User Interface) and to check the light's status.

If the research work compared with the Automatic street light system, this system saves energy up to 8 percent to 10 percent as it works only on sunrise and set timings. This system even works with the weather condition, which also saves manual effort during winter seasons. The exact location of the damaged light identified easily, whenever there is a fault in the light quickly alert message to the respective authorized person. So that the person can know about the fault and can respond to it. It also reduces the human effort of reporting. The message includes the complaint of the light and the exact location is sending in the form URL. On selecting the URL, this location will get displayed. The system will continue monitoring at all locations through the system of the cloud.

Fig. 2 describes the entire flow of the system. Initially, the LDR checks whether the light of the atmosphere is ON/OFF. No action is performed if it is in the ON state. If the weather's state is OFF, then automatically the light of the street will become ON. If there is any fault in the lamps of the street the red LED attach to each street light, will be ON. Here the color red indicates the damage.
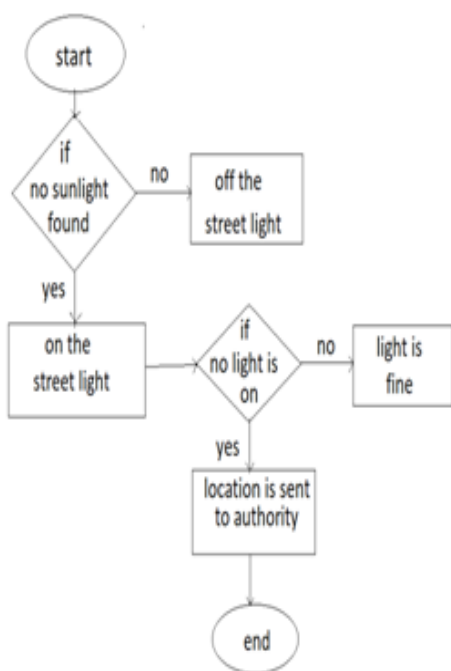
Fig. 2. Data Flow Diagram for Damaged Street Light Process.

Each street light contains their respective LDR and red LED, this LDR is used to sense the street light and ON the red LED automatically, in case of a faulty lamp. The location will send to the authorized person through the Twilio account. The Twilio Account is cloud storage, which stores the data. The description of the message and the mobile number of the authorized person is attached to this. At the time of fault detection, this message will send to that person. The exact location will be fetched through the GPS. These contains several forms of data, but in our project, we are using Bluetooth data which consists of longitude and latitude of the faulty light. After fixing the damaged lamp, the red LED will get OFF automatically.

IV. RESULT AND DISCUSSION

Initially, we design a prototype to estimate the arranging process of the whole system and can be done as part of future research, development. After completion of the entire work, we design a proposed system as displayed in Fig. 3. After developing the system, it has been tested for many months for verifying the whole functionality in real-time. The use of our methodology resulted in the fault detection accuracy and the lights are automatically ON/OFF which also saved electricity.

In the above figure, the first street light is OFF because of the fault in it. At the point of time, the red color LED will start glowing. In the success case of street light ON no action is performed. For our experimental purpose, we used basic LEDs in the place of lights.

Fig. 4 describes the message sent to the authorized person from the Twilio account. Damage is the message which we have written as the indication and the URL is fetched using GPS.
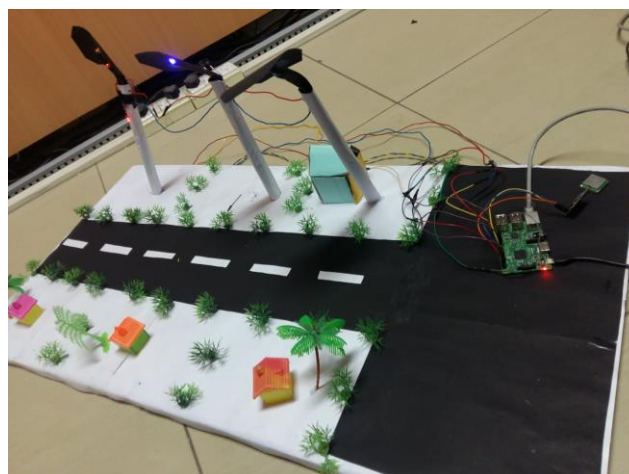
Fig. 3. Overall view of the Research Work.
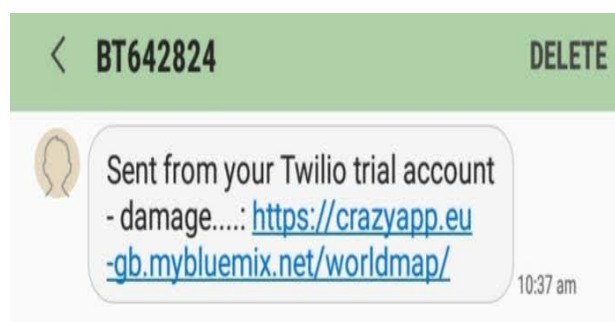
Fig. 4. Output Message from Twilio Account.

Fig. 5 shows the location of the damaged light. The indicator indicates the exact location. In the case of more lights are in one place then the damaged light can be known with the help of a red LED.

Fig. 5. Damaged Street Light Finding Location.

V. CONCLUSION

Nowadays resources (water, power, air, etc.) are very precious. This work focused to protect one such resource i.e. energy. Electricity is one of the major losses of energy. Using IoT the street lights ON/OFF is automated based on the weather condition, the working status of the street light is observed. The LDR sensor senses the environmental changes, the ON/OFF of the street lights is made automatically.

Whenever the street light got damaged or not on during night time, the LDR sensor senses it and sends the notification to the authorized person that the light is damaged and the location (using GPS) where the light is damaged. It reduces human efforts, delays in fixing the issues. The automatic control of street lights is used to find the exact location when the street light gets damaged. Further, this can be implemented for all the street lamps in rural lamps. Pre-identification of damaged street lights is done based on the expiry of lamps.

ACKNOWLEDGMENT

REFERENCES

[1] International Telecommunication Union.(2005). Internet reports 2005: The internet of things. Geneva: ITU.

[2] Issarny, V., Teixeira, T ., and Hachem, S. and (2011). Ontologies for the internet of things (pp. 1–6). New York: ACM.

[3] Suo, H., Wan, J., Li, F., and Yan, H. (2011). "Advances in cyber-physical systems research", KSII Transactions on Internet and Information Systems, 5(11), 1891–1908.

[4] Vasilakos, V., Lai, C., and Tsai, C.(2014). "Future internet of things: Open issues and challenges". ACM/Springer Wireless Networks.

[5] Morabito, G., Iera, A., & Atzori, L. and (2010). "The internet of things: A survey," Computer Networks, 54 (15), 2787–2805.doi: https://doi.org/10.1016/j.comnet.2010.05.010.

[6] Miorandi, D., Chlamtac, I., Pellegrini, F. D., and Sicari, S. (2012). "Internet of things: Vision, applications and research challenges", Ad Hoc Networks, 10(7), 1497-1516.

[7] Sayali Arkade, Akshada Mohite, Rutuj, Vikas, "IoT Based Street Lights For Smart City" International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue XII, December 2016.

[8] Agar Deo, Sachin Prakash and Asha Patil, "ZigBee-based Intelligent Street Lighting System", 2nd International Conference on Devices, Circuits and Systems (ICDCS), 2014.

[9] S. P. Raja and T. Dhiliphan Rajkumar ,"Internet of Things: Challenges, Issues and Applications", Journal of Circuits, Systems, and Computers Vol. 27, No. 9, 2018, pp. 1-16.

[10] Li, S., Zu, L.D., and He, W., (2014)," Internet of Things in Industries: A Survey" IEEE Transactions on Industrial Informatics, 10(4).doi: 10.1109/TII.2014.2300753.

[11] Atzori, L., Iera, A., Giusto, D., and Morabito, G. (2010). "The Internet of Things", Springer, 2010. ISBN: 978-1-4419-1673-0.

[12] Saifuzzaman, M., Nur, F.N., Moon, N.N., and Khan, A.H. (2017). "Smart Security for an Organization based on IoT", International Journal of Computer Applications 165(10), 33-38.

[13] Sayali Arkade, Akshada Mohite, Rutuj, Vikas, "IoT Based Street Lights For Smart City" International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue XII, December 2016.

[14] B.Abinaya, S.Guru Priya, "IOT BASED SMART AND ADAPTIVE LIGHTING IN STREET LIGHTING" Department of Information Technology Sri Sairam Engineering College.

[15] Sindhu.A.M, Jerin George, Sumit Roy, Chandra J, "Smart Streetlight Using IR Sensors" IOSR Journal of Mobile Computing & Application (IOSR-JMCA) e-ISSN: 2394- 0050, P-ISSN: 2394-0042.Volume 3, Issue 2.

[16] Vaishali Gupta, Krutika Thakur, Ritesh Thakur, "Based Smart Street Lights" International Journal of Research (IJR), Volume 2, Issue 10, October.

[17] Dhingra, V. & Cho, S. K. (2008). "Street Lighting Control Based on Lon Works Power Line Communication," IEEE In- ternational Symposium on Power Line Communications and Its Applications, Jeju City, 2-4 April 2008, pp. 396- 398.

[18] Amey Manekar, J., Dr. Kshirsagar, Dr. R .V. (2016). " Design and Implementation of Automatic Street Light Controller for Energy Optimization Using FPGA", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016.

[19] Paridhi, Holani., Chaitanya, Amin., Rahul, Kaul., and Ashutosh, Nerkar. (2013). "GSM Based Autonomous Street Illumination System for Efficient Power Management" International Journal of Engineering Trends and Technology- Volume4Issue1.

[20] Sachin Datta, N S., Abdul Latif Saleem., Sachin, H S., Raja Sagar, R., and Usha, M, S. (2015) "Street Light Monitoring and Control System" International Journal of Engineering and Techniques - Volume 1 Issue2.

[21] Monica, Pujari., Rajput, K,Y., Priyanka, Yadav., and Gargeyee, Khatav. (2013) " Intelligent Street Lighting System Using Gsm" International Journal of Engineering Science Invention Volume 2 Issue 3.

[22] Tarun Kumar, B., Sumathi, V., and Krishna Sandeep, A., (2013) "Arm Based Street Lighting System with Fault Detection" International Journal of Engineering and Technology- Vol 5 No 5 Oct-Nov 2013.

[23] Lopes, J.P., deSouza, I.H., Gobbato, C., Denardin, G.W., Kohler, S.V. (2016). "Integrated topology of DC converter for street lighting system based on LED modular drivers", Industry Applications (INDUSCON), 12th IEEE International Conference.

Lake Data Warehouse Architecture for Big Data Solutions

Emad Saddad[1]
Climate Change Information Center and Renewable Energy
and Expert System
Agricultural Research Center (ARC)
Giza, Egypt

Ali El-Bastawissy[2]
Faculty of Computer Science
MSA University
Giza, Egypt

Hoda M. O. Mokhtar[3]
Faculty of Computers and Artificial Intelligence
Cairo University
Giza, Egypt

Maryam Hazman[4]
Climate Change Information Center and Renewable Energy
and Expert System
Agricultural Research Center (ARC), Giza, Egypt

Abstract—**Traditional Data Warehouse is a multidimensional repository. It is nonvolatile, subject-oriented, integrated, time-variant, and non-operational data. It is gathered from multiple heterogeneous data sources. We need to adapt traditional Data Warehouse architecture to deal with the new challenges imposed by the abundance of data and the current big data characteristics, containing volume, value, variety, validity, volatility, visualization, variability, and venue. The new architecture also needs to handle existing drawbacks, including availability, scalability, and consequently query performance. This paper introduces a novel Data Warehouse architecture, named *Lake Data Warehouse Architecture*, to provide the traditional Data Warehouse with the capabilities to overcome the challenges. *Lake Data Warehouse Architecture* depends on merging the traditional Data Warehouse architecture with big data technologies, like the Hadoop framework and Apache Spark. It provides a hybrid solution in a complementary way. The main advantage of the proposed architecture is that it integrates the current features in traditional Data Warehouses and big data features acquired through integrating the traditional Data Warehouse with Hadoop and Spark ecosystems. Furthermore, it is tailored to handle a tremendous volume of data while maintaining availability, reliability, and scalability.**

Keywords—Traditional data warehouse; big data; semi-structured data; unstructured data; novel data warehouses architecture; Hadoop; spark

I. INTRODUCTION

Data warehouse (DW) has many benefits; it enhances Business Intelligence, data quality, and consistency, saves time, and supports historical data analysis and querying [1]. In the last two decades, data warehouses have played a prominent role in helping decision-makers. However, in the age of big data with the massive increase in the data volume and types, there is a great need to apply more adequate architectures and technologies to deal with it.

Therefore, it became crucial to enhance traditional DW to deal with big data in various fields to accommodate this evolution in volume, variety, velocity, and veracity of big data

[2], [3]. To achieve this, we propose a new DW architecture called *Lake Data Warehouse Architecture*. *Lake Data Warehouse Architecture* is a hybrid system that preserves the traditional DW features. It adds additional features and capabilities that facilitate working with big data technologies and tools (Hadoop, Data Lake, Delta Lake, and Apache Spark) in a complementary way to support and enhance existing architecture.

Our proposed contribution solve several issues that face integrating data from big data repositories such as:

- Integrating traditional DW technique, Hadoop Framework, and Apache Spark.

- Handling different data types from various sources like structured data (DBs, spreadsheet), semi-structured data (XML files, JSON files), and unstructured data (video, audio, images, emails, word, PowerPoint, pdf files).

- Capturing, storing, managing, and analyzing data volume that cannot be handled by traditional DW.

- Using recent technologies like the Hadoop framework, Data Lake, Delta Lake, and Apache Spark to decrease time spent analyzing data and to decrease storage costs are inexpensive.

- Support all users, especially data scientists, because they need to perform depth data analysis.

The rest of the paper is organized as follows: Section II explains background and preliminaries for traditional Data Warehouses and its limitations, importance of DWs, and big data characteristics and types. Section III overviews the related works to the DW architectures. Section IV presents the proposed DW architecture named Lake Data Warehouse Architecture. Section V describes the case study by applying our contributions called *Lake DW Architecture*. Finally, the conclusion is presented in Section VI.

II. BACKGROUND AND PRELIMINARIES

In this section, we will review the background of related topics, such as traditional DWs, its limitations, and why we need to redevelop it. Moreover, we will discuss various related technologies.

A. Traditional Data Warehouses(DWs)

Traditional DWs are integrated, a subject-oriented, nonvolatile, and time-variant data to support decision-makers [4], as presented in Fig. 1. Those four properties differentiate DWs from other data repository systems, such as relational database systems [5].
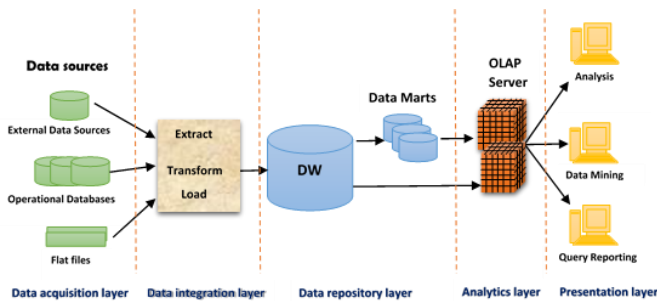

Fig. 1. Traditional Data Warehouses (DWs) architecture

Usually, in primary DW, there are data marts (DMs). Data marts (DMs) are a small DW that contains only a subset of data obtained from a central DW. The content of DMs represents information from a specific domain of interest. Many DWs servers are used to manage data. These servers present multidimensional views of data to a variety of front-end tools.

B. The limitations of Traditional DW Architecture

In [6], [7], and [8], the authors reviewed some limitations of DW, such as supporting only structured data, on the contrary, do not support semi-structured data, or unstructured data. In addition to the above, the restriction of handling data volume in terabytes, which does not scale to petabyte size, is widely available. Besides, it is costly as it depends on proprietary hardware and software.

Moreover, traditional DW performs analytic queries, which consequently affects the whole query performance, accessing, and processing of data. The decision-making process also may be affected if: 1) the correct data are not available at a suitable time, and 2) the growth of the business requires new methods for data management other than adapting traditional DW architecture.

C. The Objectives of the redeveloped traditional DW

One of our main objectives is to overcome the limitations of traditional DW architecture built on outdated technologies [4]. We initiate an overall architecture that supports the functionalities of the traditional DW with abilities to include [6], [9] :

- Meeting new business requirements.
- Depending on lower-cost infrastructure.
- Handling heterogeneous data and new data structures and formats.

- Managing customer expectations.
- Meeting growth of the business.
- Using advances in new technology and its improvements.
- Handling product existence and status.
- Improving business efficiencies and competitive advantage.
- Decreasing operational and financial risks.
- Evaluating and forecasting trends and behaviors.

D. The Age of Big data

Big Data is a data volume that is available in different levels of complication. It is generated at various velocities and levels of uncertainty; hence it is not handled using traditional approaches, traditional technologies, or traditional algorithms [6]. Today, big data is characterized by ten main characteristics namely volume, variety, velocity, veracity, value, validity, variability, visualization, volatility, and venue [3], [10], [2], [11], [12], [13] as follows:

- *Volume*: the huge amount of data generated continuously on an hourly or a daily basis from different sources. Such as terabytes generated per hour for applications like YouTube, Instagram, and Google.
- *Variety*: the types of big data that are ingested from different data sources.
- *Velocity*: the speed at which data is produced. The aspects of data may be batch data or streaming data.
- *Veracity*: the quality of data that is being handled to obtain valuable insights. Such as ambiguous, inconsistent, incomplete, anomaly, uncertain, and biased data.
- *Value*: represents the business value to be derived from big data.
- *Validity*: refers to the correctness of data used to extract outputs in the form of information.
- *Variability*: refers to the inconsistent data flow.
- *Visualization*: refers to the ability to analyze and visual insights as an output of big data analysis.
- *Volatility*: the stored data on how long it is valuable to the consumer.
- *Venue*: refers to a various platform where numerous kinds of data from different sources by several platforms.

In general, data are a set of qualitative values or quantitative variables; Big Data can be categorized into three types [2], [9]:

- *Structured data*. The data has a defined structure or a schema organized either in the form of a relational database or in some other way that is easy to operate. For example, data stored in a relational database (in the

form rows and columns), in spreadsheets (such as CSV files), and cleansed data (that have been processed with a lot of cleansing and filtering).

- *Semi-Structured Data*: The data is hard to retrieve, analyze, and store as structured data. It requires a big data software framework (such as Apache Hadoop) to achieve these operations. For example, XML files, JSON files, and BibTex files.

- *Unstructured Data*: The fully unorganized data is difficult to handle, and it requires advanced software and tools to access it. Examples include video, audio, images, emails, word, PowerPoint, pdf files, webpages, location coordinates, and streaming data.

E. Hadoop Framework and Data Lake

Hadoop is an open-source software framework. It allows the execution of the MapReduce processes for data processing. It provides massive storage for all data types, massively parallel processing, and storing data and running applications on clusters to accomplish better computation resources [14]. Hadoop has main components as follows [15], [16], [17]:

- *The Hadoop Distributed File System (HDFS)* is a file system, which manages storage and access to data spread across the different nodes of a Hadoop cluster.

- *YARN* is a Hadoop cluster resource manager used to assign system resources for applications and schedule of the jobs.

- *Map-Reduce* is a processing engine and a programming framework used to manage large-scale batch data in the Hadoop system.

- *Hadoop Common* is a set of services and institutions that provide underlying capabilities needed by the other parts of Hadoop.

Data Lake is a data store that can collect any type of data: structured, semi-structured, or unstructured data, which are stored with one another regardless of structure, format, or types [18], [19]. It is a conceptual idea that is usually implemented with one or more technologies such as Hadoop and NoSQL databases. When querying the Data Lake, only need data will transform that are relevant to business needs [20].

Creating Data Lake depends on Hadoop's technology, which is a component (as the platform) for the data lake. It is the complementary relationship between Data Lake and Hadoop [21], [22], [23]

Data Lake is similar to traditional DW in that they are both repositories for data. However, there are apparent differences in features between them [7]—the schema on reading in Data Lake, but schema on write in DW. The scale of data in Data Lake is enormous, while it is large data volumes in DW. The data sources may be semi-structured data or/and unstructured data, but it is mainly structured data in DW [24], [25], [26], [27].

F. Apache Spark and Delta Lake

Apache Spark is an open-source applied for big data analytics and distributed systems. It provides streaming libraries, SQL, graph analysis, and machine learning. It has two main components Spark streaming, which is used for managing real-time data, and the Spark engine , which directly processes each data chunks by Spark streaming [28], [29], [30].

Delta Lake is an open-source Spark storage layer. It is an extra storage layer that makes reliability to our data lakes built on The Hadoop Distributed File System (HDFS) and cloud storage [31]. Delta Lake provides a series of other features including:

- Joining streaming and batch data processing.

- Giving a scalable metadata approach.

- Providing ACID transactions that guaranteed consistency of the data stored inside the data lake through ensuring that only complete writes are committed.

- Time travel that is allowing one to access and return previous versions of the data.

- Schema evolution as data evolves, Delta allows Spark table to change in the schema and many more while we use Delta.

- Enabling a Data Lake to update data without going through the entire Data Lake repository.

III. RELATED WORKS

Several efforts have been conducted to adapt traditional DWs for handling new user requirements and changes in the underlying data sources. Many approaches focus on DWs that deal with a relational database. However, they cannot be appropriated to deal with big data. In [32], some methodologies try to solve the problem by developing the ETL (Extracting, Transforming, and Loading) process. In [33], the authors attempt to update DW Schema to reflect modifications that already took place. As mentioned in [33], [33], [34], [5], [35], [36], the authors use temporal DW and schema versions to update the DWs Structure by keeping more than one DWs version. These works do not depict how the user's needs impact the evolution changes.

Limited approaches have taken care of handling the aspect of big data development. The approach mentioned in [37] describes data schema specification and evolution processing, but it does not depend on DW. The work presented in [38] explained the method for treating the growth of the data sources in the integration area using big data integration ontology. It can handle some changes in data sources, but it does not determine how to answer all requirements.

Several types of research have concentrated on the use of DWs in big data analysis. The authors in [39] display an OLAP method for big data executed with Hadoop. They focus on multidimensional analysis for big data analysis. However, they do not address big data evolution, which applies to the research work proposed in [40] and [26].

Other researchers studied the problem of big data evolution. In [41], the authors discuss a methodology for constructing a system for big data analysis. However, it cannot be applicable in the state of the previously used data is not provided.

The authors in [42] presented the DW approach for Big Data analysis that was implemented using Map-Reduce. This approach handles two types of changes: (1) schema versions in metadata control variations of the fact table and (2) slowly changing dimensions. The approach does not process changes in big data sources that may affect the analysis process and results.

IV. The proposed Lake Data warehouse Architecture

Our contribution aims to adapt traditional DW to solve the new challenges by handling semi-structured and unstructured data. In addition to managing the growth of business requirements and treating the two main drawbacks, namely, availability and system performance. Furthermore, it enhances query performance by providing the required data from users at any time.

In this section, we present a novel DW architecture that improves the traditional DW performance to deal with these challenges. Our contribution integrates the traditional DW architecture with big data technologies like Hadoop and Apache Spark.

In the age of big data, We need to improve DW architecture to handle the new challenges imposed by big data. The Hadoop Framework and Apache Spark are a complement to handling the challenges of traditional DW; each has its advantages in different circumstances. In some cases, we need the Hadoop Framework and Apache Spark to process unstructured or semi-structured data (raw data) and large volume datasets (big data). In other words, we still depend on traditional DW for consistent and high-quality data (structured data), and low latency and interactive reports.

In Fig. 2, we explain the proposed Lake DW Architecture. Where Hadoop and Spark do not replace traditional DW and Big Data is going to change traditional DW architecture but not replacing it. Our contribution depends on integrating traditional DW techniques, Hadoop Framework, and Apache Spark into a hybrid solution.

The large amount of big data generated every minute and every hour needs a data lake that can scale to handle this volume. Therefore, we use Hadoop as a data platform for data lakes to provide extensive scalability at an acceptable cost. Besides, Data Lakes can be complemented DW besides the Hadoop framework. Also, Data Lakes can be complemented DW besides the Hadoop framework. Our proposed architecture differentiates itself from all previous work. It is a hybrid environment that upgrades traditional DW with Hadoop environment depending on the Hadoop-based Data Lake because it can extend the use and capabilities of traditional DW, as follows:

- *Collecting or capturing data from structured data sources using ETL Architecture.* DW depends on the traditional ETL process; where extract (E) data from operational databases and then, the data process, clean and transform (T) before loading (L) them into the DW or data marts or virtual data marts. DW is designed to handle and analyze read-heavy workloads. DW needs to define the data model before loading the data. Then, they call a Schema-On-Write approach, as presented in Fig. 2.

- *Collecting or capturing data from Semi-Structured or Unstructured data sources using ELT Architecture.* Big Data requires a different process to collect data where traditional ETL does not work well on semi-Structure or unstructured data. Big Data calls for ELT. The raw data will be stored in its original format. The preprocessing step will not be used until the query or other application acknowledge/ ask for these data. Where Data Lake is different from DW through the processing of data in the ELT order and utilizing the Schema-on-Read approach, as shown in Fig. 2.

Fig. 2 presents a set of essential components of our contribution model as follows:

A. Hadoop-Based Data Lake architecture in Cloud Environment

It is using Hadoop as a staging area for DW by adding Hadoop-based Data Lake that is a storage repository that is used for complementing traditional DW. It is using as a data source that passes only required data to Data Lake and ingesting unlimited amounts of raw data that related to business objectives. As shown in Fig. 2, we explore the Hadoop-based Data Lake architecture has many layers as follows:

1) Ingesting data layer: ingests raw data in native format, where the ingested data can be micro-batch, macro-batch, batch, real-time, or hybrid.

2) Landing data layer: Different data types (ingested in the previous layer) are stored in native format (without processing). In this layer, users can find the original data versions of their analysis to aid the subsequent handling.
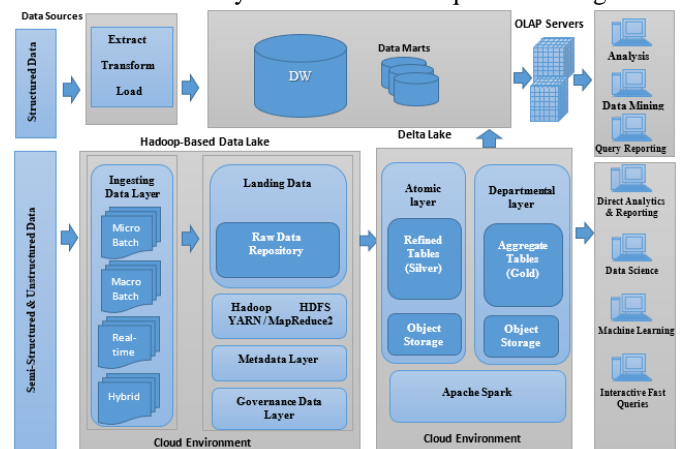

Fig. 2. The proposed Lake Data Warehouse (DW) Architecture

3) Metadata Layer: is responsible for making data easy to access and extract values from Data Lake. It helps to make

identifications, data versioning, entity and attributes, distributions, quality.

4) Governance Data Layer: applies to the other layers. It is responsible for authentications, data access, data security, data quality, and data life cycle. It determines the responsibilities of governing the right to access and handling the data.

B. Delta Lake architecture with Apache Spark Cloud Environment

Delta Lake technology is used with Apache Spark to implement our proposed model by creating a cloud data platform for solving Data Lake challenges, such (1) the data quality is low, (2) reading and writing are not guaranteed, (3) insufficient performance with growing volumes of data, and (4) updating records is hard [43]. As presented in Fig. 2, Delta Lake Architecture has two layers:

1) The atomic layer would be a Silver Delta table built using Object Storage, and

2) The Departmental layer would be any number of Gold Delta tables also built on Object Storage. We employ Spark and store all data in Apache Parquet format allowing Delta Lake to leverage the well-organized compression native to Parquet.

Apache Spark is used to read and process huge files and datasets. Spark provides a query engine capable of processing data in huge data files. Some of the most significant Spark jobs in the world run on Petabytes of data. Apache Parquet has the format as a columnar file responsible for optimizations to go faster queries [44]. It is a more efficient file format than JSON files. It is suitable for data processing in the Hadoop. It provides an efficient method to handle complex datasets.

The main difference between our contributions Lake DW Architecture over traditional DW as follows:

- Handling different data types (structured, semi-structured, and unstructured data) from various sources.

- Extracting, storing, managing, and analyzing data volume that cannot handle by traditional DW.

- Integrating between traditional DW technique, Hadoop Framework, and Apache Spark as a hybrid solution. It uses ETL or ELT processes depending on types of data sources.

- Supporting different data types in various sources.

- Determining and analyzing data from Data Lake and Delta Lake, they scale to extreme data volumes.

- Supporting all users, especially Data scientists, because they can do in-depth analysis.

V. CASE STUDY

Our goal of this section is to experiment with our contribution to prove its effectiveness in dealing with big data and analyze it for decision-makers. This case study shows how our proposed model can extract, integrate, and analyze big data that cannot be handled by traditional DW. We use Data Lake to collect and process the Internet of Things (IoT) data. We provide a demo IoT sensor dataset for demonstration purposes. The data simulates heart rate data measured by health tracker devices. Each file consists of five users whose heart rate is measured each hour, 24 hours a day, every day. We store datasets in the data lake as JSON files. We use two data files in JSON format, the first file for readings recorded by the devices in January 2020 (*health_tracker_data_2020_01.json*), and the second file for the readings recorded by the devices for February 2020 (*health_tracker_data_2020_02.json*).

We implemented on Data Lake, Delta Lake, and Apache Spark in Databricks, which provides an integrated platform for working with Apache Spark. When working with Delta Lake, Parquet files can be converted in-place to Delta files. Next, we will convert the Parquet-based data lake table we created previously into a Delta table.

A. Configure Apache Spark

1) Creating the ClustreHelthTracher cluster is computation resources used to run data science and data analytics such as the ETL process, ad-hoc analytics, and streaming analytics.

2) Configuration the Apache Spark, we will need to perform a few configuration operations on the Apache Spark session to get optimal performance. These will include creating a database to store data, as shown in the following Spark SQL script:

```
1  CREATE DATABASE IF NOT EXISTS DB_Health_Tracker;
2  USE DB_Health_Tracker
```

3) In additional to configure the number of shuffle partitions as shown in the following Spark SQL script:

```
1  SET spark.sql.shuffle.partitions=8
```

B. Importing data from a Data Lake into a Delta Lake by ETL of Apache Spark

We import data from our Data Lake and save it into a Delta Lake as Parquet files. We appended the first month of records and kept in the *health_tracker_data_2020_02.json* file by using the ETL process of Apache Spark. In additional to configure the number of shuffle partitions as presented in the following Python language scripts:

```
1  %python
2  # File location and type
3  file_location = "/FileStore/tables/health_tracker_data_2020_01.json"
4  file_type = "json"
5
6  df = spark.read.format(file_type) \
7    .option("inferSchema", infer_schema) \
8    .option("header", first_row_is_header) \
9    .option("sep", delimiter) \
10   .load(file_location)
11
12 display(df)
```

```
1  %python
2  # Create a view or table
3
4  temp_table_name = "health_tracker_data_2020_01"
5
6  df.createOrReplaceTempView(temp_table_name)
```

In Fig. 3, we visualize the sensor data over time, as displayed in the following Spark SQL script:
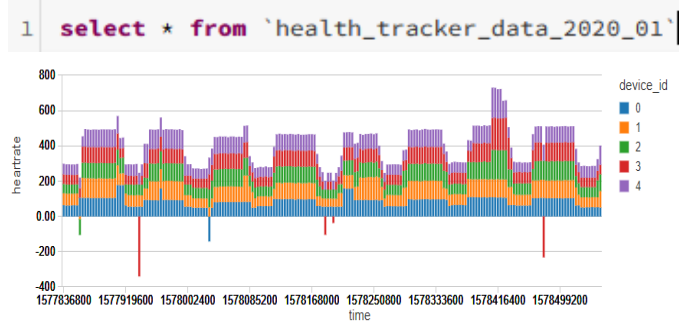

Fig. 3. The sensor data over time.

C. Create a Parquet-based Data Lake Table

We convert the existing Parquet file to The Parquet-based data lake table that will be used to Delta tables. We will be writing files to the root location of the Databricks File System (DBFS) in our cloud object storage. We create the table using the *Create Table As Select* (CTAS) Spark SQL pattern as shown in the following script:

```
1  DROP TABLE IF EXISTS health_tracker_silver;              -- ensures that if we run this again, it won't fail
2
3  CREATE TABLE health_tracker_silver
4  USING PARQUET
5  PARTITIONED BY (p_device_id)                             -- column used to partition the data
6  LOCATION "/DB_Health_Tracker/DLRS/healthtracker/silver"  -- location where the parquet files will be saved
7  AS (
8    SELECT name,                                           -- query used to transform the raw data
9           heartrate,
10          CAST(FROM_UNIXTIME(time) AS TIMESTAMP) AS time,
11          CAST(FROM_UNIXTIME(time) AS DATE) AS dte,
12          device_id AS p_device_id
13   FROM health_tracker_data_2020_01
14 )
```

Count the records in the *health_tracker_silver* table. We expect to have 3720 records: five device measurements, 24 hours a day for 31 days, as shown in the following Spark SQL script:

```
1  SELECT COUNT(*) as NumberOfRecord FROM health_tracker_silver
```

▸ (1) Spark Jobs

count(1)
3720

D. Convert an Existing Parquet-based Data Lake Table to a Delta table: The Atomic layer of Delta Lake

A Delta table consists of three things: (1) The Delta files containing the data and kept in object storage. (2) A Delta table registered in a central Hive meta-store accessible by all clusters to persist table metadata. (3) The Delta Transaction Log (an ordered record of every transaction) saved with the Delta files in object storage. To Convert an Existing Parquet-based Data Lake Table to a Delta table by using the following steps:

1) Convert the Files to Delta Files: We convert the files in place to Parquet files. The conversion creates a Delta Lake transaction log that tracks the files. Now, the directory is a directory of Delta files, as shown in the following Spark SQL script:

```
1  CONVERT TO DELTA
2    parquet.`/DB_Health_Tracker/DLRS/healthtracker/silver`
3    PARTITIONED BY (p_device_id double)
```

2) Register the Delta Table: we will register the table in the Metastore. The Spark SQL command will automatically infer the data schema by reading the Delta files' footers, as shown in the following Spark SQL script:

```
1  DROP TABLE IF EXISTS health_tracker_silver;
2
3  CREATE TABLE health_tracker_silver
4  USING DELTA
5  LOCATION "/DB_Health_Tracker/DLRS/healthtracker/silver"
```

With Delta Lake, the Delta table is ready to use the transaction log stored with the Delta files containing all metadata needed for an immediate query. We count the records in the *health_tracker_silver* table with a Spark SQL query as follows:

```
1  SELECT COUNT(*) FROM health_tracker_silver
```

▸ (2) Spark Jobs

count(1)
3720

E. Creating an aggregate Delta Table: The Departmental layer of Delta Lake

We create a new Delta table (an aggregate table) from the data in the *health_track_silver* Delta table. We create a *health_tracker_user_analytics* Delta table of summary statistics for each device. We use the Create Table As Select (CTAS) Spark SQL as shown in the following script:

```
1  DROP TABLE IF EXISTS health_tracker_user_analytics;
2
3  CREATE TABLE health_tracker_user_analytics
4  USING DELTA
5  LOCATION '/DB_Health_Tracker/DLRS/healthtracker/gold/health_tracker_user_analytics'
6  AS (
7    SELECT p_device_id,
8           SUM(heartrate) AS sum_heartrate,
9           AVG(heartrate) AS avg_heartrate,
10          STD(heartrate) AS std_heartrate,
11          MIN(heartrate) AS min_heartrate,
12          MAX(heartrate) AS max_heartrate
13
14   FROM health_tracker_silver GROUP BY p_device_id
15 )
```

F. Exploring analysis results

The health_tracker_user_analytics table could be used to define a dashboard for analyzing of the results according to business requirements as provided in Fig. 4 which describes the aggregation results such as maximum, minimum, and average data, as presented in the following Spark SQL script:
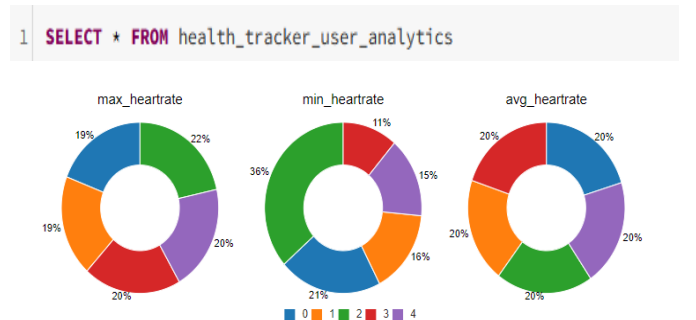

Fig. 4. The analysis of the results according to business requirements.

G. *Appending Files to an Existing Delta Table (Batch data write to Delta Tables)*

We convert the existing Parquet file to The Parquet-based data lake table that will be used to Delta tables. We will be writing files to the root location of the Databricks File System (DBFS) in our cloud object storage. We create the table using the *Create Table As Select* (CTAS) Spark SQL pattern as shown in the following script:

H. *Exploring analysis results*

We can modify existing Delta tables through appending files to an existing directory of Delta files. We append the next month of records, kept in *the health_tracker_data_2020_02* table by using *the INSERT INTO* Spark SQL command as shown in the following script:

```
1   INSERT INTO health_tracker_silver
2   SELECT name,
3          heartrate,
4          CAST(FROM_UNIXTIME(time) AS TIMESTAMP) AS time,
5          CAST(FROM_UNIXTIME(time) AS DATE) AS dte,
6          device_id as p_device_id
7   FROM health_tracker_data_2020_02
```

I. *Assessing the Missing Records*

After a batch update of the *health_tracker_silver* table, we counted the number of records in the table. We discovered that some records were missing by *Count the Number of Records per Device.*

```
1   SELECT p_device_id, COUNT(*) FROM health_tracker_silver GROUP BY p_device_id
```

J. *Assessing the Missing Records*

After a batch update of the health_tracker_silver table, we counted the number of records in the table. We discovered that some records were missing by Count the Number of Records per Device.

TABLE I. THE NUMBER OF RECORDS PER DEVICE

Device ID	Number of Records
0	1440
1	1440
2	1440
3	1440
4	1345

In Table I, device number 4 looks are missing 95 records. We run a query to discover the missing records' timing by displaying the number of records per day. We have no records for device 4 for the last few days of the month, as shown in the following Spark SQL query:

```
SELECT dte as Date, p_device_id as Device_Id,
heartrate as Heart_Rate_Reading
FROM health_tracker_silver
WHERE p_device_id IN (1, 4) and dte > "20-2-2020"
```

In Fig. 5, the absence of records from the last few days of the month shows a phenomenon that may often occur in a production data pipeline: late-arriving data. Delta Lake allows us to process data as it arrives and is prepared to handle the occurrence of late-arriving data.
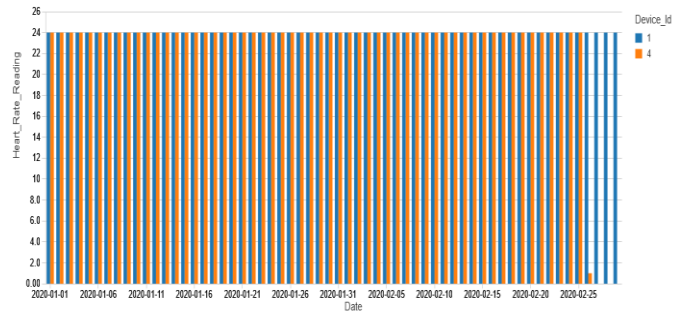
Fig. 5. The absence of records.

K. *Identify Broken Readings in the Table*

In the initial load of data into the *health_tracker_silver* table, we noted that there are broken records in the data. In specific, we made a note of the fact that several negative readings were present even though it is impossible to record a negative heart rate. Let us assess the extent of these broken readings in our table. First, we create a temporary view for the broken readings in the health_tracker_silver table, as shown in the following Spark SQL script:

```
1   CREATE OR REPLACE TEMPORARY VIEW broken_readings
2   AS (
3     SELECT COUNT(*) as broken_readings_count, dte FROM health_tracker_silver
4     WHERE heartrate < 0
5     GROUP BY dte
6     ORDER BY dte
7   )
```

Next, we sum the records in the view, as shown in the following Spark SQL query:

```
SELECT SUM(broken_readings_count)
FROM broken_readings
```

Result: 67

VI. CONCLUSION

Many companies use DWs in different areas to help in the decision-making process. Besides, DWs enhance business intelligence, data quality, and consistency, saving time, and storing historical data. In the age of Big Data, the amount of data needed for ingesting and storing is an unprecedented rate. However, the architecture of the traditional DWs cannot manage such large amounts of data. Its new types from the current data sources are autonomous, heterogeneous, scalable, and distributed, which requires modern technology and modern DW architecture to deal with it.

In this paper, we proposed a novel DW architecture called Lake Data Warehouse Architecture. Lake Data Warehouse Architecture is a new DW structure that depends on integrating between traditional DW technique, Hadoop Framework, and Apache Spark as a hybrid solution. It familiarizes a traditional DW. It employs big data technologies and tools (Hadoop, Apache Spark, Data Lake, and Delta Lake) in a complementary way to support and enhance existing architecture. Furthermore, it can improve scalability and reduce the costs of development of traditional DW architecture.

Lake Data Warehouse Architecture has many competencies over traditional DWs Architecture, such as solving several issues that face integrating data from big data repositories,

handling different data types from various sources. Like Structured data (DBs, spreadsheet), Semi-structured data (XML files, JSON files), and Unstructured data (video, audio, images, emails, word, PowerPoint, pdf files). Moreover, it captures, stores, manages, and analyzes data volume that cannot be handled by traditional DW. Furthermore, it uses the recent technology such as the Hadoop Platform, Data Lake, Delta Lake, and Apache Spark is inexpensive to minimize the time spent analyzing data and to reduce storage costs.

REFERENCES

[1] F. Silvers, Building and maintaining a data warehouse. CRC Press, 2008.

[2] T. John, Data Lake for Enterprises. Packt Publishing Ltd, 2017.

[3] M. Z. Zgurovsky and Y. P. Zaychenko, Big Data: Conceptual Analysis and Applications. Springer Nature Switzerland AG, 2020.

[4] W. H. Inmon, Building the Data Warehouse, Fourth Edition, vol. 13, no. 401. 2005.

[5] E. Saddad, A. El-Bastawissy, O. Hegazy, and M. Hazman, "Towards an alternative Data Warehouses Architecture," in 14th International Conference on Hybrid Intelligent Systems (HIS 2014), Kuwait, December 14-16, 2014 (IEEE, Poster), 2014, vol. 6, pp. 48–53.

[6] K. Krishnan, Data Warehousing in the Age of Big Data. Elsevier Inc., 2013.

[7] R. G. Goss and K. Veeramuthu, "Heading towards big data building a better data warehouse for more data, more speed, and more users," in ASMC 2013 SEMI Advanced Semiconductor Manufacturing Conference, 2013, pp. 220–225.

[8] A. Sebaa, F. Chikh, A. Nouicer, and A. Tari, "Research in Big Data Warehousing using Hadoop," J. Inf. Syst. Eng. Manag., vol. 2, no. 2, pp. 1–5, 2017.

[9] W. H. Inmon and D. Linstedt, Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault. 2014.

[10] C. L. Philip Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, vol. 275. Elsevier Inc., 2014.

[11] V. K. A. Arockia Panimalar. S, Varnekha Shree. S, "The 17 V's of Big Data," Int. Res. J. Eng. Technol., vol. 4, no. 9, pp. 3–6, 2017.

[12] N. Khan, M. Alsaqer, H. Shah, G. Badsha, A. A. Abbasi, and S. Salehian, "The 10 Vs, issues and challenges of big data," ACM Int. Conf. Proceeding Ser., no. March, pp. 52–56, 2018.

[13] R. Rialti and G. Marzi, Ambidextrous Organizations in the Big Data Era. Cham: Springer International Publishing, 2019.

[14] A. Khaleel and H. Al-Raweshidy, "Optimization of Computing and Networking Resources of a Hadoop Cluster Based on Software Defined Network," IEEE Access, vol. 6, no. c, pp. 61351–61365, 2018.

[15] I. B. Lars George, Jan Kunigk, Paul Wilkinson, Architecting Modern Data Platforms: A Guide to Enterprise Hadoop at Scale. O'Reilly Media, 2019.

[16] Z. Yang and X. Guo, "Teaching Hadoop Using Role Play Games," Decis. Sci. J. Innov. Educ., vol. 18, no. 1, pp. 6–21, 2020.

[17] J. Liu, S. Tang, G. Xu, C. Ma, and M. Lin, "A Novel Configuration Tuning Method Based on Feature Selection for Hadoop MapReduce," IEEE Access, vol. 8, pp. 63862–63871, 2020.

[18] C. Walker and H. Alrehamy, "Personal Data Lake with Data Gravity Pull," Proc. - 2015 IEEE 5th Int. Conf. Big Data Cloud Comput. BDCloud 2015, pp. 160–167, 2015.

[19] P. P. Khine and Z. S. Wang, "Data lake: a new ideology in big data era," ITM Web Conf., vol. 17, no. December, p. 03025, 2018.

[20] A. Panwar and V. Bhatnagar, "Data Lake Architecture: A New Repository for Data Engineer," Int. J. Organ. Collect. Intell., vol. 10, no. 1, pp. 63–75, 2020.

[21] P. Russom, "Data lakes: purposes, practices, patterns and platforms," pp. 1–42, 2017.

[22] H. D. Challenges, S. Gupta, and V. Giri, Practical Enterprise Data Lake Insights. Apress, 2018.

[23] A. Gorelik, The enterprise big data lake: Delivering the promise of big data and data science. O'Reilly Media, 2019.

[24] C. Madera and A. Laurent, "The next information architecture evolution: The data lake wave," 8th Int. Conf. Manag. Digit. Ecosyst. MEDES 2016, pp. 174–180, 2016.

[25] Lei Zhang, "What are the differences between a database, data mart, data warehouse, a data lake and a cube?," quora, 2016. [Online]. Available: https://www.quora.com/. [Accessed: 15-Feb-2019].

[26] M. Y. Santos, B. Martinho, and C. Costa, "Modelling and implementing big data warehouses for decision support," J. Manag. Anal., vol. 4, no. 2, pp. 111–129, 2017.

[27] F. Ravat and Y. Zhao, "Data Lakes: Trends and Perspectives," in Springer Nature Switzerland AG 2019, 2019, vol. 2, no. Umr 5505, pp. 304–313.

[28] A. T. Spark, "Apache Spark," no. 1, pp. 39–53, 2018.

[29] M. M. Rathore, H. Son, A. Ahmad, A. Paul, and G. Jeon, "Real-Time Big Data Stream Processing Using GPU with Spark Over Hadoop Ecosystem," Int. J. Parallel Program., vol. 46, no. 3, pp. 630–646, 2018.

[30] T. Mahapatra and C. Prehofer, Graphical Flow-based Spark Programming, vol. 7, no. 1. Springer International Publishing, 2020.

[31] Microsoft Team, "Introduction to Delta Lake - Azure Databricks | Microsoft Docs," 2020. [Online]. Available: https://docs.microsoft.com/en-us/azure/databricks/delta/delta-intro. [Accessed: 26-May-2020].

[32] A. Wojciechowski, "ETL workflow reparation by means of case-based reasoning," Inf. Syst. Front., vol. 20, no. 1, pp. 21–43, 2018.

[33] F. Bentayeb, C. Favre, and O. Boussaid, "A user-driven data warehouse evolution approach for concurrent personalized analysis needs," Integr. Comput. Aided. Eng., vol. 15, no. 1, pp. 21–36, 2008.

[34] G. Thakur and A. Gosain, "DWEVOLVE: a requirement based framework for data warehouse evolution," ACM SIGSOFT Softw. Eng. Notes, vol. 36, no. 6, pp. 1–8, 2011.

[35] W. Ahmed, E. Zimányi, and R. Wrembel, "A logical model for multiversion data warehouses," in International Conference on Data Warehousing and Knowledge Discovery, 2014, pp. 23–34.

[36] M. Thenmozhi and K. Vivekanandan, "An ontological approach to handle multidimensional schema evolution for data warehouse," Int. J. Database Manag. Syst., vol. 6, no. 3, p. 33, 2014.

[37] C. Olston et al., "Nova: continuous pig/hadoop workflows," in Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, 2011, pp. 1081–1090.

[38] S. Nadal, O. Romero, A. Abelló, P. Vassiliadis, and S. Vansummeren, "An integration-oriented ontology to govern evolution in big data ecosystems," Inf. Syst., vol. 79, pp. 3–19, 2017.

[39] J. Song, C. Guo, Z. Wang, Y. Zhang, G. Yu, and J.-M. Pierson, "HaoLap: A Hadoop based OLAP system for big data," J. Syst. Softw., vol. 102, pp. 167–181, 2015.

[40] W. Chen, H. Wang, X. Zhang, and Q. Lin, "An optimized distributed OLAP system for big data," in 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), 2017, pp. 36–40.

[41] R. Tardio, A. Mate, and J. Trujillo, "An iterative methodology for big data management, analysis and visualization," in 2015 IEEE International Conference on Big Data (Big Data), 2015, pp. 545–550.

[42] S. Chen, "Cheetah: a high performance, custom data warehouse on top of MapReduce," Proc. VLDB Endow., vol. 3, no. 1–2, pp. 1459–1468, 2010.

[43] K. Akepanidtaworn, "Is 'Delta Lake' Replacing 'Data Lakes'? - Dev Dream Team - Medium," 2020. [Online]. Available: https://medium.com/dev-dream-team/is-delta-lake-replacing-data-lakes-49f1caddc646. [Accessed: 26-May-2020].

[44] Databricks Team, "Parquet files — Databricks Documentation," 04/03, 2020. [Online]. Available: https://docs.databricks.com/data/data-sources/read-parquet.html#parquet-files. [Accessed: 25-May-2020]

An Improved Ant Colony Optimization Algorithm: A Technique for Extending Wireless Sensor Networks Lifetime Utilization

Ademola P. Abidoye[1], Elisha O. Ochola[2], Ibidun C. Obagbuwa[3], Desmond W. Govender[4]

Department of Information Technology, Cape Peninsula University of Technology[1]
Cape Town, South Africa
School of Computing, University of South Africa, Pretoria, South Africa[2]
Department of Computer Science Education, University of KwaZulu-Natal, Pinetown, South Africa[3, 4]

Abstract—**Wireless sensor networks (WSNs) are one of the most essential technologies in the 21st century due to their increase in various application areas and can be deployed in areas where cable and power supply are difficult to use. However, sensor nodes that form these networks are energy-constrained because they are powered by non-rechargeable small batteries. Thus, it is imperative to design a routing protocol that is energy efficient and reliable to extend network lifetime utilization. In this article, we propose an improved ant colony optimization algorithm: a technique for extending wireless sensor networks lifetime utilization called AMACO. We present a new clustering method to avoid the overhead that is usually involved during the election of cluster heads in the previous approaches and energy holes within the network. Moreover, fog computing is integrated into the scheme due to its ability to optimize the limited power source of WSNs and to scale up to the requirements of the Internet of Things applications. All the data packets received by the fog nodes are transmitted to the cloud for further analysis and storage. An improved ant colony optimization (ACO) algorithm is used to construct optimal paths between the cluster heads and fog nodes for a reliable end-to-end data packets delivery. The simulation results show that the network lifetime in AMACO increased by 22.0%, 30.7%, and 32.0% in comparison with EBAR, IACO-MS, and RRDLA before the first node dies (FND) respectively. It increased by 15.2%, 18.4%, and 33.5% in comparison with EBAR, IACO-MS, and RRDLA before half nodes die (HND) respectively. Finally, it increased by 28.2%, 24.9%, and 58.9% in comparison with EBAR, IACO-MS, and RRDLA before the last node dies (LND) respectively.**

Keywords—*Sensor nodes; advanced nodes; fog nodes; data centre; cloud computing; ant colony optimization; visual sensor networks*

I. INTRODUCTION

Wireless sensor networks (WSNs) are one of the new technologies of the 21st century. A WSN is a group of spatially dispersed sensor nodes, which are interconnected through wireless communication. Sensor nodes jointly measure environmental conditions in an area of interest [1]. Recent improvements in wireless communication and camera sensors technologies have brought about the design of a new kind of sensor networks called "visual sensor networks VSNs". These networks can provide multi-perspective visual data which may be highly valuable for many types of monitoring applications. Although, VSNs developed from WSNs; they have extended the range of WSNs applications in health assistance, vehicular networks, Internet of Things, detection and prediction of natural calamities, smart cities, industrial, video surveillance for security systems such as pipeline monitoring, home automation, immersive entertainment oil, gas exploration, and real-time crop monitoring [2-3]. Conversely, in contrast to conventional WSNs, VSNs require more energy consumption, high bandwidth, high packet loss rate, and more processing capability to process sensed data and deliver them to a base station. Sensor nodes are usually deployed on large-scale in the area of interest and generate large volumes of data. However, they have many challenges in terms of data reliability and communication due to the limited capabilities of the individual nodes. It is expected that sensed data generated by the nodes are reliably delivered at the destination node for processing and storage. Sensor nodes are expected to remain functional for a longer period while providing a continuous stream of image data. These nodes are powered by small batteries and their life is dependent on the amount of initial power loaded onto the batteries and to the way they are dissipated during network operation. A sensor node dissipates energy during data acquisition, formatting, pre-processing, and data forwarding [4]. Thus, data transmission is the main energy consumption of a sensor node [5]. Therefore, there is a need to minimize sensor nodes' energy consumption to extend the network lifetime.

A. The Significance of the Routing Protocol for Wireless Sensor Networks

The nodes relay their sensed data towards the base station through their neighbouring nodes [6]. Data routing in WSNs is very important unlike transitional networks based on the following features [7-8]:

- Power to the WSNs is usually provided by small batteries.

- Traffic pattern in WSNs is many-to-one data transmission.

- Scalability in WSNs to the large scale of distribution.

- The ability of the WSNs to quickly adjust itself to change in the topology is considered its responsiveness.

- Each sensor node transmits data to its neighbouring node.

Designing an efficient, trustworthy routing protocol, and provisioning quality of service (QoS) for data routing in WSNs is a challenging task due to resource-constraints of sensor nodes, dynamic nature of the networks, and mode of the transmission medium. Several approaches have been proposed in the literature based on data routing to meet the QoS requirements [9-10]. Most of the proposed routing protocols for WSNs are based on the single-path routing algorithm, where each source node transmits its sensed data to the base station through a single path. Although, single-path routing approach is simple because it can be implemented with minimum computational complexity and scalability but there is no adequate consideration of traffic load-balance or reliability along the selected paths [11-12]. It is simple because paths between source nodes and the base station can be constructed in a short time. Similarly, it is scalable because as the number of sensor nodes becomes large, the method and complexity to create paths between the source nodes and the base station do not change [13].

Conversely, a multi-path routing protocol transmits copies of the sensed data to the base station through different paths. This addresses the throughput, load balancing, reliability, and security challenges of the single-path routing protocol [11, 14]. In the event of the main path being unavailable due to the low energy of individual sensor nodes or congestion, other less congested paths of the network will be used for data transmission. This increases the throughput of a network by transmitting sensed data in parallel through many paths and delivering the entire data at the destination with the expectation of achieving high throughput, reliability, and low data packets loss rate [10]. Many multipath routing protocols have been developed in the literature to ensure load balancing, congestion avoidance, fault-tolerance, and QoS [11, 15-16].

Braided multipath routing techniques have been proposed to relax the requirement for node disjointedness with the expectation of addressing the energy issues of node disjoint paths [17]. This scheme creates a backup node for every sensor node on the primary path. If a node in the main route fails, the backup node is used to connect other nodes. However, this routing protocol still builds around reliability requirements only, disregarding the throughput maximization and energy efficiency. Therefore, to achieve reliable wireless communications in WSNs, there is a need to have a reliable routing protocol.

Continuous developments in wireless communication and the Internet of Things (IoT) technologies have enabled WSNs to transmit raw sensed data to the cloud for processing, analysis, and storage. Cloud computing is considered as a promising solution to provide applications with elastic resources and deliver services to end-users at a low cost. However, cloud computing has its drawbacks and cannot solve all WSNs routing challenges [18].

Moreover, if sensor nodes are to transmit directly to the cloud, the process will need a high bandwidth network and quickly dissipates sensor nodes' energy [19]. Some applications such as real-time streaming, health care data, real-time gaming, and augmented reality are too latency-sensitive to be deployed directly to the cloud. Similarly, in large scale sensor networks, sensor nodes have to constantly sense and transmit a huge amount of data. The processing of such data needs extra efforts and time, which dissipates more energy from the sensor nodes. In addition, sensor nodes are resource-constrained, they can neither perform complex analytical computations nor machine learning tasks. Therefore, instead of sensor nodes to transmit directly to the cloud, a new technology called fog computing can be deployed at the edge of the network to provide cloud services closer in a sense [20-21]. The main aim of integrating fog into the WSNs is to improve their reliability and minimize the redundancies associated with data transmission to the cloud for processing.

Fog computing is composed of networking devices such as gateways, routers, proxy servers, set-top boxes. These devices store frequently used the information to provide the services to edge users [22]. They have higher processing capability and storage than typical sensor nodes. The devices can be placed between WSNs and cloud computing to receive, process, and temporarily store the sensed data. This technique will greatly conserve sensor nodes' energy consumption because the nodes will only need to transmit through a short distance. It can efficiently reduce the amount of bandwidth that is required due to its ability to minimize the needed back and forth communication with the cloud and the various sensors [23]. The term "fog computing" was developed by Cisco to overcome limitations in cloud computing [24-25]. It consists of fog nodes (FNs) which provide resources at the edge of the network. They play an important role in the overall working of fog computing as they aggregate the data from source nodes for processing. They act as decentralized local access, thus reducing the dependency on the cloud platform for analyzing the sensed data. This new technology offers several benefits to end users including efficient network bandwidth usage, data security, fewer bottlenecks, the solving of high latency on the network, increased reliability of transmitted sensed data, and a higher speed of analysis [26]. Fogging can effectively string everything together without reducing the overall performance of the processes or devices. Leveraging the advantages of this new technology, we integrate fog computing into the proposed scheme to address some of the constraints of WSNs.

Sensor nodes are resource-constrained in terms of power and communication bandwidth [27]. As a consequence, reducing the energy dissipation of an individual sensor node is a critical issue for WSNs. Ant colony optimization (ACO) algorithm has been applied in WSNs to find optimal paths to save energy consumption during transmission. However, the algorithm is prone to converge at local optima. Therefore, we propose an improved ACO algorithm that can be applied to construct the optimal paths between layer 1 and layer 2 of the proposed scheme architecture in Fig. 1 for efficient and reliable data transfer.
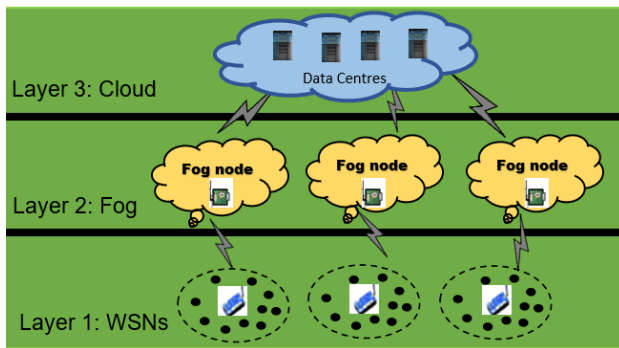
Fig. 1.   Proposed Scheme Architecture.

ACO is developed by Dorigo [28]. The idea is based on the natural behaviour of real ants in their colonies and to emulate the cooperative behaviour of an ant colony, which discovers the shortest path to a food source. The behaviour of real ants and sensor nodes share similar attributes. Thus, the algorithm has been applied to solve many routing problems in WSNs [29-31].

B. Contributions

This research investigates existing routing protocols for WSNs in terms of their reliable performance and proposes a new routing protocol. In this paper, we present an energy-efficient routing scheme combined with clustering and fog nodes technology. We develop a model to partition the network into different clusters. An algorithm is designed to distribute evenly advanced nodes within the network and assign only one advanced node (cluster head) to each cluster. We developed a model based on the distance and the residual energy of a node to associate it to a cluster head, this enables the nodes to communicate with their associated cluster head using a single or multi-hop communication following the optimal energy consumption. An improved ACO algorithm is designed to construct optimal and efficient paths between the cluster heads and the fog nodes to minimize the total number of broadcast transmission between layer 1 and layer 2 of the proposed architecture. Thus extending the life cycle of the WSNs, and effectively improving network congestion and minimizing the average delay. Finally, through numerous performance comparisons, we show the effectiveness of our approach for the proposed routing in WSNs.

The rest of the paper is organized as follows: Section II presents some related work. Section III presents the assumptions and system models. The proposed scheme is presented in Section IV. Section V presents the analysis and simulation results for different scenarios. Lastly, the conclusion and our future research directions are presented in Section VI.

II. Related Work

In recent years, many routing techniques have been proposed to minimize energy consumption in WSNs. Multipath routing techniques are one of the most prominent and efficient routing schemes developed for WSNs. By leveraging the many routing paths, multipath routing can improve the reliability for end-to-end data transmission, minimize energy consumption and transmission delay. The following are some of the multipath routing protocols that have been proposed for WSNs.

Radi et al. [32] proposed an interference-minimized multipath routing protocol for WSNs. The scheme aims to discover a sufficient number of minimum interfering paths with high QoS between source nodes and base stations. The approach consists of three phases: i) initialization phase; ii) path establishment phase; iii) data packet transmission phase. These phases control the traffic rate of each constructed path. Performance evaluations of the proposed method show improvements in terms of packet reception ratio, packet delivery latency, and energy consumption.

However, the authors did not explain in detail how the clusters are formed. They only use an assumption for the formation of the clusters.

Deepa and Suguna [33] proposed Optimized QoS-based Clustering with Multipath Routing Protocol (OQoS-CMRP) for WSNs. The authors applied improved particle swarm optimization (PSO)-based clustering algorithm to form clusters within the network. Also, they used the Single Sink-All Destination algorithm to determine near-optimal multi-hop routing path from the sink to sensor nodes in order to choose the next-hop neighbour nodes and Round-robin Paths Selection algorithm is used for sending sensed data to the sink. Performance evaluations of the proposed scheme show that it performs better in terms of energy conservation, transmission delay, and communication overhead than selected related algorithms.

But, in this approach, a lot of energy get consumed because re-clustering and re-routing happen every time, when residual energy of a node goes below a particular threshold.

Sharma and Jena [34] suggested a cluster-based multipath routing protocol for WSNs (CMRP), which uses both the multipath and clustering methods to decrease the energy consumption of the sensor nodes and increase data packets reliability. The idea centres on the reduction of individual sensor nodes load by assigning more responsibility to the destination node. The performance evaluations show that CMRP is more energy-efficient and reliable compared to existing protocols.

However, the clustering scheme in the approach is not scalable enough to facilitate cluster maintenance, therefore it can result in domino effects.

Gupta and Jha [35] proposed integrated clustering and routing protocol for WSNs using an improved cuckoo and harmony search based metaheuristic techniques. The approach uses a novel multi-objective function for uniform distribution of cluster head nodes and a modified harmony search based routing protocol is used for routing of a data packet from cluster heads to the sink node. The performance of the proposed protocol is evaluated using metrics such as average energy consumption, number of dead nodes, number of alive nodes, and network lifetime. The evaluation results show significant improvement over the state-of-art protocols.

However, the authors failed to explain how the clusters are formed and cluster heads selected. There is a high possibility

that cluster heads are selected from one side of the network and thus results in energy holes.

An energy-efficient algorithm for reliable routing of WSNs based on distributed learning automaton (RRDLA) algorithm is proposed in [11]. The author models the scheme as a multi-constrained optimal path problem. The scheme exploits the features of distributed learning automaton (DLA) to determine the smallest number of sensor nodes to maintain the desired QoS requirements. The scheme considered many QoS routing metrics which include end-to-end reliability, throughput, and delay to select an optimal path in the network. The evaluation results show that the proposed scheme performs better than related proposed algorithms in terms of energy efficiency, throughput, and end-to-end delay.

An improved ant colony optimization-based approach with a mobile sink (IACO-MS) is proposed to solve the energy hole problem in WSNs [36]. The approach divided the network into different clusters and each cluster has only one cluster head (CH). The conventional ACO algorithms construct paths between the CHs while the mobile sink node finds an optimal mobility path to communicate with CHs under the improved ACO algorithm. The authors claimed that the sink mobility approach can be used to solve many routing problems in WSNs such as hot spots problem, reduction in energy consumption of the whole network, and improve the network in terms of transmission latency and network throughput compared to conventional routing algorithms.

However, the approach is compared with only one approach and it may not be efficient if it is compared with two or three existing approaches.

Krishnan et al. [37] stated that existing clustering with static sink approaches for WSNs creates an energy hole problem and untimely death of sensor nodes leads to data packets loss. The authors proposed a novel dynamic clustering approach with PSO-based mobile data collectors for information gathering to solve these problems. The performance evaluations show that the proposed scheme performs better in terms of reducing energy consumption for sensor nodes, improving throughput, and extending the network lifetime compared with the existing schemes.

However, this approach is only implemented on a small number of sensor nodes. It may not be energy-efficient if the network size is large.

Arjunan and Sujatha [38] proposed a lifetime maximization of WSNs using fuzzy-based unequal clustering and ACO based routing hybrid protocol. The protocol consists of three phases: CH phase, inter-cluster routing phase, and cluster maintenance phase, and it aims to eliminate hot spot problems and prolong the network lifetime. The Fuzzy logic part of the approach chooses CHs efficiently and splits the network into different clusters based on the distance to its neighbouring nodes, residual energy, node degree, node centrality, and distance to the base station (BS). Thereafter, it uses the ACO algorithm based on a hybrid routing protocol to construct optimal paths between the CHs and BS. The authors employed a threshold concept to transmit or intimate sudden changes in the environment in addition to periodic data

transmission. Finally, the cluster maintenance phase is used to uniformly distribute the traffic load. The performance evaluation results show that the scheme solves hot spot problems, efficiently balances the energy consumption among all sensor nodes, and prolongs network lifetime.

However, this approach requires a significant amount of overhead energy.

Li et al. [39] proposed an energy-efficient load balancing ant-based routing algorithm (EBAR) for WSNs. The approach adopted the following methods - a pseudo-random path discovery algorithm and an improved pheromone trail algorithm. The pseudo-random path discovery algorithm based on a greedy algorithm is used to optimize the route establishment and the pheromone trail is used to balance the energy consumption of the sensor nodes. The proposed scheme exploits an energy-based opportunistic broadcast scheme to minimize the energy consumption of sensor nodes caused by the control overhead. The authors used metrics such as energy efficiency, energy consumption, and predicted network lifetime to evaluate the proposed scheme. The performance results show that EBAR performs better compared to selected related work.

However, the approach is based on static network and cannot be applied in a scenario with multiple base stations.

III. ASSUMPTIONS AND SYSTEM MODELS

In this section, we present the assumptions and system models for the proposed scheme in detail.

A. Assumptions

- The initial energy of all sensor nodes is equal.

- Each sensor node in the network is both a transmitting node and a source node.

- Data centres in cloud computing are final destinations for sensed data.

- The network topology is static after the nodes have been deployed in the network area.

- Data centres power sources are unlimited.

- The coordinates of the sensor nodes are known.

- The medium access control (MAC) layer provides the facility to determine link quality such as the packet reception ratio (PRR).

- Every node knows the PRR of its neighbouring nodes.

B. Network Model

In this paper, N sensor nodes are randomly distributed in a M x M network area as shown in Fig. 2. All sensor nodes deployed in a target area can be represented as a set of nodes $\{n_1, \ldots, n_v\}$, where n_i denotes node i for all $i = 1, 2, \ldots, v$. Every node has a sensing range (R_S) that allows it to monitor the events in the area of deployment and a transmission range (r) which allows it to communicate with other nodes within the network such that the nodes $n_i, n_j \in N$, and $n_i \neq n_j$.
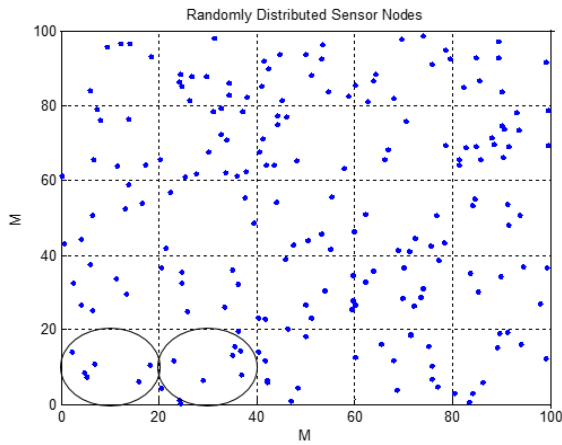
Fig. 2.   Network Model.

MAC sub-layer is part of the Data Link layer in WSN's protocol stack. The energy consumption of sensor nodes is greatly affected by the MAC protocol which controls the node radio functionalities [40]. The reliability of links quality is obtained using the PRR.

We model the routing problem as an undirected graph $G = (N, L)$, where N is a set of sensor nodes and L denotes the set of two-way edges that link two sensor nodes such that each link is contained in G for all $w_i, w_j \in L$.

We assume that nodes n_i and n_j can communicate with each other if and only if the distance between them is less than or equal to the sensor node transmission range. The Euclidean distance $d_{n_i,n_j} = D$ between two nodes is expressed as follows.

$$D = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{1}$$

where (x_i, y_i) is the coordinate of node i such that $n_i = (x_i, y_i)$, $n_j = (x_j, y_j)$.

Node n_j is a neighbour of node n_i if it can be expressed as follows

$$n_j = \{ n_j \text{ such that } (n_i, n_j) \in N, D \leq r\} \tag{2}$$

C. Energy Model

Each sensor node in a WSN has a radio communication subsystem consisting of transmitter/receiver electronics, antennae, and an amplifier [40]. We adopt the radio energy model presented in [41]. The energy of the transmitter E_{Tx} can be determined by these equations based on their transmission distance with a neighbouring node or a cluster head. We assumed two models as shown in Equation (3) for data packets transmission. A free space model is used if the distance between two communicating nodes is less than a threshold distance value d_0, otherwise, a multi-path fading model is used to compute the energy consumption of the node. The models are presented as follows.

$$E_{Tx}(l, d) = \begin{cases} q * E_{elec} + q * \varepsilon_{fs} * d^2, & \text{if d} < \text{d}_0 \\ q * E_{elec} + q * \epsilon_{amp} * d^4, & \text{if d} \geq \text{d}_0 \end{cases} \tag{3}$$

and the energy consumed by a node to receive a $q - bit$ of sensed data from a node is expressed as follows.

$$E_{Rx}(q) = q * E_{elec} \tag{4}$$

where E_{elec} is the electronic energy that depends on factors such as the spreading of the signal, modulation, and digital coding. Friss free space is denoted as ε_{fs} and ϵ_{amp} denotes multi-path fading which depends on the transmitter amplifier model, d is the distance between the nodes.

The threshold distance value d_0 is determined as follows.

$$d_0 = \left(\frac{\varepsilon_{fs}}{\epsilon_{amp}}\right)^{\frac{1}{2}} \tag{5}$$

D. Sensor Node Coverage Area

A sensor node can only sense the environment and perceive the event if and only if the event is within its transmission range as shown in Fig. 3. A target node is said to be covered if it is contained in the sensing area (S_A) of a sensor node. The network area is partitioned into 25 grids and S_A is an area of a circle $S_A = \pi r^2$. The probability of target (P_B) detection in a grid of the network area is expressed as $P_B = \frac{S_A}{M^2}$. Thus, the probability (P_T) for detecting a target by at least a sensor node in the network area is expressed as follows.

$$P_T = 1 - (1 - P_B)^N \tag{6}$$

where R denotes an average transmission range of a sensor node, R_{max} denotes maximum transmission range of a sensor node, and dr is a small increment in r.

The probability that a sensor node n_i senses an event within its transmission range can be expressed as:

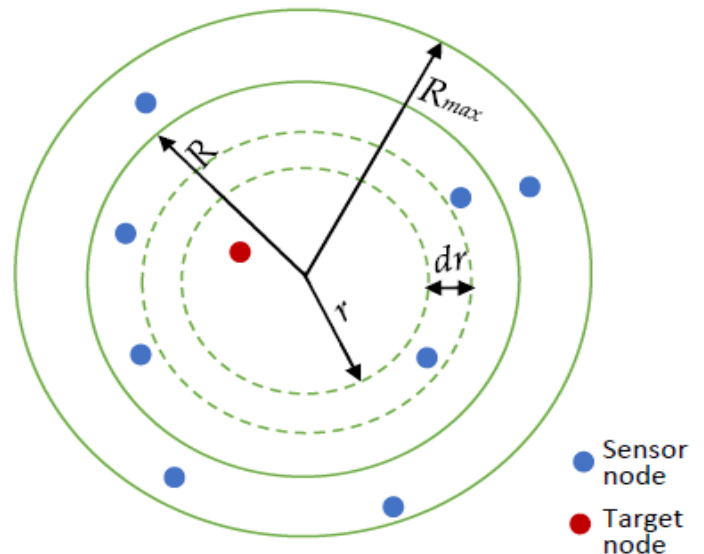$$P_B(r) = P\left(X_\sigma + \gamma + 10\Phi log_{10}\left(\frac{r}{R}\right)\right) \tag{7}$$

Fig. 3.   Sensing Area for a Sensor Node Coverage.

where X_σ connotes a zero-mean, Gaussian random variable (measured in decibels) with standard deviation σ (in decibels), γ is a random variable that connotes Rayleigh fading to model the multiple path effects in propagation path and Φ denotes path loss exponent [42]. Thus, $P_B(r)$ can be expressed as follows.

$$P_B(r) =$$

$$\int_0^\infty \int_{10\Phi log_{10}\left(\frac{r}{R}\right)-y}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\left(x^2/2\sigma^2\right)\right]} \frac{y}{\nabla^2} e^{\left[-\left(y^2/2\nabla^2\right)\right]} dxdy \quad (8)$$

$$= \int_0^\infty C\left(\frac{10\Phi log_{10}\left(\frac{r}{R}\right)-y}{\sigma}\right) \frac{y}{\nabla^2} e^{\left[-\left(y^2/2\nabla^2\right)\right]} dy \quad (9)$$

Applying the numerical integration of the Gauss–Laguerre $\int_0^\infty e^{-x} f(x) dx$, the detection probability can be expressed as:

$$P_B(r) = C\left(\frac{10\Phi log_{10}\left(\frac{r}{R}\right)-\nabla\sqrt{2}}{\sigma}\right) \quad (10)$$

Thus, the probability that a sensor node detects a target node placed at a particular point of the network area M^2 denoted as A can be expressed as:

$$P_B(r) = \frac{2}{A} \int_{r=0}^{Rmax} P_B(r) * \pi r dr \quad (11)$$

Substituting (10) into (11), the probability that a sensor node sensed a target in the specified area of the network can be expressed as follows.

$$P_B(r) = \frac{2}{A} \int_{r=0}^{Rmax} C\left(\frac{10\Phi log_{10}\left(\frac{r}{R}\right)-\nabla\sqrt{2}}{\sigma}\right) * \pi r dr \quad (12)$$

E. Coverage Rate Computation

A target node B at point $g(x_B, y_B)$ is within the transmission radius of the node n_i, if its sensing rate $r_t(n_i, B)$ is within the coverage of n_i as:

$$r_t(n_i, B) = \begin{cases} 1, D(n_i, B) \leq r \\ 0, D(n_i, B) \geq r \end{cases} \quad (13)$$

where $D(n_i, B)$ denotes the distance between the target node and node n_i. The probability that the target node B can be covered by nodes in N is presented as follows.

$$r_t(N, B) = \prod_{n_i \in N} [1 - r_t(n_i, B)] \quad (14)$$

Thus, the coverage rate (C_t) of the target area can be computed as follows.

$$C_t = \frac{\sum_{B \in U} r_t(N,B)}{U} \quad (15)$$

IV. Our Proposed AMACO Scheme

The architecture of the proposed scheme is composed of three layers: Layer 1- WSNs, layer 2- fog nodes, and layer 3- remote cloud storage as shown in Fig. 1.

Layer 1-WSNs: This layer consists of various types of sensor nodes connected through wireless technologies and are randomly distributed in the network area. The layer consists of two types of nodes: normal sensor nodes (N_S) and advanced

nodes (N_A). These nodes are given specific responsibility to perform within the network. Each sensor node monitors its area, obtain sensed data, and transmits it to a neighbouring node so as to achieve a global detection objective.

Advanced nodes are included in the architecture to avoid overhead during cluster formation and support various types of nodes for improving routing capability. The advanced nodes are responsible to receive, aggregate, and retransmit sensed data received from the normal sensor nodes. They have more processing power and energy source than normal sensor nodes.

Layer 2–Fog computing: Layer 2 consists of fog nodes with high computing, storage, and network connectivity. They receive data from advanced nodes, aggregate, analyse the data so that only the essential data gets forwarded further to the cloud, and decreases the bandwidth used [43]. Besides, sensitive data can be processed at these fog nodes.

Layer 3–Cloud computing: This is the uppermost part of our architecture. Pre-processed data is transmitted from fog nodes to the cloud for extensive processing and analysis using cloud computing platforms and storage is done in data centres. The cloud can process and store a large amount of data being transmitted from the lower layers. Many security issues such as data integrity, data authentication, privacy are addressed at this layer [44].

A. Clustering

Wireless sensors networks are usually composed of a large number of low-cost sensor nodes connected through a wireless network that sense data to be relayed to the data centres through multi-hop wireless transmission. Many methods have been proposed in the literature to reduce the traffic into the network. Clustering algorithms have been used to minimize communication distance among the nodes to conserve their limited energy. It involves dividing the network into different groups (clusters) and select a node as the group leader (cluster head) for each cluster in the network.

Our main aim for clustering is to minimize each sensor node transmission distance to save its energy. In addition, it eliminates energy holes problems within the network.

B. Setup Phase

During the setup phase, sensor nodes are randomly distributed into the network area as shown in Fig. 2. Each node exchanges its information with neighbouring nodes. Unlike in previous approaches, in which normal nodes are selected as CHs using various algorithms. In our scheme, advanced nodes (NA) are added to the network and they automatically become cluster heads (CHs) due to their higher specifications in terms of energy source, transmission range, and computational power than normal nodes. Algorithm 1 is used to evenly distribute the NA and the desired number A_N represented as K of NA in the network is determined as follows.

$$A_N = \left\lfloor \sqrt{N_{per} * N} \right\rfloor \quad (16)$$

where N_{per} is the percentage of N_A and N is the number of sensor nodes.

Algorithm 1. Algorithm to distribute advanced nodes evenly within the network

Begin
1: Given the set of $N_A = \{N_A^1, N_A^2, \ldots., N_A^K\}$
   2: Initialization: $x(k) = y(k) = 0$;
3: Input: use Equation (16) to compute desired number of N_A;
   4: **for** $k = 1:K$ // K is A_N
5: $x(k) = \text{rand}(1,1) * A$;
6: $y(k) = \text{rand}(1,1) * A$; // $A = Lenght$ and $Width$ of the network
7: Voronoi $((x(k), y(k))$
8: **end for**
9: Determine the center of the network
10: **for** $R_{ow} = 1: L_y$
11: **for** $C_{ol} = 1: L_x$
10: $O_x = (C_{ol} - 1) * 0.5$;
12: $O_y = (R_{ow} - 1) * 0.5$;
13: $C_{tr}(R_{ow}, C_{ol}) = (O_x, O_y)$;
14: **end for**
15: **end for**
16: $return \ C_{tr}(R_{ow}, C_{ol})$;
End

C. Intra-Cluster Communication

Now the advanced nodes (CHs) have been distributed evenly within the network, next we divide the network into different clusters. We assume that the number of advanced nodes is equal to the number of clusters. Long-distance transmission usually dissipates the energy of a sensor node and shortens the network lifetime. Thus, a node determines the cluster it is to join by selecting a CH that is within its radio transmission range in which the received signal strength (RSS) is strongest. A Boolean variable X_{ij} is used to denote whether a node i is belongs to N_A^j to form a cluster as follows.

$$X_{ij} = \begin{cases} 1, \text{if node } i \text{ close to } N_A^j \\ \forall i, j: 1 \leq i \leq N, 1 \leq j \leq K \\ 0, \text{Otherwise} \end{cases} \quad (17)$$

where N denotes the number of nodes and K is the number of advanced nodes. In the proposed scheme, energy consumption for a sensor node to transmit to its associated advanced node N_A^j is calculated as follows.

$$E_1(n_i, N_A^j) =$$
$$\begin{cases} q * E_{elec} + q * \varepsilon_{fs} * d(n_i, N_A^j)^2, \text{if } d(n_i, N_A^j) < d_0 \\ q * E_{elec} + q * \epsilon_{amp} * d(n_i, N_A^j)^4, \text{if } d(n_i, N_A^j) \geq d_0 \end{cases} \quad (18)$$

Moreover, if a node is far from its leader, it will transmit to the CH through a neighbouring node n_j. The energy consumption for data transmission can be expressed as follows.

$$E_2(n_i, n_j, N_A^j) = E_{Tx}\left(q, d(n_i, n_j)\right) + E_{Rx}(q)$$
$$+ E_{Tx}\left(q, d(n_j, N_A^j)\right)$$
$$= 3qE_{elec} + \varepsilon_{fs} \, d(n_i, n_j)^2 + \varepsilon_{fs} \, d(n_j,, N_A^j)^2 \quad (19)$$

D. Average Packet Delay

The average packet delay (D_{pctk}) is the average interval between data packets transmitted at the source node n_i and delivered at the destination node n_j or N_A^j for a given period of data transmission and expressed as follows.

$$D_{pctk} = \frac{\sum_{j=1}^{K} \sum_{i=1}^{N} n_i \left(t_{ij}^{depat.} - t_{ij}^{arrial}\right)}{\sum_{j=1}^{K} q_j} \quad (20)$$

where t_{ij}^{arrial} is the time it takes data packets to arrive at the next node media access control (MAC) layer and $t_{ij}^{depat.}$ is the time it takes the data packets to be delivered at the next node MAC sublayer.

E. Data Packets End-to-End Reliability

In the proposed scheme, medium access control (MAC) layer is used to estimate the routing paths quality, that is, the packet reception ratio (PRR), every node knows the PRR of its neighbouring nodes. In a multi-path transmission, the probability of successful data packets delivery (*DPD*) to the CH through m-hop of a path *o* between two neighbouring nodes can be expressed as follows.

$$PDR = \coprod_{\forall (n_i, n_j) \in o} P_r\left((n_i, n_j)\right) \quad (21)$$

where *PDR* denotes the data packets delivery ratio. Based on the value of PDR and the number of data packets transmitted H_{sent}, we can compute the number of data packets successfully delivered $H_{success}$ to the CH as follows.

$$H_{success} = H_{sent} * PDR \quad (22)$$

F. Improved ACO Algorithm

In this section, we use an improved ACO algorithm to construct an efficient and reliable path for inter-cluster communication between the advanced nodes in layer 1 and fog nodes in layer 2 of the proposed architecture.

Both advanced nodes and fog communicate through wireless medium and many routing paths exist between these nodes but not all are efficient and reliable for data packet transmission. Therefore, it is necessary to construct efficient and reliable paths through which the advanced nodes would transmit their aggregated data to the fog nodes for processing and temporary storage. Algorithm 2 is developed to construct optimal and reliable paths for data transmission between the advanced node N_A and fog node N_F. ACO algorithm has been used in WSNs to construct shortest paths, thus save energy consumption in the network. However, this algorithm converges slowly and leads to local optima.

ACO is a branch of optimization modelled algorithms based on the behaviour of real ants in a colony and is a subclass of computational intelligence (IC) paradigms that aid in determining optimal solutions to optimization problems [30].

The idea comes from the foraging behaviour of ants in their colony when searching for food. They first randomly explore the surrounding area and deposit a chemical substance called pheromone on their paths as they search for food, forming pheromone trails. The pheromone trails can be sensed by other ants in the colony and enhance the pheromone deposited on the paths. They tend to choose a path marked by strong concentrations of pheromone. This traditional ACO algorithm does not distinguish between the different types of data packets transmitted over the network. All data packets are sent in the same way to a destination; this algorithm is mainly designed to find the shortest path between the source nodes and destination node. On the other hand, WSNs in the IoT environment has to support multimedia data transmission, requiring a different level of quality of service (QoS). This work provides three different QoSs.

Guaranteed service (Q_{s_1}): Q_{s_1} provides safe end-to-end delay guarantees. It guarantees that data packets will reach the destination node at the right time with zero data packets loss.

Control load-balance service (Q_{s_2}): Q_{s_2} service is applied where there is a possibility that delay will happen and when there is congestion in the network, the Q_{s_2} can provide a service just as if there was no congestion in the network.

Best-effort service (Q_{s_3}): This an Internet delivery service where the network does not provide any guarantees on the time delay limit the data packets will be delivered in the network.

If a link $l_u \in E$ denotes a single hop between a source (an advanced) node and destination (fog) node, then the multi-hop, T_{path}, is the sum of the hops the nodes and expressed as follows.

$$T_{path} = \sum_{u=1}^{w}(l_u) \tag{23}$$

Equation (23) shows that the fewer T_{path}, the nearer it is to the fog node. $R(i,j)$ is used to represent the relationship between a node (i) and receiver node (j) as shown in equation (24).

$$R(i,j) = \begin{cases} 1 \; if \; l_u \in T_f(i) \\ -1 \; if \; l_u \in T_r(i) \\ 0, others \end{cases} \tag{24}$$

where $T_f(i)$ and $T_r(i)$ denote forward link and reverse link respectively of the node i.

It is assumed that the success transmission rates of all the links between the advanced nodes and fog nodes are all 100%. If the data packets transmission success rate is less than 10%, then the advanced node has low communication performance. However, if the data packets transmission rate is more than 90%, then it has high communication performance.

G. Local Pheromone update

Pheromone is very important in the performance of the ACO algorithm. It is used for updating both local and global trails in traditional ACO. It makes the paths visited by ants to depend on the objective value. Modification made to the traditional ACO algorithm is presented as follows. The choice

of constructing an optimal path between the source node and the receiver node is made based on the probability decision rule in equation (25).

$$c = \begin{cases} \arg \max_{j=1,2,..,m} \left\{ \left[\lambda_{kj}^{Q_{s\_a}}\right]^\alpha * \left[\eta_{kj}^{Q_{s\_a}}\right]^\beta * R(k,j), \rho \le \rho_o \\ C, \rho > \rho_o \end{cases} \tag{25}$$

$$C = \begin{cases} \dfrac{\left[\lambda_{kj}^{Q_{s\_a}}\right]^\alpha * \left[\eta_{kj}^{Q_{s\_a}}\right]^\beta * R(k,j)}{\sum_{j \in T_{path,}} \left[\lambda_{kj}^{Q_{s\_a}}\right]^\alpha * \left[\eta_{kj}^{Q_{s\_a}}\right]^\beta}, j \in \text{allowed}_u \\ 0, j \notin \text{allowed}_u \end{cases}$$

where $\lambda_{kj}^{Q_{s\_a}}$ is the amount of pheromone deposited on the path, $\eta_{kj}^{Q_{s\_a}}$ represents the local heuristic value of the path between the sender node and the receiver node, a denotes the QoS type, α and β are control parameters used to regulate the concentration of the pheromone trail and the heuristic value respectively. The ρ is a random variable range between 0 and 1 (i.e [0, 1]) and ρ_o ($0 \le q_o \le 1$) is a given parameter. $j \in \text{allowed}_u$ for all $u = 1,2,3, \dots, w$ are paths that can be selected by node k in the next step.

H. Global Pheromone updating Rule

Once the communication paths between nodes have been constructed by the search ants, equation (22) is used to choose an optimal path.

$$Z_O^{Q_{s\_a}} = max \left\{ \left(Z_{path}^{Q_{s\_a}}\right)_u \right\} \tag{26}$$

where $\left(Z_{path}^{Q_{s\_a}}\right)_u$ is the path utility value. $\Delta\lambda_{kj}^{Q_{s\_a}}$ is the pheromone increment on the path between the advanced node k and fog node j. Pheromone is updated between the nodes as follows.

$$\Delta\lambda_{kj}^{Q_{s\_a}} = \begin{cases} \dfrac{\varpi * Z_{path}^{Q_{s\_a}}}{T_{path}}, j \in \text{allowed}_u \\ 0, \; j \notin \text{allowed}_u \end{cases} \tag{27}$$

where ϖ is an adjustment coefficient, T_{path} denotes path length. Equation (24) is a pheromone update rule for the forward ants used to create the paths between the advanced nodes and the fog nodes.

I. Pheromone Deterioration

Advanced nodes will use the optimal paths constructed to transmit their data packets to the fog nodes. However, continuous data packets transmission along the optimal paths will lead to (a) congestion along the paths (b) inability to discover other paths. These two points cannot be overlooked in a dynamic network because (i) an optimal path may become non-optimal if it is congested; (ii) It may also lead to loss of data packets due to network failure or energy holes problem. To overcome these challenges, pheromone control is used as a measure to reduce the impact of earlier experience and encourages the search for alternative paths that were non-optimal through evaporation. For a constant ρ of pheromone deterioration (decay), the pheromone concentration on the path usually varies with time t and expressed as follows:

$$\lambda_{kj}^{Q_{s\_a}}(t) = \lambda_0^{Q_{s\_a}} e^{-\rho t} \tag{28}$$

where $\lambda_0^{Q_{s\_a}}$ denotes the initial pheromone concentration. If ρt is much less than 1 (i.e $\rho t \ll 1$) then.

$$\lambda_{kj}^{Q_{s\_a}}(t) \approx \lambda_0^{Q_{s\_a}}(1 - \rho t)$$

If the time increment is 1, then the evaporation can be approximated as follows.

$$\lambda_{kj}^{Q_{s\_a}(t+1)} \leftarrow (1 - \rho)\lambda_{kj}^{Q_{s\_a}t} \tag{29}$$

Thus, the general pheromone update formula can be expressed as:

$$\lambda_{kj}^{Q_{s\_a}(t+1)} = (1 - \rho)\lambda_{kj}^{Q_{s\_a}t} + \Delta\lambda_{kj}^{Q_{s\_a}t} \tag{30}$$

where ρ denotes the rate of pheromone evaporation; its value ranges between 0 and 1. The increment $\lambda_{kj}^{Q_{s\_a}t}$ is the amount of pheromone deposited at time t along with path k and j.

Algorithm 2. Construction of optimal path based on improved ACO algorithm

---

Begin
Input: Advanced nodes (N_A)
Output: Construct an optimal path between N_A and N_F

1:    Initialize $\lambda_{kj}^{Q_{s\_a}}$ and $a \leftarrow 1$
2:    **While** termination conditions are not met **do**
3:     $a \leftarrow a + 1$
4:    **for** $k = 1$ to K
5:     k is positioned at F_L
6:     $F(k) \leftarrow F_L$
8:     $R^k \leftarrow \emptyset$; $\varphi^k \leftarrow \emptyset$
9:    **While** $F(k) \neq F_L$ **do**
10:    **if** $C(F(k)) - \varphi^k \neq \emptyset$ **then**
11:     Choose $(F(j)$ from $C(F(k)) - \varphi^k$ to move based on the probabilistic transition rule in equation (25)
12:     $R^k \leftarrow R^k \cup \{F(i)\}$; $\varphi^k \leftarrow \varphi^k \cup \{F(i)\}$; $k \leftarrow j$
13:    **else**
14:    Back to the previous-hop of $F(k)$
15:    **end if**
16:    **end while**
17:    Compute the paths between the nodes using equation (24)
18:    Determine the value of $Z_O^{Q_{s\_a}}$ using equation (26)
19:    **end for**
20:    Update the number of pheromone values using equation (27)
21:    Compare the values of solutions obtained
22:    **end while**
23:    Return the optimal solution
24:    Choose the best solution as the output
25:    **End**

V. PERFORMANCE EVALUATION

To evaluate the performance of the proposed scheme, the AMACO algorithm is compared with three related algorithms namely RRDLA, IACO-MS, and EBAR. The following metrics are used to measure the performance of these algorithms: network lifetime, the sum of energy consumption, the average number of hops, packet delivery ratio, and success rate. Network Simulator- 2 (NS-2) is used to perform the experiment and 200 sensor nodes are randomly distributed in a $100 * 100$ m^2 network area. We set the initial energy of each sensor node to 0.5J and other relevant parameters are presented in Table I.

A. Network Lifetime

The network lifetime t_N can be defined as:

$$t_N = \frac{\text{sum of the initial energy of all the sensor nodes}}{\text{total energy dissipated in one round}}$$

The proposed scheme is implemented to determine the network lifetime as shown in Fig. 4. The network consists of 200 nodes randomly distributed and 5 advanced nodes are uniformly distributed within the network. The proposed and three other selected schemes were implemented and run for 1200 rounds. All the schemes dissipate their energy slowly as the simulation time increases. The result shows that the proposed, AMACO, has more number of alive nodes than EBAR, IACO-MS, and RRDLA. The first node dies (FND) at 263rd round in RRDLA, at 268th round in IACO-MS, at 302nd round in EBAR, and 387th round in AMACO. Similarly, the last node dies (LND) at 391th, 714th, 683th, and 951th in RRDLA, IACO-MS, EBAR, and AMACO respectively. In all the scenarios, AMACO had the highest number of alive nodes and better performance in the network lifetime compared to RRDLA, IACO-MS, and EBAR. The reason is that during intra-cluster communication, sensor nodes are only responsible for sensing and short distance transmission which conserve their limited energy. Additionally, inter-cluster communication in AMACO only CHs (advanced nodes) communication with fog nodes and transmit through optimal paths.

TABLE I.   SIMULATION PARAMETERS AND THEIR VALUES

| Parameters	Values
Network area	100 x 100 m^2
Initial energy of nodes	0.5 – 2.0J
Number of nodes	200
Energy consumption for sending unit of data	0.01J
Transmission range	75m
Number of Advanced nodes	5
Path loss exponent (Φ)	4
Multipath component (∇)	0 dB to 10dB
Standard deviation ($X_\sigma \sigma$)	0 dB to 12dB
Energy consumption on circuit ((E_{elec})	50 nJ/bit
Free-space model (ε_{fs})	10 pJ/bit/m^2
Multi-path model (ϵ_{amp})	0.0013 pJ/bit/m^4
Data packet size	500 bytes
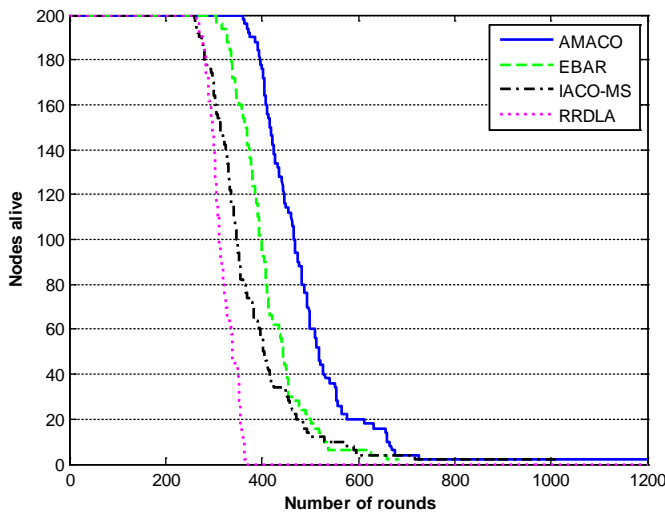
Fig. 4.   Nodes Alive Against the Number of Rounds.

B. Comparison of the Network Lifetime

The network lifetime is one of the important metrics of WSNs. The energy of a sensor node is mostly consumed on data transmission during network operation. If the limited energy of these nodes is completely exhausted, energy holes will be created and some nodes will not be able to communicate with the destination nodes. This will affect the normal operation of the entire network. We performed various simulations to compare the network lifetime of the four schemes. We varied the number of nodes in the network area from 40, 80, 120, 160, and 200 for each of the algorithms and present the simulation results in Fig. 5. We can see that the AMACO scheme has the longest network lifetime among all the four schemes. The reason is that the AMACO algorithm only chooses nodes in which their energy is above the energy threshold as neighbouring nodes and avoids the low energy nodes in the process of data transmission to the CHs. Thus, the proposed algorithm selects the optimal and efficient paths in the energy-abundant nodes.
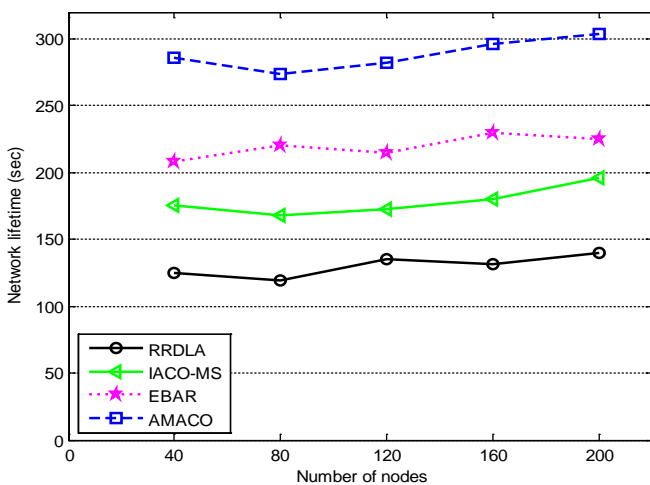

Fig. 5.   Network Lifetime Against the Number of Nodes.

C. Sum of Energy Consumption

The sum energy consumption of the sensor network is defined as the total energy consumed by each sensor node of the whole network. The lesser the sum energy consumption of the network, the lesser the routing cost, thus the longer the network lifetime. Fig. 6 presents the sum of the energy consumption against the number of rounds for all the algorithms; the energy of the network increases, with the number of rounds. RRDLA consumes more energy than the other algorithms. The reason is that the scheme adopts a multi-path for data transmission and did not consider the residual energy of nodes that are selected as relay nodes. There is a high possibility for the nodes to transmit their data through long-distance either to their neighbouring nodes or CHs and hence more energy will be consumed. Conversely, the sum of energy consumption in AMACO is less than the other three schemes the reason is that the improved ACO algorithm considers the dynamic optimization of the ants while taking into account the load balance characteristics, thus it prevents the intermittent ant diffusion, and conserves the energy consumption of the network, therefore achieving the purpose of prolonging the network lifetime.

D. The Average Number of Hops

We define the average number of hops as the number of links traversed by a data packet in the network. The longer the average number of hops is, the heavier the data packet traffic along the path is. So, there is a positive association between the end-to-end delay and the average number of hops. In Fig. 7, we can see that the performance of RRDLA algorithm is worst, the reason is that the algorithm uses single-hop communication for data packet transmission to the base station. Our proposed AMACO dynamically considers both single-hop and multi-hop communication to transmit data to the fog nodes. Each sensor node uses the energy model presented in Section 3 to transmit data to its associated cluster head.
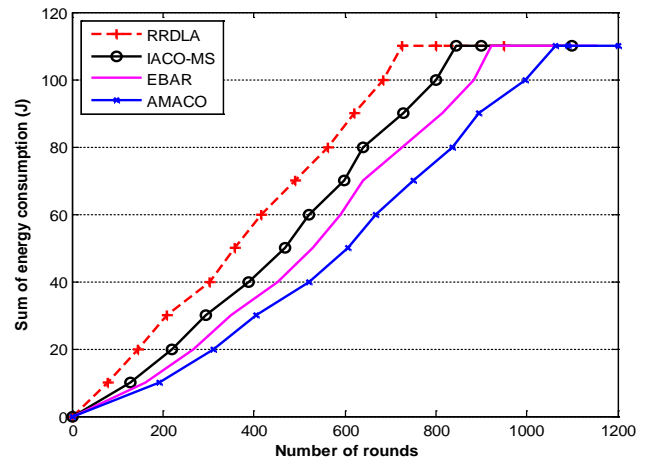

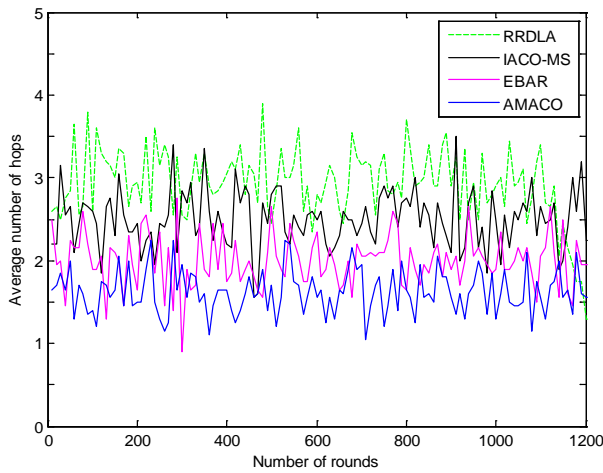Fig. 6.   Sum of Energy Consumption Against many Rounds.

Fig. 7.   The Average Number of Hops Against the Number of Rounds.

E. *Packet Delivery Ratio*

A Packet delivery ratio PDR: It can be measured as the ratio of the total number of data packets delivered to the total number of data packets transmitted from the sender node to the receiver node in the network. PDR of various algorithms is presented in Table II and Fig. 8. We observed that as the number of rounds increases, the PDR decrease linearly, the reason is that sensor nodes dissipate energy slowly as the number of rounds increases. There is a high likelihood of energy holes to be created among the nodes. RRDLA has the least PDR because its approach is similar to LEACH protocol. The method used for CH selection incurs high overhead. Similarly, IACO-MS algorithm is developed in a way that if a sensor node transmits a data packet to the base station, the data pass through all the sensor nodes as it is moving to the base station. Therefore, the overhead in the delivery of the data packet to the base station increase for every sensor node present in the network. AMACO has the highest PDR compared to the three algorithms. The reason is that it avoids overhead that is usually involved during cluster formation and selection of cluster heads in the previous approaches.

F. *Success Rate*

The average time per iteration of all the schemes and the success rate of searching for the optimal solution in 75 iterations of the four schemes are presented in Table II. AMACO algorithm has a higher data packet delivery rate compared to the other three schemes in computing time. The simulation runs for 75 random topologies to ensure consistency of the results to assess the success rate of searching for the optimal solution of all the schemes. We observe that AMACO has great potential to search for the global optimal solution. Its success rate is more than compared of algorithms. It can reliably and efficiently improve the routing computation and high success rate of searching for the global optimal solution. Thus, it shows that the AMACO algorithm is effective and reliable.

TABLE II.   AVERAGE TIME PER ITERATION AND THE SUCCESS RATE OF SEARCHING FOR THE OPTIMAL SOLUTION OF THE ALGORITHMS

Name of algorithms	Average time for one iteration (ms)	Success rate (%)
AMACO	4.38	99.23
EBAR	6.17	93.60
IACO-MS	10.54	85.49
RRDLA	16.92	54.17

	Packet delivery ratio (%)			
Number of rounds	RRDLA	IACO-MS	EBAR	AMACO
200	84.2	86.4	92.4	97.2
400	84.6	88.6	91.3	95.8
600	85.8	88.1	92.7	97.4
800	84.9	88.5	92.4	95.7
1000	83.4	86.7	92.1	96.5
1200	82.7	86.2	91.5	96.7
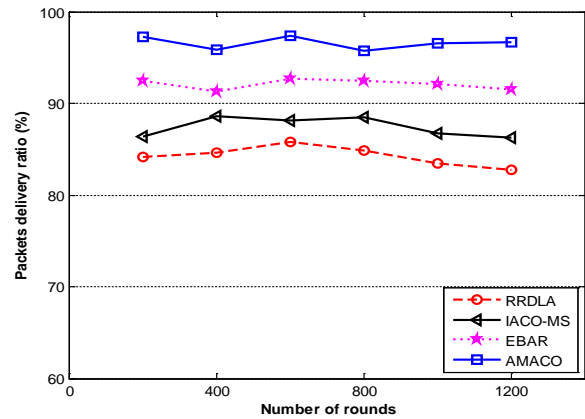
Fig. 8.   Packet Delivery Ratio Against the Number of Rounds.

VI. CONCLUSION

Clustering and fog nodes integration technologies are effective techniques to improve the performance of WSNs. In this paper, we presented an improved ant colony optimization routing algorithm to prolong the network lifetime of WSNs. We first partition the entire network area into different clusters. An algorithm is designed to ensure that advanced nodes which are automatically selected as CHs are evenly distributed within the network and assign a CH to a cluster. Each node belongs to a cluster based on its residual energy and distance from the CHs. An improved ACO algorithm is developed to construct an optimal path between the CHs and fog nodes for efficient and reliable data transmission. NS-2 Simulator is used to measure the performance against three related algorithms namely RRDLA, IACO-MS, and EBAR using the following metrics: network lifetime, the sum of energy consumption, the average number of hops, packet delivery ratio, and success rate. The simulation results show that AMACO performs better in terms of energy consumption and data packet delivery than the other three algorithms.

VII. Future Work

In the future, we intend to implement all the algorithms in a real test-bed using the above metrics.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Authors' Contributions

All authors contributed and approved the final manuscript.

Data Availability

The raw data of the IoT devices used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] L. Shkurti, X. Bajrami, E. Canhasi, B. Limani, S. Krrabaj, and A. Hulaj, "Development of ambient environmental monitoring system through wireless sensor network (WSN) using NodeMCU and "WSN monitoring"," in: Proceedings - 6th Mediterranean Conference on Embedded Computing (MECO), pp. 1-5, Bar, Montenegro 11-15 June, 2017.

[2] J. Aranda, D. Mendez, and H. Carrillo, "Multimodal wireless sensor networks for monitoring applications: a review," Journal of Circuits, Systems and Computers.vol. 29, no. 02, pp. 203-216, 2020.

[3] B. Bhushan and G. Sahoo, Routing protocols in wireless sensor networks. In Computational intelligence in sensor networks, Springer, 2019; pp. 215-248.

[4] M. Elshrkawey, S. M. Elsherif, and M. E. Wahed, "An enhancement approach for reducing the energy consumption in wireless sensor networks," Journal of King Saud University-Computer and Information Sciences.vol. 30, no. 2, pp. 259-267, 2018.

[5] A. Sarkar and T. S. Murugan, "Cluster head selection for energy efficient and delay-less routing in wireless sensor network," Wireless Networks.vol. 25, no. 1, pp. 303-320, 2019.

[6] F. A. Aoudia, M. Gautier, M. Magno, M. Le Gentil, O. Berder, and L. Benini, "Long-short range communication network leveraging LoRa™ and wake-up receiver," Microprocessors and Microsystems.vol. 56, pp. 184-192, 2018.

[7] X. Bai, Z. Wang, L. Sheng, and Z. Wang, "Reliable data fusion of hierarchical wireless sensor networks with asynchronous measurement for greenhouse monitoring," IEEE Transactions on Control Systems Technology.vol. 27, no. 3, pp. 1036-1046, 2018.

[8] C. Mallick and S. Satpathy, "Challenges and Design Goals of Wireless Sensor Networks: A State-of-the-art Review," International Journal of Computer Applications.vol. 179, no. 28, pp. 42-47, 2018.

[9] Z. Zou and Y. Qian, "Wireless sensor network routing method based on improved ant colony algorithm," Journal of Ambient Intelligence and Humanized Computing.vol. 10, no. 3, pp. 991-998, 2019.

[10] G. Samara and M. Aljaidi, "Efficient energy, cost reduction, and QoS based routing protocol for wireless sensor networks," arXiv preprint arXiv:1903.09636. 2019.

[11] H. Mostafaei, "Energy-efficient algorithm for reliable routing of wireless sensor networks," IEEE Transactions on Industrial Electronics.vol. 66, no. 7, pp. 5567-5575, 2018.

[12] V. Srivastava, S. Tripathi, and K. Singh, "Energy efficient optimized rate based congestion control routing in wireless sensor network," Journal of Ambient Intelligence and Humanized Computing.vol. 11, no. 3, pp. 1325-1338, 2020.

[13] X. Liu and J. Wu, "A method for energy balance and data transmission optimal routing in wireless sensor networks," Sensors.vol. 19, no. 13, pp. 216-228, 2019.

[14] M. Z. Hasan, H. Al-Rizzo, and F. Al-Turjman, "A survey on multipath routing protocols for QoS assurances in real-time wireless multimedia sensor networks," IEEE Communications Surveys & Tutorials.vol. 19, no. 3, pp. 1424-1456, 2017.

[15] P. M. Chanal, M. S. Kakkasageri, A. A. Shirbur, and G. S. Kori, "Energy aware multipath routing scheme for wireless sensor networks," in: Proceedings - 2017 IEEE 7th International Advance Computing Conference (IACC), pp. 313-317, 2017.

[16] S. Othmen, F. Zarai, A. Belghith, M. S. Obaidat, and L. Kamoun, "Secure and Reliable Multi-Path Routing Protocol for Multi-Hop Wireless Networks," Adhoc & Sensor Wireless Networks.vol. 36, no. 4, pp. 617-629, 2017.

[17] Z. Li, M. Xu, T. Liu, and L. Yu, "A Network Coding-Based Braided Multipath Routing Protocol for Wireless Sensor Networks," Wireless Communications and Mobile Computing. 2019.

[18] A. Aljumah and T. A. Ahanger, "Cyber security threats, challenges and defence mechanisms in cloud computing," IET Communications.vol. 14, no. 7, pp. 1185-1191, 2020.

[19] S. K. Gawali and M. K. Deshmukh, "Energy Autonomy in IoT Technologies," Energy Procedia.vol. 156, pp. 222-226, 2019.

[20] X. Chen, Y. Zhou, B. He, and L. Lv, "Energy-efficiency fog computing resource allocation in cyber physical internet of things systems," IET Communications.vol. 13, no. 13, pp. 2003-2011, 2019.

[21] A. V. Dastjerdi and R. Buyya, "Fog computing: Helping the Internet of Things realize its potential," Computer.vol. 49, no. 8, pp. 112-116, 2016.

[22] A. Yassine, S. Singh, M. S. Hossain, and G. Muhammad, "IoT big data analytics for smart homes with fog and cloud computing," Future Generation Computer Systems.vol. 91, pp. 563-573, 2019.

[23] B. Marques, I. M. Coelho, A. D. C. Sena, and M. C. Castro, "A network coding protocol for wireless sensor fog computing," International Journal of Grid and Utility Computing.vol. 10, no. 3, pp. 224-234, 2019.

[24] 'Cisco Visual Networking Index: Forecast and Methodology, 2014-2019', http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf, Accessed March 2016.

[25] S. P. Singh, A. Nayyar, R. Kumar, and A. Sharma, "Fog computing: from architecture to edge computing and big data processing," The Journal of Supercomputing.vol. 75, no. 4, pp. 2070-2105, 2019.

[26] S. Yi, Z. Hao, Z. Qin, and Q. Li, "Fog computing: Platform and applications," in: Proceedings - 3rd IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb), pp. 73-78, Washington, DC, USA, 12-13 Nov., 2015.

[27] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," IEEE Journal on Selected Areas in Communications.vol. 37, no. 6, pp. 1205-1221, 2019.

[28] M. Dorigo. Ottimizzazione, apprendimento automatico, ed algoritmi basati su metafora naturale. Unpublished doctoral dissertation, PhD thesis, Dipartimento di Elettronica, Politecnico di Milano, 1992.

[29] S. Abi, B. Benhala, H. Bouyghf, and M. Fakhfakh, "A Comparative Study between ACO and DE Techniques by Numerical Functions Optimization," in: Proceedings - 5th Int'l Conf. on Optimization and Applications (ICOA), pp. 1-6, 2019.

[30] A. Gonzalez-Pardo, J. Del Ser, and D. Camacho, "Comparative study of pheromone control heuristics in ACO algorithms for solving RCPSP problems," Applied Soft Computing.vol. 60, pp. 241-255, 2017.

[31] V. Sharma and A. Grover, "A modified ant colony optimization algorithm (mACO) for energy efficient wireless sensor networks," Optik-International Journal for Light and Electron Optics.vol. 127, no. 4, pp. 2169-2172, 2016.

[32] M. Radi, B. Dezfouli, K. A. Bakar, S. A. Razak, and T. Hwee-Pink, "IM2PR: interference-minimized multipath routing protocol for wireless sensor networks," Wireless Networks.vol. 20, no. 7, pp. 1807-1823, 2014.

[33] O. Deepa and J. Suguna, "An optimized QoS-based clustering with multipath routing protocol for wireless sensor networks," Journal of King Saud University-Computer and Information Sciences. pp. 1-12, 2017.

[34] S. Sharma and S. K. Jena, "Cluster based multipath routing protocol for wireless sensor networks," ACM SIGCOMM Computer Communication Review.vol. 45, no. 2, pp. 14-20, 2015.

[35] G. P. Gupta and S. Jha, "Integrated clustering and routing protocol for wireless sensor networks using Cuckoo and Harmony Search based metaheuristic techniques," Engineering Applications of Artificial Intelligence.vol. 68, pp. 101-109, 2018.

[36] J. Wang, J. Cao, R. S. Sherratt, and J. H. Park, "An improved ant colony optimization-based approach with mobile sink for wireless sensor networks," The Journal of Supercomputing.vol. 74, no. 12, pp. 6633-6645, 2018.

[37] M. Krishnan, Y. M. Jung, and S. Yun, "An Improved Clustering with Particle Swarm Optimization-Based Mobile Sink for Wireless Sensor Networks," in: Proceedings - 2nd Int'l Conference on Trends in Electronics and Informatics (ICOEI), pp. 1024-1028, Tirunelveli, India, May, 2018.

[38] S. Arjunan and P. Sujatha, "Lifetime maximization of wireless sensor network using fuzzy based unequal clustering and ACO based routing hybrid protocol," Applied Intelligence.vol. 48, no. 8, pp. 2229-2246, 2018.

[39] X. Li, B. Keegan, F. Mtenzi, T. Weise, and M. Tan, "Energy-Efficient Load Balancing Ant Based Routing Algorithm for Wireless Sensor Networks," IEEE Access.vol. 7, pp. 113182-113196, 2019.

[40] M. Sefuba and T. Walingo, "Energy-efficient medium access control and routing protocol for multihop wireless sensor networks," IET Wireless Sensor Systems.vol. 8, no. 3, pp. 99-108, 2018.

[41] Heinzelman, A. Chandrakasan, and H. Balakrishnan, "An application - specific protocol architecture for wireless microsensor networks," IEEE Transactions on Wireless Communications.vol. 1, no. 4, pp. 660-670, 2002.

[42] S. Kumar and D. Lobiyal, "Sensing coverage prediction for wireless sensor networks in shadowed and multipath environment," The Scientific World Journal.vol. 2013, 2013.

[43] Y. Yang, X. Luo, X. Chu, and M.-T. Zhou, Fog Computing Architecture and Technologies. In Fog-Enabled Intelligent IoT Systems, Springer, 2020; pp. 39-60.

[44] S. Namasudra, "An improved attribute‐based encryption technique towards the data security in cloud computing," Concurrency and Computation: Practice and Experience.vol. 31, no. 3, pp. 4352-4364, 2019.

A Flood Forecasting Model based on Wireless Sensor and Actor Networks

Sheikh Tahir Bakhsh[1], Mohammed Basheri[4]

Information Technology Department
King Abdulaziz University
Jeddah, Saudi Arabia

Naveed Ahmed[2], Basit Shahzad[3]

Faculty of Engineering and Computer Sciences
National University of Modern Languages
Islamabad, Pakistan

Abstract—**Flood forecasting is a challenging area of research that can help to save precious lives by timely intimating about the flood possibilities. The role of advancements in computing and allied technologies has moved the research towards a new horizon. Researchers from all over the world are using Artificial Neural Networks (ANN), Global Information Systems (GIS), and Wireless Sensor Networks (WSN) based schemes for flash flood forecasting and hydrological risk analysis. ANN and GIS-based solutions are much costly whereas the analysis and prediction using WSN require much less cost for infrastructure deployment. It will enable the use of flood prediction mechanisms in the third world and poor countries. New variation in storage capacity can be a vital source to eliminate or reduce the chance of flood. By considering this observation, it is proposed to develop a generic flood prediction scheme that can manage the system as per resources and environmental conditions. A heterogeneous WSN has considered where powerful Collector Nodes (CN) continuously takes values from member sensor nodes in the region. CN transmits the alerts to Administration Unit (AU) when threshold values are crossed. It is proposed to save the threshold values from all water sources like storage capacity, water inflow, and outflow in the repository for decision making. Moreover, environmental parameters including water level, humidity, temperature, air pressure, rainfall, soil moisture, etc. are considered for the analysis of flood prediction. We have also considered the evaluation of these parameters in specified boundary values that were not considered in existing schemes. In this research study, we have used Arc GIS and NS2 simulation tools to analyze the parameters and predict the probability of the occurrence of a flood.**

Keywords—*Flood forecasting; GIS; remote sensing; hydrology; particle swarm optimization*

I. INTRODUCTION

Flood forecasting is the real-time weather forecasting metrological technique used in the rainfall-runoff process leading towards the flow of water levels in the specific region of interest. Flood prediction helps the hydrologists to investigate when and where the extreme event of Flood will occur [1]. It depends on the intensity of rainfall occurred based on the rate of change in the environmental conditions. Accurate flood prediction helps scientists and researchers to understand the relationship between flood forecasting and rainfall [2]. The web-based technologies play a vital role in the prediction of floods and the management of other disasters [3]. Wireless sensor networks (WSN) consist of small low power nodes that can collectively sense the environment and transfer the desired readings to a central server. WSN provides an early flood prediction scheme, especially in underdeveloped countries. The rainfall gauges installed on the wireless sensor network also plays a vital role in calculating flood flow intensity and thrust value. Fig. 1 provides in deep detail the role of wireless sensor networks used especially in emergency circumstances [4-6]. The sensor nodes are responsible to transfer the disaster-related information (rainfall intensity including Flood water level) to the cluster head nodes. The cluster nodes exist in the middle of the multiple sensor nodes. The cluster nodes are capable to communicate with each other and to the gateway node. The gateway node is connected to the centralized global communication infrastructure.

Artificial neural networks (ANN) provide a mechanism to analyze and make a decision based on existing data records. Flood prediction schemes based on ANN are discussed under the following two categories. Particle swarm optimization schemes [7, 8] are based on a computational method that optimizes the complex problem which results in the improvement of a candidate solution. ANN-based fuzzy logic-based Schemes [9, 10] are based on the integration of neural networks and fuzzy logic. For the implementation of an Artificial neural fuzzy inference system, we can select the best parameters for flood forecasting based on genetic algorithms. Time series based data mining schemes [11] are also used in flood forecasting. It has been used to predict the continuous values of data based on historical trends. The results had been obtained based on hydro metrological data based on the successfully obtained records stored in the database. WSN and GIS technologies are also merged to provide a solution to flooding prediction techniques. A geo sensor network plays a vital role in acquiring spatial information. They can be successfully deployed in air, buildings, underwater, and on the body having onboard processing capability [12].

The research report emphasizes its focus on the implementation of direct diffusion protocol in case of flood events. The floods are one of the major events of environmental disasters. The major objective of this research is the water level monitoring during the extreme event of floods. The simulation setup in the proposed network model has been implemented by WSN configuration by using OMNET++ discreet event simulator. The scalability is an important factor in WSN which includes end to end transmission delay versus throughput. Throughput is the classification of network performance transmitted over bits per second on the selected communication channel. The simulation results in Fig. 2 had

been obtained graphically in which the x-axis represents the number of wireless sensor nodes and the y-axis represents the average energy cost per sensor node [13].

Flood prediction schemes are discussed categorically to identify the level of work already done in this area of research. The review is generally based on ANN, WSN, and GIS-based schemes used in managing the flood-related risks and doing effective forecasting in the areas of disaster management. ANN is used to extract the required data either from existing datasets containing flood-related details for previous years or live data capturing form radars and sensor networks. Existing schemes are discussed in subsections and shown in Fig. 3. The network architecture for the flood forecasting system consists of a distributed prediction algorithm. The optimum intermediate deployment scheme has been used for WSN based communication infrastructure. The fuzzy risk analysis scheme is implemented employing the probability distribution method [1].
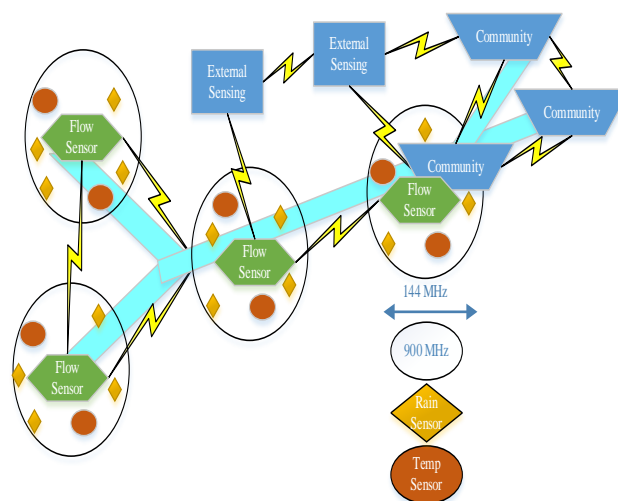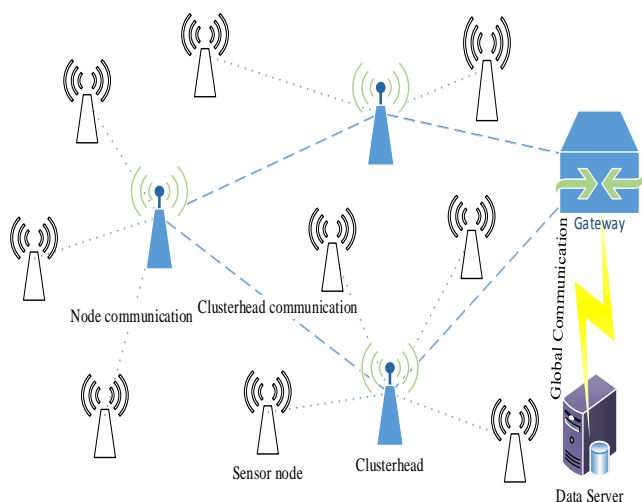

Fig. 1. A Real-Time Wireless Sensor Network used in Flood Forecasting.
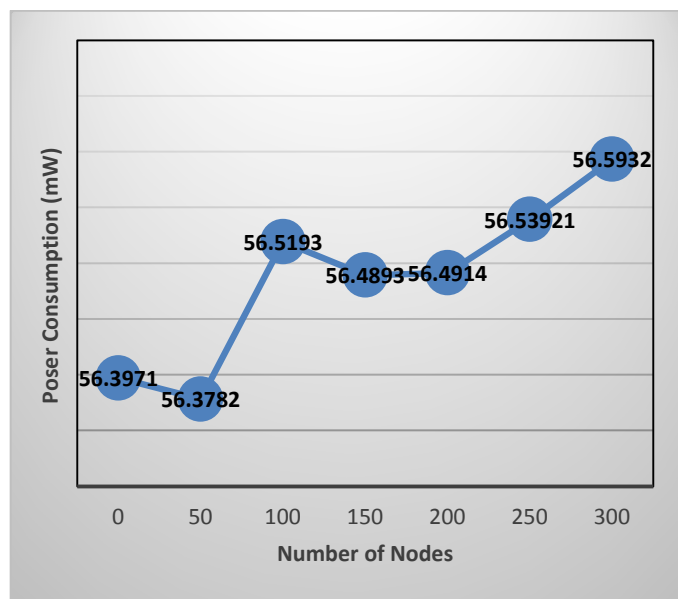
Fig. 2. Wireless Sensor Nodes Versus Power Consumption Overhead.

Fig. 3. Deployment of Wireless Sensor Nodes for Flood Forecasting.

The rest of the paper is organized into the following sections. Section 2 provides a literature review based on flood forecasting techniques. Section 3 provides a problem statement and research questions. Section 4 and 5 provide the proposed solution and research discussions respectively. Finally, the paper is concluded in Section 6.

II. LITERATURE REVIEW

The mathematical formulation of the Flood forecasting scheme is designed based on the geometrical distribution of wireless sensor nodes [14-20]. Fig. 4 explains in deep detail about the deployment scheme related to the different categories of wireless sensor nodes. The computational nodes are responsible for strong processing capability. For every two computational nodes in between the middle exists duplicate node. Duplicate node is responsible to ensure connectivity between the left and right computational node. The Intermediate nodes are responsible for transferring the river flow condition from the computational nodes to a centralized office node, the head node in this case. The researchers had also proposed the distributed prediction model based on air temperature, rainfall, and water discharge level [21].

Fig. 4. Distribution of Wireless Sensor Node for the Scheme.

The analysis of flood management is being done by using the multi-layer perception, neural network model. The parameters being used for this purpose are the velocity of water in meter per second, humidity in air measured in per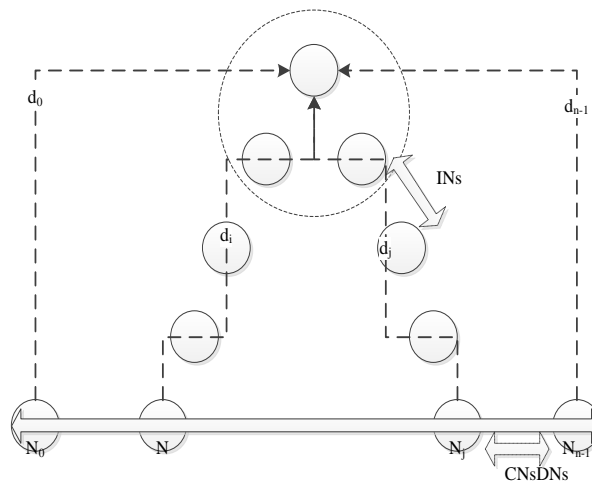centage, the rain in millimeters, the speed of wind in kilometers in each hour, and the pressure of the atmosphere is measured in its respective units. Though there are several tools for implementing the artificial neural networks (ANN) but the for clarity and simplicity the implementation of ANN is being done by using the Matlab simulation tool [10] as shown in Fig. 5. The formation of the artificial neural network model is such that it consists of three layers, the layer that deals with the inputs are called the input layer, the layer dealing with the outputs is called output layer, while the third layer that deals with the processing of the ANN are called the hidden layer. The output of the ANN model demonstrates the warning situation about the possible prediction of flood at low, high, or very high levels.

In hydrological modeling, it is very common to use particle swarm optimization technique. Particle swarm optimization is a technique based on stochastic groups and is considered to be a segment of evolutionary computing algorithms that were developed by Kennedy and Eberhart. In PSO techniques there exists a group of particles that are formed at random and are located in the positions which are already specified and move through the entire space to potentially locate the expected solution. The particles learn by themselves while living within a group by using the neural network and also work for solution identification. The emphasis of the PSO is on the effective and appropriate fitting of the particles in the space and are called as personal best (pbest) and global best (gbest) collectively depending on the speed of the movement towards the velocity of the particle in the hyperspace. PSO uses two equations [7]. The equation (1) is known as movement equation:

$$Loc_{Present} = Loc_{Previous} + V_i \Delta t \qquad (1)$$

The location of the particle changes and the current position of the particle can be determined by using the past position ($Loc_{Previous}$) of the particle in the described vector space along with the velocity of that specific particle in a specific time frame ($V_i \Delta t$).

$$V_i = wV_{i-1} + C_1 * rand() * (Pbest - preslocation) + C_2 * rand() * (gbest - prelocation) \qquad (2)$$

Equation (2) records and elaborates on the change in velocity for the identified particles within the whole space of the movement known as the personal best and global best. In the previous equation V_i is the initial velocity of the particles, Δt is the time interval for the movement of particles with n a hyperspace, V_{i-1} is the previous velocity, random() is the random number value for example (0,1,2,3..........) and C_1, C_2 are the acceleration of the particles within the entire space. The core of this research work is based on identifying the application of particle swarm optimization algorithm for improving the learning process.

Artificial neural networks fuzzy logic-based scheme is a multilayer feedforward and backpropagation algorithm which deals in forwarding the weighted connections from input layers

towards the output layers [9]. The artificial neural network model finds a set of attributes that formulates the basis for the If-then fuzzy rules based on the relevant number functions. To maintain efficiency in mathematical modeling and optimization technique Sugneo interface system has been adopted. The 1st-order Sugeno fuzzy model can be described mathematically using the following equations. Fig. 6 demonstrates the formation of the ANFIS system which consists of the output layer as Layer 5, the input layer as Layer 1, and the hidden layer as Layers 2, 3, and 4.

Rule 1: *If X is A_1 and Y is B_1, then $f_1 = P_{1x+q1y+r1}$* (3)

Rule 2: *If X is A_2 and Y is B_2, then $f_2 = P_{2x+q2y+r2}$* (4)

The above mentioned Rule-based equations depend on the output function f corresponds to the input vector value *X* and *Y*. The values *p*, *q*, and *r* represent the constant quantities.
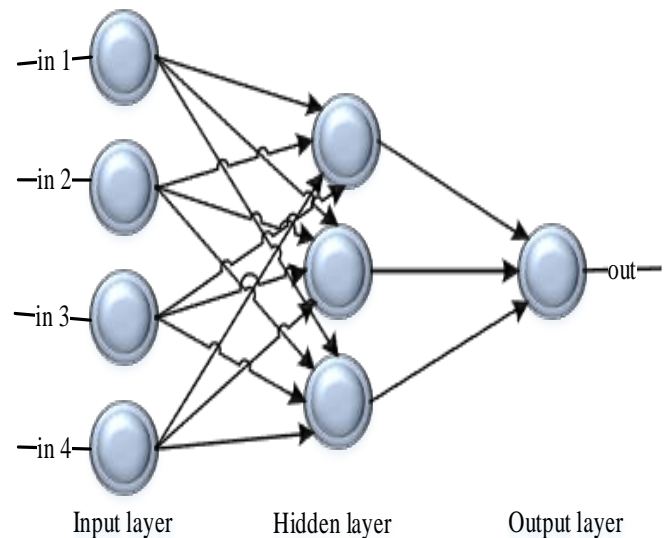

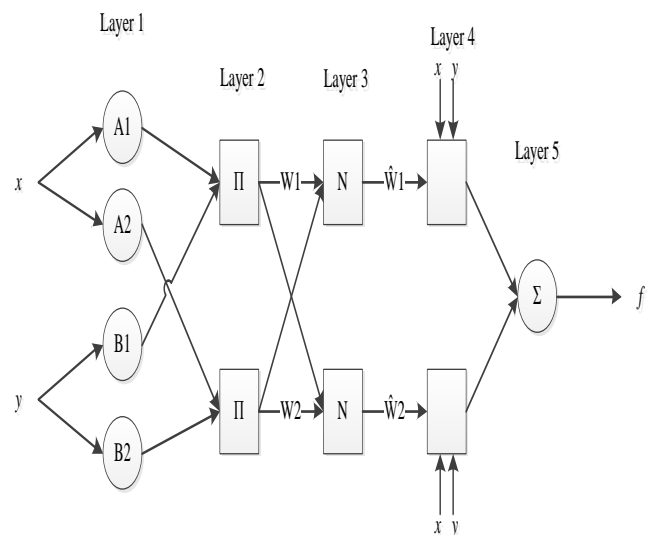Fig. 5. Model of Artificial Neural Network for Flood Prediction.

Fig. 6. Five-Layer Feed forward ANFIS System.

The use of Ad hoc WSN is pretty common nowadays in the real-life application in the areas like telemedicine, underwater wireless sensor networks and disaster management, and vehicle Ad hoc networks. Disaster management consists of many such applications that are of direct benefit to mankind and includes floods, earthquakes, fire, hurricanes. WSN's can be seen as energy-efficient multi-hopping systems having the capability to transfer information across nodes using Ad hoc relay station [22-29]. The model that has been proposed in this paper uses web-based GIS. The prediction model takes live data from the gateway in the locality of the flood-affected area. The node on the gateway is connected to sensor nodes working on p-to-p architecture. The adaptive sampling results are based on different time events that demonstrate the working of sensors in allocated time gaps along with the intensity of variance of its signals at discrete events [31]. Fig. 7 provides detailed information about the fact that how the WSN and the GIS can be integrated and as a result, a framework for environmental modeling has been proposed and consists of the following parameters.

1) Identification and selection of the desired region on the map.

2) Application of the WSN on the map according to communication and sensing range.

3) Tabulation of data from sensors to the database.

4) Improving the database's capability to handle the increasing amount of data.

5) Deployment of the proposed framework by using the postgre SQL database.

The critical evaluation of the research study in Table I which is extracted from the literature discussed. It describes the role of GIS and WSN used in Flood forecasting along with their pros and cons. The Author [1] had discussed the Fuzzy logic based scheme for flood forecasting and disaster risk analysis. A distributed modeling technique is used for weather forecasting. The Metrological data set (COSMO) is used to calculate the rainfall peak discharge ratio for a relatively short duration [2]. The hydrological model is used to calculate flash fold events. The temporal and spatial resolution is used to calculate the rainfall-runoff process [5]. The authors [2, 5] had focused their research on using metrological models for extreme weather conditions including rainfall. ANN-based flood forecasting model has been designed based on input,

output, and hidden layers [10]. According to hydrologists the real-time flood analysis and prediction are performed atkaluganga river srilinka [33].

WSN based Statistical Model is designed to predict the extreme event of floods [17]. The authors [17, 26] had proposed the WSN based scheme to forecast extreme events of flood in case of disaster circumstance. The hydrological modeling technique is used to predict floods based on numerical weather prediction [28]. The authors [28, 30] had proposed the mathematical model for flood forecasting in case of disaster circumstance. GIS-based flood forecasting Model is designed to calculate the flood series frequency [31-42]. The authors [10, 35] had focused their research on designing the conceptual model for flood forecasting. The regression-based flood forecasting model is designed by [43] which is used to reduce the power consumption overhead by using WSN. The authors [26, 43] had proposed the WSN based solution for flood forecasting.
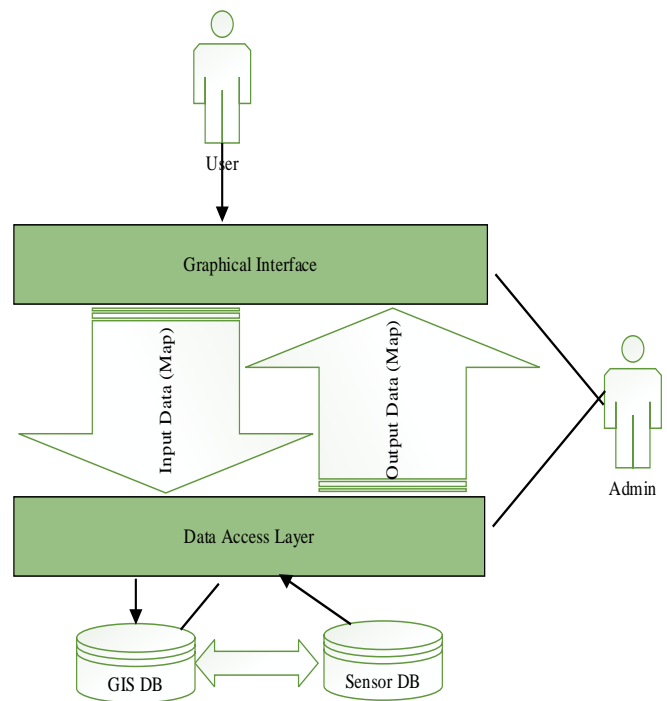

Fig. 7. Deployment of WSN with GIS Scenario.

TABLE II.   TABULAR REPRESENTATION OF GIS, WSN, AND FLOOD PREDICTION TECHNIQUE USED ALONG WITH PROS AND CONS

Paper Ref	WSN and GIS-based Flood Forecasting schemes	Frameworks/Models discussed	Strengths	Limitations
[1]	Flood Prediction Model for Fuzzy Logic has been discussed	Flood Risk Model is designed for a Probability distribution	Flood Risk probability is calculated in a specific region	Mathematical Model is complex
[2]	Probabilistic Model for Flood forecasting has been designed	Weather forecasting is based on Distributed Systems	Peak discharge ratio is calculated by using a Metrological data set (COSMO)	Test duration period is relatively short, Seventeen months
[5]	Runoff Process and peak-flow discharge process is numerically calculated	hydrological models are used to predict flash flood events	Temporal rainfall process and spatial resolution is used to calculate flood runoff process	Complex Framework Model is designed
[10]	ANN-based rainfall-runoff Model has been designed	System Model has been designed based on input, output and hidden layer	Input parameters for flood forecasting have been designed	Simulation results are not obtained
[17]	WSN based Flood Forecasting Model has been designed	Flash Flood events are predicted based on the Statistical Model	Graphical results are based on instantaneous water levels	Node deployment scheme is based on Hybrid WSN
[26]	Monitoring for early warning Flood system based on WSN	Flood prediction Algorithm is designed	An algorithm is implemented in Mat lab	Difficult to implement in a real-life scenario
[28]	Ensemble Prediction System (EPS) is based on Flash Flood Forecasting	Hydrological Modeling (deterministic prediction) technique has been used	Numerical weather prediction is used for Flood warning situations	un certainty in future rainfall Modeling has been identified
[30]	Bayesian Flood Forecasting systems (BFS) has been designed	The stochastic distribution Model is designed for the prediction of Flash Flood disasters.	predictive distribution scheme is used for accurate calibration	Flood Forecast Model generated uncertain results
[31]	Flood assessment is Scheme is based on GIS-based WSN	Mathematical Model is designed for calculating Flood series frequency	Flood estimation methodology has been designed	Environmental parameters are nor identified properly for Flash Flood Forecasting.
[35]	Dynamics of a drainage system is based on Sewer System Simulation	Conceptual Model in case of Flood foresting is designed for Urban areas	Rain Fall and Runoff Analysis for a hundred-year storm has been calculated	Statistical Framework has not been designed
[43]	WSN based Algorithm is designed for flood forecasting	The flood forecasting model is based on linear regression	Mathematical Model is designed to reduce power consumption overhead	Percentage error is obtained based on the predicted water level

III.   PROBLEM STATEMENT AND RESEARCH QUESTIONS

Flood forecasting schemes using ANN require extensive computations, a large amount of data for training, large storage, and high processing capabilities. The numbers of input parameters are fixed for different regions that make it a rigid one. The major limitation of the existing research is that the river basin and other water reservoirs towards the dam or barrage are considered static entities. The impact of environmental parameters on flood prediction is based on existing readings taken from the old infrastructure of water reservoirs and rivers etc. The number of river basins and canals towards the dam may change over the passage of time and new entities can be added in the existing infrastructure. It can vary the flow, speed, storage, and level of water entering in the dams. Due to these dynamic parameters, the accuracy of flood prediction is highly reduced and even can produce false results. System reliability reduces as there are no redundancy checks to identify and correct the false calculations. The resultant values calculated based on false or out of bound environmental parametric values can generate wrong results. It could be much worst to produce opposite results in case negative values are generated instead of positive ones. These false results can be further saved in a database for future calculations for flood prediction. These false calibrations may result in inaccurate predictions which may cause extreme damage.

The following are the research questions that will be solved during this work:

1) How to design a Flood Forecasting Model by using GIS-based Wireless Sensor Networks?

2) How to collect the Flood Forecasting input Parameters by using Ad-hoc wireless Sensor architecture.

3) To design a Mathematical Model for Flood Forecasting by using Polynomial based approximation.

4) To generate the results of the Flood Forecasting technique by plotting graphs and pie charts.

5) Finally, perform the comparative analysis of the Actual and predicted Results.

The significance of the proposed research study is based on the critical evaluation of the literature studied during this research. This research study investigates how to build standard rules and regulations to be followed by developing nations in case of any flood emergency circumstance. The core issue for the research mainly depends upon the classification of determining flood level duration which directly affects the STRUCTURE of the buildings in the residential area. Almost the majority of Authors had discussed the applications of GIS, ANN, and WSN in flood event circumstances, but some of the Authors had proposed that using special-purpose sensors and hydrology based parameters plays a significant role in flood risk assessment and adopting precautionary measurements. The

Remote sensing approach can also be adopted for the potential impact of flood risk depending on the specified geographical region. This work will be effective in managing the water level in dams and open the spillways for small intervals of time as per the actual picture of input water to either avoid flood situations or keep the water level to a lower value in flood. It will also help in analyzing the water levels and environmental parameters to predict the flood before time. It will save or reduce the damage in rural and agricultural areas in terms of crops, livestock, human lives, and house holdings, etc.

IV. PROPOSED SOLUTION

This section presents an improved Flood warning system based on different parameters by using WSN and Artificial Neural Networks. Improvements in the low powered tiny sensor nodes working in Ad hoc infrastructure is one of the key factors in this research work. WSN used in flood forecasting is based on the efficient utilization of sensor nodes based on different clustering schemes. The proposed system consists of multiple wireless sensor nodes organized in a cluster group hierarchy deployed near to the river bed. The collector node is responsible for reading the flood characteristics from multiple sensor nodes. The collector node is also responsible for processing capability based on ANN-based polynomial approximation. In the system model, different nodes are involved to perform mandatory duties to fulfill the flood prediction scenario using WSN. Fig. 8 provides detailed information about the framework related to the proposed solution. The proposed framework consists of three clusters working in Flood forecasting and disaster risk assessment. Each cluster within the network consists of a collector node surrounded by multiple sensor nodes. The sensor nodes are responsible to transfer the disaster information to the collector node. The collector nodes within each cluster are capable to communicate with each other using Ad hoc wireless infrastructure. Each collector node is also capable to transmit the disaster-related information to the administration unit. The Administration Unit is finally responsible to update disaster information in its online database management system.
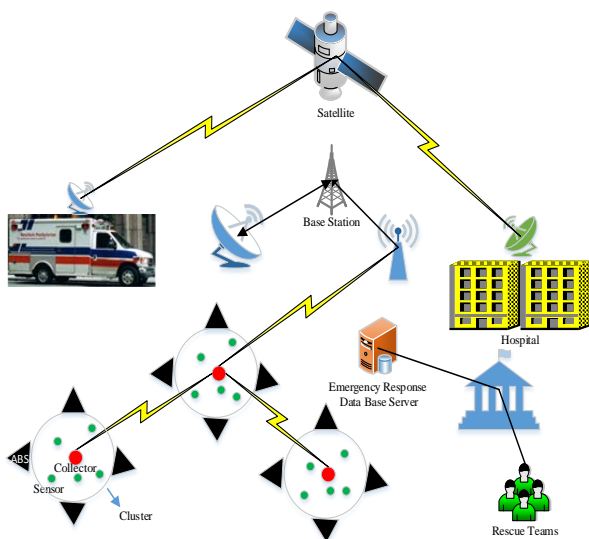

Fig. 8. Proposed Framework for Flood Forecasting.

The main responsibility of a sensor node is to sense the data from the surrounding and store in flash memory. The sensor can also GENERATE an alert when sensor readings are higher than the threshold values. The sensor can identify the query request and can send a reply to the collector node as per query requirements. These nodes are responsible to transmit the alarming messages to the AU. Moreover, CN can also collect the data in response to query from all nodes to transmit towards neighboring CNs that forwards towards the AU. The administrative activities and threshold values sent by AU are also coordinated via CN to sensor nodes. The Administration Unit is responsible to collect the disaster-related information (Flood forecasting parameters) from the collector nodes. The Administration unit has strong processing capabilities. The administration unit is capable to store the Flood forecasting parameters in its database. The Flood prediction Algorithm has been deployed on the Administration unit. In this research study, the most appropriate flood prediction technique has been used and also design its mathematical model using polynomial based approximation and artificial neural networks. The flood prediction input parameters would be very helpful for us in estimating the rate of flood occurrence.

During the applicability of the proposed solution, the following assumptions have been considered;

- An intrusion detection system should be implemented in a server at the Administration Unit (AU).

- AU should be capable of performing extensive parameter calculations regarding flood prediction.

- The sensor node should be able to keep a record of temperature, humidity, velocity, etc. It should also be able to generate alerts when the threshold value is crossed.

- Collector Node (CN) which is a sensor device should have sufficient resources in terms of energy, computation, and communication powers.

- CN should be able to collect, aggregate, and compress the water, rainfall, etc. data from member nodes.

- The power consumption factor in the case of wireless sensor nodes working in sleep mode should be reduced to a great extent.

Each deployed sensor node is a battery-powered device. The WSN consists of even hundreds or thousands of sensor nodes. The major constraints in WSN are based on communication bandwidth related to the sender node and receiver node. Sensor nodes are limited to battery-powered devices. In a multi-hop environment, a sensor node can act as a sender and as well as the receiver node. The major objective of this research related to the communication cost is dependent on using a limited size of data. This data can easily be transferred from the transmitter node to the receiver node to reduce the communication overhead. The storage unit in the wireless sensor node plays a very important role in data collection and aggregation. The size of data should be limited enough to be easily stored in the wireless sensor node. Sensor nodes can use the limited amount of energy in performing computations and transmissions. The sensor node consists of a sensing unit,

processing unit, and the transmission unit. Energy-efficient routing protocols also plays a strong role in data transmission and aggregation. The energy consumption factor is measured during the information exchange of flood-related parameters. WSN is capable enough to transfer the Flood forecasting information safely in a minimum period from the transmitter node to the receiver node. The time analysis is performed for transferring the information from the sensor to AU. The input parameters used for the WSN simulation are water velocity, water level, Air moisture, wind velocity, rainfall, and Atmospheric pressure. The resulting values are checked from sensors by comparing them with saved Boundary values of different water resources.

V. RESULTS AND DISCUSSION

The flood forecasting techniques are based on artificial neural networks (ANN), Particle swarm optimization (PSO), Regression-based technique, Fuzzy logic, and Time-series data mining. The SIMULATION scenario is created by using a network simulator (NS 2). The simulation results are collected by using ARC GIS simulation tools. The simulation of flood forecasting mathematical equations are used by using ANN-based polynomial approximation. The performance evaluation of simulated results has been identified. Arc GIS is a desktop-based application specially designed for working with maps based on geographic information. ARC GIS tool is capable to create the layered maps based on spatial data [22]. Network Simulator is an open-source simulation tool. Simulation of wireless and wired networks including routing algorithms TCP and UDP are implemented by using network simulator tools. NS 2 is based on C++ and Tool command language (TCL). It creates a trace file that stores the activities and events generated by devices under simulation. Finally, the results are extracted from the values printed in the trace file. The actual flood classification has been obtained from discharges at important river sites. The Analysis of actual flow depends upon actual inflow and outflow depending on reservoir designed capacity. The Real-time rainfall data set is very much helpful for us in identifying the flood forecasting parameters which include the intensity of rainfall in mm, temperature, air pressure, humidity along with wind speed. For this reason, the Ad hoc Wireless Sensor Network Architecture has chosen as a Model to collect flood forecasting parameters via Ad hoc wireless sensor Nodes.

The simulation results have been obtained by using a Linear regression model via using the sklearn library. Spyder version 3.3.6 (Scientific Python Development Environment) has been used to generate Flood Prediction results via using Anaconda Navigator. The structure of results is based on training and testing data sets in green and blue color. Fig. 9 X-axis represents the values that represent the number of months and along Y-axis are Flood prediction accuracy. Positive values represent the true positive and False positive values and negative values represent False-negative and true Negative values. The maximum convergence of positive training and testing values predicts the accuracy of the Prediction Model to Maximum extent.
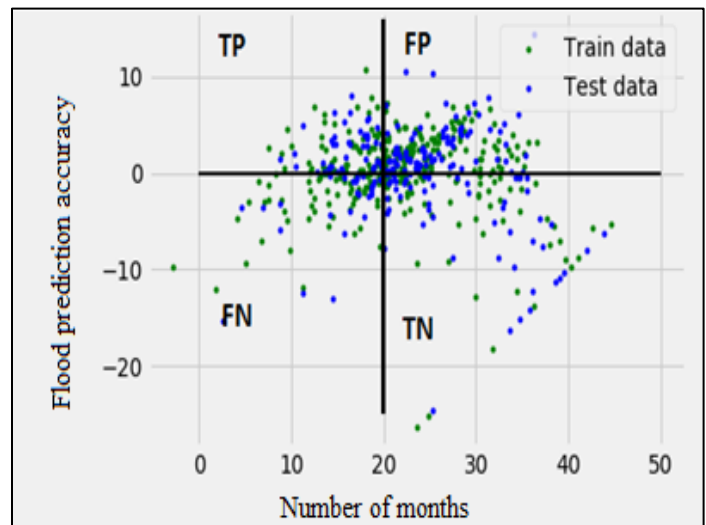

Fig. 9. Simulation Results for Flood Forecasting using Linear Regression.

The Accuracy (Classification rate) is obtained given by the following relationship:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (5)$$

The Recall is obtained given by the following relationship as mentioned as follows:

$$Recall = \frac{TP}{TP+FN} \qquad (6)$$

The Recall is obtained given by the following relationship mentioned as follows:

$$Precision = \frac{TP}{TP+FP} \qquad (7)$$

The F- measures are obtained given by the following relationship mentioned as follows:

$$F-measure = \frac{2*Recall*Precision}{Recall+Precision} \qquad (8)$$

VI. CONCLUSIONS AND FUTURE WORKS

This research work investigates the development of the Flood Forecasting Model based on linear regression and confusion Matrix. The existing schemes have thoroughly studied to identify that the flood prediction solutions are rigid against the input parameters, infrastructure, and scalability. The existing solution becomes useless if there is a change in the surrounding like a new canal, stream, or dam is included in the system. The objective of this study is to generate the damage maps of the flood-affected areas by using remote sensing and social and infrastructure data sets. The core objective in this research study focuses on Remote Sensing which is an integral application of GIS and is used to create damage maps from social and infrastructure data sets in case of floods events. The Rainfall anomaly maps had also been used for peak water level analysis and compare the results with RADARSAT images for remote sensing. Along with RADARSAT images LANDSAT images have also been used for digitized image classification and comparison for flood monitoring and mapping. The proposed flood hazard model focuses on the Feed Forward Neural based approach in predicting the occurrence of Flood

on a monthly and yearly basis. Based on the obtained results it is analyzed that accurately predicts the intensity of flood. The proposed model can also be implemented in the future via using Convolution Neural Networks by obtaining more accurate results.

REFERENCES

[1] L. Feng and G. Luo, "Practical Study on the Fuzzy Risk of Flood Disasters," Springer Journal. Acta Applicandae Mathematicae, vol. 106, no. 3, pp. 421-432, 2009.

[2] L. Alfieri, J. Thielen, and F. pappenberger, "Ensemble hydro-metrological simulation for flash flood early detection," Elsevier Journal of Hydrology, pp. 143-153, 2012.

[3] S. Tahir, S. T. Bakhsh, M. F. Abulkhair, and M. O. Alassafi, "An energy-efficient fog-to-cloud Internet of Medical Things architecture," International Journal of Distributed Sensor Networks, vol. 15, pp. 1-13, 2019.

[4] P. A.A and B. U.U, "Flood detection system using wireless sensor networks," International Journal of Advanced Research in computer science and software Engineering, vol. 5, no. 2, 2015.

[5] R. S, M. M and Y. Y, "Flash Flood Prediction using an uncelebrated hydrological model and radar rainfall data in a Mediterranean watershed under changing hydrological conditions," Journal of Hydrology, vol. 394, pp. 245-255, 2010.

[6] A. N, H. M and R. N, "Adhoc wireless sensor Network Architecture for Disaster Survivor detection," International Journal of Advance Science and Technology, vol. 34, 2011.

[7] M. Imran, S. T. Bakhsh, S. Tahir, M. Basheri, and M. Shoaib, "A reconfigurable scatternet formation and maintenance scheme with heterogeneous services for smart Bluetooth devices," Sustainable Cities and Society, vol. 40, pp. 589-599, 2018.

[8] Wang and Wen-Chuan, "A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series," Journal of hydrology, vol. 374, no. 3, pp. 294-306, 2009.

[9] G. Yasser and M. A. Serhani, "A WSN-driven service discovery technique for disaster recovery using mobile ad hoc networks," Wireless Days (WD), IFIP. IEEE, 2011.

[10] J. K. Roy, D. Gupta, and S. Goswami, "An improved flood warning system using WSN and Artificial Neural Network," India Conference (INDICON), 2012 Annual IEEE, 2012.

[11] S. Kar, A. D. Sarkar and N. Mukherjee, "An Integrated Framework in Geographic Information System using Wireless Sensor Network," International Journal of Computer Applications, 2012.

[12] Argany, Meysam, M. A. Mostafavi, F. Karimipour and C. Gagné, "A GIS based wireless sensor network coverage estimation and optimization: A Voronoi approach," In Transactions on Computational Science XIV, pp. 151-172, 2011.

[13] M. Abdullah, "Simulation of Wireless Sensor Network for Flood Monitoring System," Design, User Experience, and Usability. User Experience Design for Everyday Life Applications and Services. Springer International Publishing, pp. 255-264, 2014.

[14] E.A.Basha, S.Ravela and D.Rus, "Model based Mentoring for early warning Flood Detection," Proceedings of the 6th ACM conference on Embedded network sensor systems, vol. 5, no. 7, pp. 295-308, Nov 2008.

[15] J. Y. Z. X. H. B. Zhang Hailin, "GIS-based Risk Assessment for Regional Flood Disaster," IEEE International Conference on Environmental Science and Information Application Technology, 2010.

[16] S.Rozali, E.Morin, Y.Yair and C.Price, "Flash Flood Prediction using an uncelebrated hydrological model and radar rainfall data in a Mediterranean watershed under changing hydrological conditions," Elsevier Journal of Hydrology, vol. 394, pp. 245-255, 2010.

[17] V.Seal, A.Raha, S.Maity, S.Mitra, A.Mukrjee and M.K.Naskar, "A Real Time Multivariate robust Regression Based Flood Prediction Model using Polynomial Approximation for wireless sensor network based flood forcasting system," Springer Journal of Advances in Computer science and information Technology, vol. 86, pp. 432-441, 2012.

[18] Y. M. Xiaomeng, L. Xiaojuan, L. Y. Sun and X. Li, "Rapid Assessment of Flood Disaster Loss in Sind and Punjab Province, Pakistan Based on RS and GIS," IEEE Trans. (ICMT) International Conference on Multimedia Technology, pp. 646-649, July 2011.

[19] R. A, M. S, S. K. Mitra, J. R and M. K. Naskar, "A Low Complexity Multivariate Regression Based Flood Forecasting Model Using an Optimized WSN Deployment Scheme," Advanced Materials Research, vol. 403, pp. 3484-3494.

[20] Aziz, N. A. Ab and K. A. Aziz, "Managing disaster with wireless sensor networks," Advanced Communication Technology (ICACT), 2011 13th International Conference on. IEEE, 2011.

[21] Chen, Jian, A. A, Hill and L. D. Urbano, "A GIS-based model for urban flood inundation," Journal of Hydrology, vol. 1, no. 373, pp. 184-192, 2009.

[22] M. N.Ahmad, "Flood Prediction and Disaster Risk Analysis Using GIS based wireless sensor Networks," Journal of Basic and Applied Scientific Research, vol. 3, no. 8, pp. 632-643, 2013.

[23] " National Disaster Management Authority".

[24] X. Chen, X. Lu, Z. Liu, S. Fang, D. Jin, and L. Zeng, "A Heterogeneous High Speed Wireless Body Sensor Network Based on SC-UWB and ZIGBEE," IEEEGlobal Telecommunications Conference (GLOBECOM 2011), pp. 1-5, 2011.

[25] S. M. J. S. Samarasinghea, H. K. Nandalalb, D. P. Weliwitiyac, J. S. M. Fowzed, M. K. Hazarikad and L. Samarakoond, "Application of Remote Sensing and GIS for flood Risk Analysis:A Case study at Kalu-Ganga River SriLanka," International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science, p. 101-109, 2010.

[26] S. M. J. S. Samarasinghea, H. K. Nandalalb, D. P. Weliwitiyac, J. S. M. Fowzed, M. K. Hazarikad, and L. Samarakoond, "GIS-based Risk Assessment Model for Flood Disaster in China," IEEE Trans. 18th International Conference on Geoinformatics, vol. 18, no. 20, pp. 1-5, June 2010.

[27] S. M. J. S. Samarasinghea, H. K. Nandalalb, D. P. Weliwitiyac, J. S. M. Fowzed, M. K. Hazarikad and L. Samarakoond, "Adaptive space–time sampling with wireless sensor nodes for flood forecasting," Journal of Hydrology, vol. 414, pp. 136-147, 2012.

[28] H. L. Cloke and F. Pappenberger, "Ensemble flood forecasting: A review," Elsevier Journal of Hydrology, vol. 375, pp. 613-626, September 2009.

[29] Y. Chen, W. Shen, HongweiHuo and Y. Xu, " A Smart Gateway for Health Care System Using Wireless Sensor Network," IEEE Trans. Fourth International Conference on Sensor Technologies and Applications (SENSORCOMM), vol. 18, no. 25, p. 545 – 550, 2010.

[30] B. Biondi and D. L. D. Luca, "Performance assessment of a Bayesian Forecasting System (BFS) for real time flood forcasting," Elsevier Journal of Hydrology, vol. 479, pp. 51-63, 2013.

[31] G. M. Dawood, M. N, M. Khalid, and A. A. Ghamdi, "GIS-Based Spatial Mapping of Flash Flood Hazard in MakkahCity, Suadi Arabia," Journal of Geograhic Information System, vol. 3, pp. 225-231, 2011.

[32] S. T. Bakhsh, R. AlGhamdi, A. H. Altalhi, S. Tahir, and M. A. Sheikh, "Adaptive Sleep Efficient Hybrid Medium Access Control algorithm for next-generation wireless sensor networks," EURASIP Journal on Wireless Communications and Networking, vol. 2017, pp. 84-94, 2017.

[33] M. A. Saud, "Assesment of Flood Hazard of Jeddah Area 2009, Suadi Arabia," Journal of Water Resource and Protection, pp. 839-847, 2010.

[34] J. Zhang, L. Song, F. Feng, and H. Gong, "Hydrologic Information Extraction for Flood Disaster Risk Assessment in Pearl River Basin and Luan River Basin, China," IEEE Trans. International Conference on geoinformatics, pp. 1-4, 2010.

[35] J. Chen, A. A, Hill and L. D and Urbano, "A GIS-based model for urban flood inundation," Elsevier Journal of Hydrology, pp. 184-192, 2009.

[36] J. Liu, J. Wen, K. Yang, Z. Shang and H. Zhang, "GIS-Based Analysis of Flood Disaster Risk in LECZ of China and Population Exposure," IEEE

Trans. 19th International Conference on Geo informatics, Shanghai, China, 2011.

[37] H. Zhuowei, L. Xiaojuan, S. Yonghua, and Z. Liying, "Flood Disaster Response and Decision-making Support System Based On Remote Sensing and GIS," Flood Disaster Response and Decision-making Support System Based On Remote Sensing and GIS, pp. 2435-2438, 2007.

[38] S. Tahir, S. T. Bakhsh, R. Alghamdi, and M. Abulkhair, "Fog-based healthcare architecture for wearable body area network," Journal of Medical Imaging and Health Informatics, vol. 7, no. 6, pp. 1409-1418, 2017.

[39] N. Sulaiman, F. Husain, K. A. Hashim, and A. M. Samad, "A Study on Flood Risk Assessment for Bandar Segamat Sustainability Using Remote Sensing and GIS Approach," IEEE Control and System Graduate Research Colloquium (ICSGRC 2012), pp. 386-391, 2012.

[40] R. Jeberson, R. Raj and T. Sasipraba, "Disaster Management System based on GIS Web Services," IEEE Trans. Recent Advances in Space Technology Services and Climate Change (RSTSCC), p. 252 – 261, 2010.

[41] S. T. Bakhsh, R. A. A. AlGhamdi, A. H. Altalhi, S. Tahir, and M. A. Sheikh, "Adaptive Sleep Efficient Hybrid Medium Access Control algorithm for next-generation wireless sensor networks," EURASIP J. Wireless Comm. and Networking, vol. 2017, p. 1-15, 2017.

[42] C. Nie, H. Li, L. Yang, S. Wu, Y. Liu and Y. Liao, "Spatial and temporal changes in flooding and the affecting factors in China," Springer Journal of Natural Hazards, vol. 31, pp. 425-439, 2012.

[43] S. V, R. A, M. S, M. S. K, M. A, and N. M. K, "A simple flood forecasting scheme using wireless sensor networks," arXiv preprint arXiv:1203.2511, 2012.

Automatic Extraction of Rarely Explored Materials and Methods Sections from Research Journals using Machine Learning Techniques

Kavitha Jayaram[1], Prakash G[2]
Department of Computer Science and Engineering
Amrita School of Engineering
Amrita Vishwa Vidyapeetham, Bengaluru, India

Jayaram V[3]
Solid States and Structural Chemistry Unit
Indian Institute of Science, IISc
Bengaluru, India

Abstract—The scientific community is expanding by leaps and bounds every day owing to pioneering and path breaking scientific literature published in journals around the globe. Viewing as well as retrieving this data is a challenging task in today's fast paced world. The essence and importance of scientific research papers for the expert lies in their experimental and theoretical results along with the sanctioned research projects from the organizations. Since scant work has been done in this direction, the alternative option is to explore text mining by machine learning techniques. Myriad journals are available on material research which throws light on a gamut of materials, synthesis methods, and characterization methods used to study properties of the materials. Application of materials has many diversified areas, hence selected papers from "Journal of Material Science" where "Materials and Methods" sections contains names of the method, characterization techniques (instrumental methods), algorithms, images, etc. used in research work. The "Acknowledgment" section conveys information about authors' proximity, collaborations with organizations that are again not explored for the citation network. In the present articulated work, our attempt is to derive a means to automatically extract methods or terminologies used in characterization techniques, author, organization data from "Materials and Methods" and "Acknowledgment" sections, using machine learning techniques. Another goal of this research is to provide a data set for characterization terms, classification and an extended version of the existing citation network for material research. The complete dataset will help new researchers to select research work, find new domains and techniques to solve advanced scientific research problems.

Keywords—Data-mining; rule-based; machine-learning; term extraction; classification; materials and methods; acknowledgment

I. Introduction

Citation networks have been well analyzed both syntactically as well as structurally but there is a strong need for semantic analysis for these networks. Citation analysis as the most significant area of bibliometric that has been studied using the Page Ranking algorithm for a long time and there has been a great deal of research work in this direction [1]. Citation sentiment analysis is used to determine sentiment polarity of clinical trial papers using n-gram and sentiment lexicon features on annotated corpus [2]. A summary of a corpus of research papers, domain-independent structural relations between abstracts and domain of scholarly medical articles,

state-of-the-art deep learning baseline was constructed and has been reported [3]. Given a particular paper of interest, CiteSeer can display the context of how the paper is cited or indexed in subsequent publications with a summary of the paper in electronic format [4]. The semantic analysis of paper abstracts is a good start for annotating papers using Natural Language Processing (NLP) with semantic metadata and for increasing the general representation and visualization of the key concepts within a given domain [5]. Here they discuss and analyze the text mining techniques and their applications in diverse fields [6]. The collaboration of productive authors based on the topics, collaborative effort, highly cited articles, etc. would identify the relationship between two specific nodes that can reveal scholarly communication patterns (i.e., collaboration or knowledge diffusion, copyright transfer) with finer granularity [7]. The author proposes a mathematical model that matches empirical acknowledgment data closely for citation patterns which give cognitive interdependence among disciplines [8]. A function of appreciation using the acknowledgment section within academia of the instrumental and normative significance has been presented [9]. A content-based image retrieval system that extracts, image features from journal papers using a supervised learning algorithm has been explained in [10]. An overview of the principles and methods of automatic term recognition of significant elements have been presented [11]. From conference proceedings and journal papers information extracted like dataset, content, and basis of extraction summarized in Appendix I. Scientometrics researchers use structural/syntactic information from a bibliographic network for qualitative analysis of the same network [12]. Some common aspects like the dataset used, methods, the most focused problem in a particular field, frequently used algorithms, hot areas such as analysis of research trends have been extracted [13]. Fig. 1 represents the existing citation network focuses on citation count and co-authorship hence it mainly contains four nodes namely Venue, Paper, Term, and Author/co-authors [12].

The "Abstract" section contains the best ratio of keywords per total of words, which contains research, material, methods used and challenges faced, but many times they do not include methods. Hence, the next most important findings of the research are expressed in "Materials and Methods" section such as experimental techniques, instrumental methods, algorithm, figures, etc. Acknowledgments section is used to express

appreciation between researchers, direct or indirect collaborators, and the contribution of external people or organizations. These aspects of the citation network are important and are needed to be explored to improve author proximity, affiliations, and funding organizations that contribute to academic or industrial research. It is found that the automatic identification of methods and acknowledgment influences the citation network. The modified citation network where few other nodes like "organization" have been included to study the author collaboration, method and dataset nodes from Materials and Methods section to analyze compounds or materials is shown in Fig. 2. It is clear that extracting the above mentioned important information from the "Journal of Material Science" and incorporating it into the existing citation network give us new ways to look at the authors' communities, collaborators from organizations and institutions in material research. This paper describes the automatic extraction of materials, characterization techniques, instrument-related terminologies, acknowledged by authors (organization) using Machine Learning (ML) techniques.

This paper is organized as follows: Section 2 covers the implementation of the algorithm, tools, framework, and work executed in the present research. Section 3 explains the experiments, results, and discussion. Section 4 summarizes the present work and future research which can be laid upon the work. In the last Section, we acknowledge the research collaborators. Appendix I include a list of reference papers where the information is extracted from the present published research papers. Appendix II gives lists of sample research papers from the research journals with title, materials, and characterization techniques. Appendix III gives a list of characterization methods used to investigate the results presented in the materials and methods section of the journal publications.
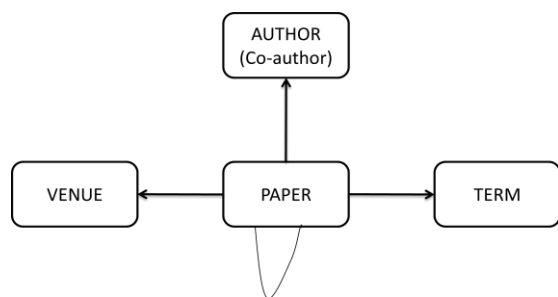

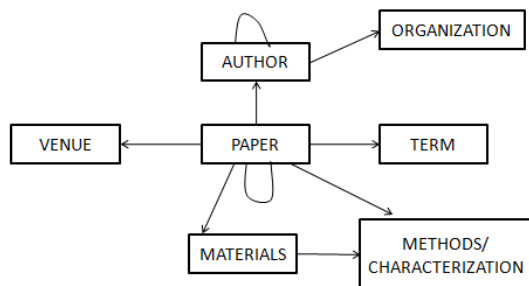Fig. 1.   Current Citation Network Analysis.

Fig. 2.   Modified Citation Network for Materials and Methods from Scientific Journals.

II.   IMPLEMENTATION

The building of a heterogeneous network includes different types of entities and incorporating into the currently existing citation network. The main work comprises finding out the entity mentioned (characterization techniques and organization names) from the research work. The implementation details mainly focuses on the "Materials and Methods" section which includes materials, methods, figures, micrographs, images, and material characterization obtained from different instrumental methods. To analyze "Material and Methods" section, it is required to convert the extracted information into a format compatible with usual heterogeneous citation networks. New node types such as "Algorithm, technique, method", "characterization, measurements, instruments", "Images or figures" and "Organization funding" form the key semantic components of the research work. To the best of our knowledge, no such work has been done with "Materials and Methods" and "Acknowledgment" sections from material research journals using machine learning techniques.

The statistical approach generally uses information such as term frequency, term-document frequency, inverse term-document frequency, etc. for extracting the important entities and mentioned phrases. Named entity recognition (NER) is also a problem that attempts to find out mention, author, organization, place, etc. Although their extraction is very good, it is limited to particular classes and does not have any model to mention terminologies and acronym. Although a lot of work has been done on domain-specific term extraction and named entity extraction for particular classes, the method keywords extraction has not been explored. Both rule-based and ML approaches to find methods were mentioned in a scientific research document to extract important techniques and methods used in biomedical research [14].

Scientific documents are mostly available in PDF format, which is semi-structured and not tagged, unlike HTML, also 'text' in them is usually arranged in multiple rows and columns. Many tools are available to extract text from PDFs, but when documents come with multiple rows and columns like tables, figures, etc., text extraction tool is not good enough[11]. A rule-based approach that is leveraged to extract the required sections is proposed using regular expressions and was reported in [15]. Single-word does not represent an entity, but a sequence of words does, support vector machine (SVM), linguistic-based techniques for entity extraction generally uses part-of-speech (POS) tagging and the dependencies of the words upon each other [16]. The ML algorithms used are Naïve Bayes' classifier, decision tree, and maximum entropy classifier. Extraction of a vast number of terminologies and acronyms from the "Materials and Methods" section is not an easy task. New methods and techniques are being used and named with new emerging problems. Using PDFBox and TET tools, the extraction of spatial co-ordinates and formatting information of text has been completed. In the present research work, automatic extraction of entities like text, single nouns and compound nouns has been carried our using a machine learning approach instead of linguistic methods.

---

[1] http://www.pdflib.com/products/tet/

Primarily the 'text' has been extracted using PDFBox text extraction tool, and then the co-ordinates of words and lines in the documents were calculated. This helps in calculating the coordinates of the line where the section name is extracted using regular expression, starting from one section to the next section, using a regular expression. Further "Materials and Methods" and "Acknowledgment" sections were also extracted individually from PDF into a text format using a section extraction algorithm as presented in Fig. 3. After extracting the required sections the terminologies like names of the materials, characterization techniques, methods, authors, organization, etc. are extracted from the text file. Terminology and acronym for materials and characterization techniques from the sample journals are listed in Appendix II.

```
Open PDF document
While lines: start to end do
if line_text == secName then
    set co-ordinates and page; break
end
end
while lines : sec_pagedo
if line_coord>= sec_coordthen
extract lines
end
if (line_txt == secName) and  (line_coord>=
sec_coordOR
line_page>sec_page) then
    break;
end
end
```

Fig. 3.   Section Extraction Algorithm [12].

The following two categories were considered for extracting acronym:

Category 1: Methods ending with keywords (such as analysis or scope) eg.: Energy Dispersive X-Ray Analysis/spectroscopy (EDX or EDS).

Category 2: Methods do not have any keywords. eg.: X-ray Diffraction (XRD).

New dataset was created by selecting data from nearly 800 research papers, where it contains methods and characterization techniques. The method mentions in the dataset representing the characterization techniques (or Instrumental methods), algorithm, theory, model are considered in the form of nouns. In category 1, methods are extracted using regular expressions to create the training dataset. Hence POS (Part of Speech) was used for tagging to extract names of all methods, though they fall into any of these two categories. When data falls into category 2, supervised machine learning algorithms have been used, for which a good quantity of training dataset is required. All the relevant characterization techniques and abbreviations are listed in Appendix III. Different classification algorithms were run over datasets and evaluated by precision, recall, and F1-score techniques using the following formulas [17]:

$$precison$$
$$= \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

$$recall$$
$$= \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

$$Precision = \frac{t_p}{t_p + f_p}, \qquad Recall = \frac{t_p}{t_p + f_n}$$

$$F = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Balance\ accuracy = \frac{TPR + TNR}{2}$$

$$TPR = \frac{tp}{tp + fn} \ and \ TNR = \frac{tp}{tp + fp}$$

Whereas for classification, the following terms are used to compare the results of the classifier: the term t_p is true positives, t_n is true negatives, f_p is false positives, f_n is false negative, further, TPR is term positive rate and TNR is term negative rate. Precision is the fraction of relevant instances among the retrieved instances, while recall fraction of the total amount of relevant instances that were actually retrieved. F1-score is the harmonic mean of precision and recall. TPR and TNR are statistical classification for a confusion matrix or error matrix. The terms positive and negative refer to the classifier's prediction (expectation) and true and false terms refer to the prediction corresponds to the external judgment (observation) [18].

III. RESULTS AND EXPERIMENTS

Present experiments were performed on system configuration having 128 GB RAM by using Python 3.0 with nltk and also Java as the programming language.

Data pre-processing was performed before collecting training data, such as removing all stop words, commas, semicolons, newlines (which were unnecessarily present because the data was extracted from pdfs). The papers were downloaded from an official website of "Journal of Material Science (JMS)". The text contained in the documents was extracted using PDFBox tool. Even this tool is not found to be very promising in retaining the structure of the extracted text. Since the data is in PDF format, it is a difficult task to use all the information available in the research documents. Therefore, data pre-processing becomes an important and time consuming task. Spatial coordinates of the words to form the lines and to keep the lines in correct order are also an important task. After working on many methods, good results were achieved by a supervised classification method approach. Summary of noun phrases from journal papers and Wikipedia entries term sequence are listed in Appendix III.

Both "Materials and Methods" and "Acknowledgment" sections are derived using the regular expression based rules. Previous work on section extractions shows that regular expressions achieve 100% precision and 67% recall for extracting Acknowledgment section [19]. The proposed analysis shows that same approach works for the Materials and

Methods section too. Hence regular expressions and spatial co-ordinates are used to extract both sections of the research paper. NLP technique is used to extract sentences having the materials name, algorithm methods or characterization, measurement, etc. words. StandfordCoreNLP tool is used for named entity recognition [20]. Using these entities a list of the most widely used methods or simulation work done in material research are listed in Appendix III. Named Entity Recognition (NER) is used to extract sentences from different papers to find out methods, characterization, algorithms, people / authors, and organizations.

Noun phrases available in the research paper are searched from the Wikipedia entries. The summaries of the term sequences were collected while rejecting the sequences that were not available in Wikipedia entry. Along with these entries documents were clustered into five classes using Linear Discriminant Analysis (LDA), the list of methods predicted in a paper (w.r.t. materials used for research) is shown in Fig. 4(a). Using nearest neighbor and supervised methods the corresponding classes were assigned to dominant topics. With the important extracted information, the Citation Network is extended to provide dataset related to collaborators and authors due to newly introduced nodes in the network. The results obtained also include a new dataset for characterization techniques from the research paper.

The main goal is to extract the characterization or methods from the research papers. Initially, about 100 term sequences from various research papers were manually tagged as methods (characterization methods). These 100 terms were extracted from research papers and Wikipedia entries terms using rule-based regular expressions techniques based on machine learning and NLP methodologies. Subsequently all the relevant stop words, commas, semicolons, newlines (which are unnecessarily present because of the data extracted from pdf's) were removed from the extracted text. Although many of the problems that arose owing to the pdf's extraction were addressed, few problems remain unsolved. Few problems like, unnecessary spacing between few words, some non-ASCII characters, and distortion of table data are attributed as the primary reasons for messing up in text data. These mistakes could have a detrimental effect on the output. The features extracted from the text are automatically run by the program where the positive and negative class term sequences are encountered.

The process of searching term sequences in the whole document set for creating the training data manually while considering positive and negative term sequences with a ratio of 5:3, but the ratio obtained was about 1:9. This is the class imbalance problem that occurred due to the specificity of the positive term sequences and all the general noun phrases coming into negative class. This class imbalance problem was resolved by applying entity clustering for sampling negative class instances, where the ratio was about 6:4. Fig. 4(a) shows the list of materials classified as carbon, graphite, silica, electronic, and high-temperature materials (HTC) along with a sub-classified list of few compounds selected from the Journals. Fig. 4(b) shows the list of characterization techniques from different instrumental methods like XRD, SEM, TEM, etc. including few simulation and ML done on the selected

materials. Summary of the characterization methods used to analyze the material selected from the "Materials and Methods" section of the Journal are listed in Table I. All these classifications are considered while solving problems using machine learning techniques.

The materials and characterization techniques extracted from the Journal gives the following conclusion. The bar graph in Fig. 5(a) shows that TiO_2, Graphene materials appear in more research papers compared to other compounds. However, in Fig. 5(b) shows XRD and SEM instruments used extensively as characterization techniques for material analysis. Our results from the machine learning techniques reveals the statistics of materials not explored by the researchers and the type of methods not used for characterization of materials including simulation work. The present research work provides a dataset for materials and methods for selecting particular area of research by the scientific community.
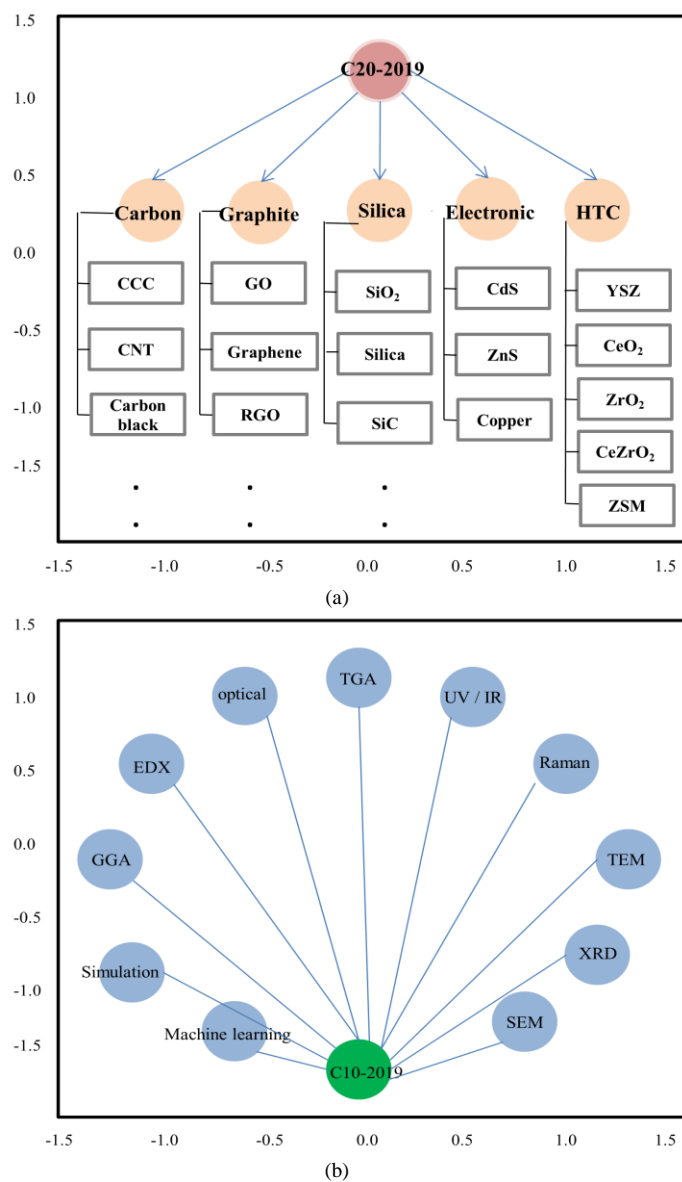

(a)

(b)

Fig. 4.   Dataset Selected for the Classification from the Journal of Material Science (a) Materials (b) Methods.

Sl no.	Material	Characterization techniques
1	Graphene	TEM, HRTEM, EELS, ITC, XPS, ES, Raman, UV-Vis, DSC, TGA, Vickers hardness indenter, Zeta potential, Material Studio (MD simulation)
2	SiO_2	SEM, TEM, Raman, XRD, XPS, Raman, Simulation
3	TiO_2	TEM, SEM, TIFR, Raman, XRD, XPS
4	YSZ	CALPHAD, TA, SEM, EDX, XRD, XPS,
5	Carbon black	XRD, SEM, TEM, Raman, IR camera, UV-Vis-NIR
6	Copper	Raman, SEM, UV-Vis, FESEM, XRD, BET, TEM, EIS, MD Algorithm
7	Organic Polymer	NMR spectra, FTIR, TGA, XPS, XRD, SEM, BET, UV-Vis, Fluorescence spectra
8	SiC	SEM, Raman, SRIM simulation
9	Nanofibers	SEM, XRD, TEM, BET, TGA, Raman
10	Ceramics	Gibbs energy, DSC, TGA, XRD, SEM, RADIANT

added in the training dataset. Top 14 organizations which are acknowledged in Material Research Journal were analyzed for research publication. An analysis of the acknowledgment section along with organization and country names extracted from the Journal of Material Science for the past three years is shown in Fig. 6(a). Few selected funding agency are listed in Table II. According to the graph, NNSFC (National Natural Science Foundations of China) is the most acknowledged Chinese organization, involved in funding the most research projects. The analysis shows China published most research papers followed by the USA and other countries as shown in Fig. 6(b).

In summary, the results show that China published more research papers, and NNSFC funded the maximum project in past three years. The comparison shows the number of the research paper published by different countries in the past three years. This period can be extended to more number of years to validate our machine learning techniques. Once the author and organization parameters are extracted, built a social network of the acknowledgment section, and the snapshot of the social network is shown in Fig. 7.
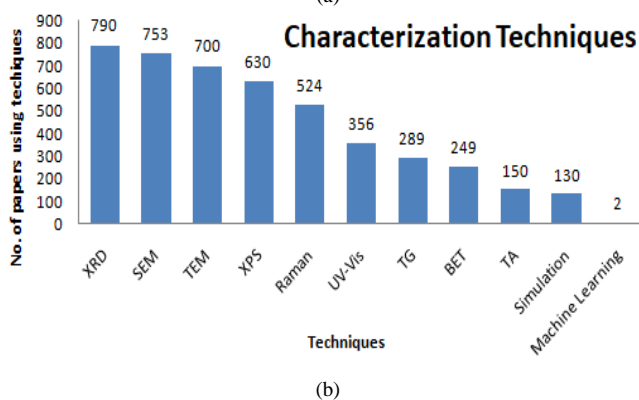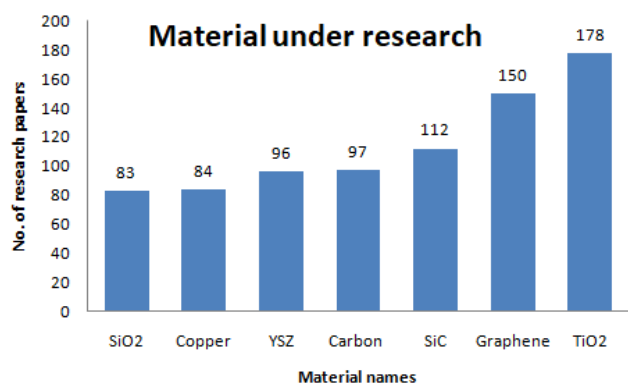

(a)

(b)

Fig. 5.   Data Extracted from Journal of Material Science Papers (a) Top 7 Materials, (b) Top 10 Characterization Techniques.

The clustering is a frequency, which contains names, short names, and abbreviations using different alternatives for some well-known organization in acknowledgment section are

(a)
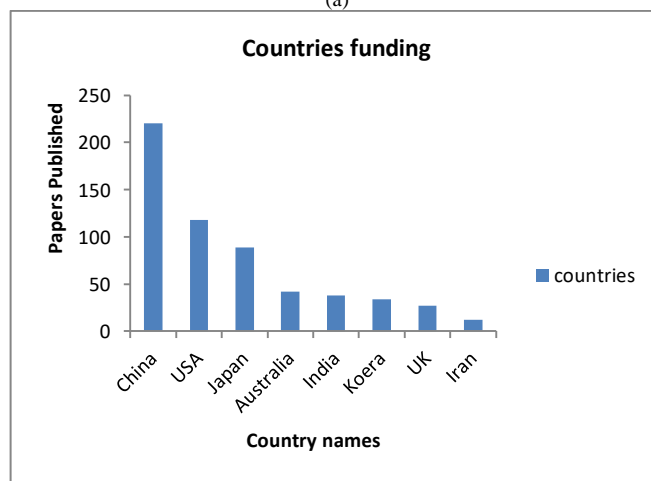
(b)

Fig. 6.   Projects Funded from different Countries are Selected from Acknowledgment Section, (a) Bar Graph of different Organization Funded the Project from Past 3 Years Vs Number of Projects, (b) Bar Graph of Project Funding Countries Vs Number of Research Paper Published.

TABLE II.   LIST OF FUNDING AGENCY

| Sl no.	Funding agency
1	National Natural Science Foundation
2	Yong Teachers Scientific Research
3	College and University Key Project
4	National Funds for Distinguished Young Scientists
5	Research Fund of University
6	United Innovation
7	National Research Agency
8	Jiangxi Scientific / Education Fund
9	Innovative Research Team in University
10	China Scholarship Council
11	Fundamental Research Funds for Central University
12	National Key Research and Development Program
13	Japan Society for the Promotion of Science
14	National Science and Technology cooperation Funds
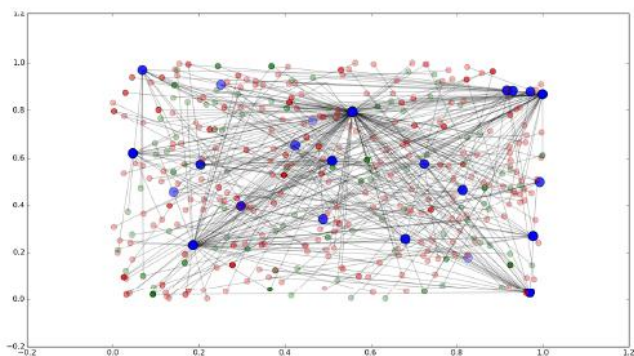
Fig. 7.   Social Network of Acknowledgment Section Processed for 100 Published Research Papers. Blue = Papers, Red = Organization, Green = Person.

Novelty and evaluation is a very important part of research work. Precision and recall are two extremely important model for evaluation metrics. While precision refers to the relevant percentage of results, recall refers to the percentage of total relevant results correctly classified by the algorithm. F-1 score is the harmonic mean of precision and recall. Both precision and recall are important to solve problems; one can select a model that maximizes the F1 score. To check the correctness of the predicted method terms, 20 research documents from the same journal were selected at random. Manually extracted methods and characterization techniques were used from the "Materials and Methods" sections and the results were compared with classification algorithms [21, 22]. Precision, Recall and F1-scores are different classification algorithms used to predict characterization techniques and organization names. Also we have computed dataset using LDA, NBS and LIBLINEAR to evaluate classification algorithms. LDA (Latent Dirichlet Allocation) is a generative probabilistic model for collections of discrete data such as text corpora [23]. NBC (Neighborhood Based Clustering) discovers clusters based on the neighborhood characteristics of data [24].

LIBLINEAR (Library for Large Linear Classification) uses logistic regression and linear support vector machines which is very efficient on large sparse datasets [25]. The values listed in Table III are the measured dataset from the classification algorithms; which concludes **LDA**, NBC and LIBLINEAR (SVM) and gives better F1 scores. Overall results show the novelty of our research work which generates and establishes tagged training dataset extracted from the materials and methods section using ML technique, which supports researchers to select advanced research topics.

TABLE III.   EVALUATED DATASET MATRIX OF THE CLASSIFICATION ALGORITHMS

Method list	Recall	Precision	F1-Score
NBC	0.68	0.32	0.41
LIBLINEAR	0.68	0.30	0.41
Decision Tree	0.83	0.24	0.37
MaxEnt	0.72	0.26	0.38
Random	0.48	0.16	0.24
LDA	**0.54**	**0.34**	**0.41**

IV. CONCLUSION

Our analysis shows there are plenty of hidden information in each section of research journal papers. The extracted information can be used to extend the currently existing Citation Network. "Materials and Methods" and "Acknowledgments" are the least explored aspects of Scientometrics of Material Science Research papers. The methods and characterization from the "Materials and Methods" section, people and organizations acknowledged from the "Acknowledgments" section were extracted from "Journal of Material Science" and revealed important insight. A new researcher or a beginner can get an idea of material as well as the characterization methods used for completion of the research work. They can also understand which material, techniques are least explored for new research domains to proceed. Gives adequate information about the ongoing research problems, researchers are interested to find out the country, collaborators, and to propose new joint research project form different funding agencies. Future work involves extracting the "Abstract" and "Results" sections from scientific research journals. These two sections helps in summarizing the classification of completed research work, figures can be classified according to image quality and instrumental methods used for characterization of the materials. The complete dataset will help new researchers to select research work, find new domains and techniques to solve advanced scientific research problems.

REFERENCES

[1] Elliot J Yates and Louise C Dixon, "PageRank as a method to rank biomedical literature by importance", Source Code of Biology and Medicine, vol. 10, pp. 16-25, 2015.

[2] Jun Xu, Yaoyun Zhang, Yonghui Wu, Jinggi Wang M S, Xiao Dong M D and Hua Xu, "Citation sentiment analysis in clinical trial papers", AMIA Annual Symposium Proceedings Archive, pp. 1334-1341, November 2015.

[3] Arthur Brack, Jennifer D Souza, Anett Hoppe, Soren Auer and Ralph Ewerth, "Domain-independent extraction of scientific concepts from research articles", Springer Nature Switzerland AG, pp. 251-266, 2020.

[4] C Lee Giles, Kurt D Ballacker and Steve Lawrence, "Citeseer: An automatic citation indexing system", Proceedings of the third ACM conference on Digital libraries, New York, pp. 89-98, 1998.

[5] Ionut Cristian Paraschiv, Mihai Dascalu, Stefan Trausan-Matu, Philippe Dessus, "Analyzing the semantic relatedness of paper abstract", 20th International Conference on Control System and Science, pp. 759-764, 2015 [IEEE].

[6] Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha and Fakeeha Fatima, "Text mining: techniques, applications and issues", International Journal of Advanced Computer Science and Applications,Vol. 7, No. 11, pp. 414-418, 2016.

[7] M E J Newman, "Coauthorship networks and patterns of scientific collaboration, colloquium", The National Academy of Sciences of the USA, Vol. 101, pp. 5200-5205, 2004 [PNAS].

[8] Charles H Davis and Blaise Cronin, "Brief communication acknowledgments and intellectual indebtedness: A bibliometric conjecture", Journal of the American Society for Information Science, vol. 44, no..10, pp. 590-592, 1993.

[9] Blaise Cronin, "Acknowlegdement trends in the research literature of informantion science", Journal of Documentation, 2001, Vol. 57, No. 3, pp. 427-433, 2001.

[10] B Akshaya, S S Sruthi Sri, A Niranjana Sathish, K Shobika, R Karthika and Latha Parameswaran, "Content-based image retrieval using hybrid feature extraction techniques", Springer Nature Switzerland AG, pp. 583-593, 2019 [ISMAC-CVB].

[11] Bin Umino, "Methods of automatic term recognition – A review", Proceedings of COLING, Mumbai, pp. 1211-1222, 2012 [Technical Papers].

[12] Madian Khabsa, Pucktada Treeratpituk, and C Lee Giles, "Ackseer: a repository and search engine for automatically extracted acknowledgements from digital libraries", Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, USA, pp. 185-194, 2012.

[13] Hospice Houngbo and Robert E Mercer, "Method mention extraction from scientific research papers", Proceedings of COLING, pp. 1211-1222, 2012.

[14] Bei Yu, "Automated Citation Sentiment Analysis: What can we learn from biomedical researchers", ASIS&T Proceedings, vol. 50, no. 1, pp. 1-9, 2013.

[15] Xiaoyu Tang, Qingtian Zeng, Tingting Cui and Zeze Wu, "Regular expression-based reference metadata extraction from the Web", 2nd symposium on web society, pp. 346-350, 2010 [IEEE].

[16] Didier Bourigault, "Surface grammatical analysis for the extraction of terminological noun phrases", Proceedings of the COLING, pp. 977-981, August 1992 [NANTES].

[17] David M W Powers, "Evaluation: From precision, recall and F-factor to ROC", Informedness, Markedness & Correlation, Technical report SIE-07-001, pp. 1-24, 2007.

[18] C Lee Giles and Isaac G Councill, "Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing", The National Academy of Sciences of the USA, vol.101, no. 51, pp. 17599-17604, 2004.

[19] Christopher D Mining, Mihai Surdeanu, Sohn Baner, Jenny Finkel, Steven J Bethard and David McClosky, "The Standford corenlp natural language processing toolkit", Proceedings of 52nd annual Meeting of Association for Computational Linguistics, USA, pp. 55-60, June 2014.

[20] Kamal Sarkar, Mita Nasipuri and Suranjan Ghose, "Machine learning based keyphrase extraction: comparing decision trees, Navie Bayes and artifical neural networks", J Inf Process Syst, India, vol. 8, no. 4, pp. 693-712, December 2012.

[21] Kavitha Jayaram, Sangeeta , "A review: Information extraction techniques from research papers", ICIMIA, India, pp. 56-59, February 2017 [IEEE].

[22] David M Blei, Andrew Y Ng and Michael Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research,vol. 3, pp. 993-1022, 2003.

[23] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang and Chih-Jen Lin, "LIBLINEAR: a library for large linear classfication", Journal of Machine Learning Research, 2008, 9: 1871-1874.

[24] Shuigeng Zhou, Yue Zhao, Jihong Guan and Joshua Huang, "A neighborhood-based clustering algorithm", Springer-Verlag Berlin Heidelberg, pp. 361-371, 2005 [LNAI 3518].

APPENDIX I. LIST OF EXTRACTED INFORMATION FROM SELECTED JOURNAL PAPERS

Sl no	Paper title	Dataset	Extracted Content	Basis of Extraction
1	IE from Biomedical Literature: Methodology, Evaluation and an Application (2003)	Biomedical dataset	Biological terms in doc, identify the common concepts in group of genes	Dictionary, clustering of genes,
2	Automatic extraction of titles from general documents using machine learning (ACM 2005)	Internet of Microsoft, DotGov, DotCom	Titles	Improving of search ranking results in doc retrieval by extracted titles
3	Mining knowledge from text using information extraction (2005)	Book amazon.com	Abstract knowledge, concrete data	Knowledge, patterns
4	Extracting procedures from text (2007)	Public recipe web site	Procedures and graphs	Building large knowledge
5	Automatic extraction and processing of document references (2007)	CRF-based approach	References , names of related documents	Search these doc in sys DB, create links to respective documents
6	Opinion holder extraction from author and authority viewpoints (2007)	MPQA corpus	Named entity extraction	Opinion of author and authority viewpoint
7	Entity categorization over large documents collections (ACM KDD, 2008)	Web-data	Entities (people, movies, painter, writer)	Categorizing extracted entities
8	Corpus study of kidney-related experimental data in scientific papers (2009)	Quantitative kidney Database (QKDB)	Experimental data	Automate extraction of experiment and result section
9	Automatic creation of a technical trend map from research papers and patents (ACM 2010)	NTCIR-8 patent mining task (Japanese data)	Technology (algo, tools, materials, data)	Creating Technical trend map (recall and precision)

10	Link Analysis in mind maps: a new approach to determining doc relatedness (2010)	Maps	Reference	Doc recommender if doc A and B are refereed by mind map then related
11	Contextual Information Extraction in Research Articles: A case of developing contextual RDF data for ESWC papers (ACM 2011)	European Semantic Web Conference (ESWC)	Paragraphs containing citations and classified sentences	Automatic context identification. Author work, cited work used by various authors, searching citation sentences, export data in different formats
12	Citation analysis and keyword mining based on full text extraction of scientific literature (ACM 2012)	ACM digital library	Citation, bibliometric analysis (domain context graph)	Domain knowledge graph and analyzing interrelation of publications for research direction
13	A comparison of metadata extraction techniques for crowdsourced bibliographic metadata management (2012)	e-prints, Mendeley dataset	Authors	Conditional Random Fields and SVM
14	Machine Learning based keyphrase extraction: comparing decision trees, naïve bayes, and artificial neural network (2012)	Downloaded from websites of journals Springer, Elsevier, (economics, law, medical)	Comparing different keyphrase extraction	Comparison of methods
15	ARTIC: metadata extraction from scientific papers using a two-layer CRF model (2014)	100 scientific papers from IEEE, Elsevier, Springer and ACM	Title, author names, emails, affiliations and venue info	Conditional Random Fields to extract metadata from scientific papers
16	Recommendation of newly published research papers using belief propagation (2014)	DBLP dataset	Citation information	Recommend most interesting newly published paper
17	Scientific monitoring by mining scientific papers (2014)	PDF and HTML	Semantic annotation	Relation between organizations or topics
18	Automatic extraction of main thesis documents fields using decision trees (2015 International conference on Computational Science and Computational Intelligence)	Downloaded thesis, 65 documents (converted to word format)	Title, abstract, authors	Facilitate solving the research problem process. Structuring thesis document to help research to access knowledge easily
19	Competing Algorithm Detection from Research Papers (ACM 2016)	DBLP (small dataset)	Algorithm names (name entity extraction)	Competing algorithm (ranking based on number of comparisons
20	Insights from mining eleven years of scholarly paper publications in requirements engineering (RE) series of conferences (2016)	551 papers from Requirements Engineering (RE)	Topics frequently co-occurring and connected terms, co-author	University-industry collaborations, internal and external collaborations
21	PDFFigures 2.0: mining figures from research papers (2016) (more fig more citations)	Introduce a new dataset of comp science papers	Figures and tables, captions	Component analysis, online search engine, correlates figure citation.
22	Extracting code segments and their descriptions from research articles (2017 IEEE/ACM international conference on mining s/w repositories (MSR)	IEEE digital libraries	Code example	Functionality and properties

APPENDIX II. LIST OF MATERIALS AND CHARACTERIZATION TECHNIQUES SELECTED FROM "MATERIALS AND METHODS" SECTION FROM THE SAMPLE JOURNALS

Sl. No.	Title	Material	Characterization techniques
1	Evolution of phase during heating of metastable beta titanium alloy Ti-15Mo	Ti	TEM
2	Mesoscale simulations of shockwave energy dissipation via chemical reaction	Polymer	MD Simulation (ChemDID), shock tube
3	Oxygen ion mobility and conductivity prediction in cubic yttria-stabilized zirconia single crystals	Yttria-stabilized zirconia (YSZ)	CALPHAD
4	Nitrogen-doped porous carbon derived from imidazole-functionalized polyhedral oligomeric silsesquioxane	3-Chioropropyltrimethoxysilane	NMR, FTIR, TEM, SEM, XRD, XPS and Raman spectrometer
5	Ultrahigh strength in nanocrystalline materials under shock loading	Copper	Shockwave, simulation, TEM
6	Facile preparation of Mg-doped graphitic carbon nitride composites as a solid base catalyst for Knoevenagel condensations	Mg-doped g-C_3N_4	XRD, SEM, TEM, XPS
7	Rational design of CuO nanostructures grown on carbon fiber fabrics with enhanced electrochemical performance for flexible supercapacitor	CuO	FESEM, XRD, BET, EIS
8	Magnetoresistance of graphite intercalated with cobalt	Pyrolytic graphite	XRD
9	Non-catalytic behavior of SiO_2 fine powders in presence of strong shockwaves for aerospace application	SiO_2	Material Shock Tube (MST), XRD, SEM, TEM, HRTEM, XPS

10	Thickness dependence of photoresponsive properties at $SrTiO_3$-based oxide heterointerfaces under different strains	$SrTiO_3$-based oxide	Atomic force microscopy (AFM), X-ray reflectivity (XRR)
11	Response of microstructure to annealing in-situ Cu-Nb microcomposite	Nb with Cu-Nb microcomposite	SEM, TEM
12	Poly (ε-caprolactone)/cellulose nanocrystal nanocomposite mechanical reinforcement and morphology: the role of nanocrystal pre-dispersion	Cellulose nanocrystal (CNC)	Young's modulus, TEM
13	Transparent heat insulation coatings with high selective shielding ability designed with novel superstructures of copper sulfide nanoplates	CuS superstructures	SEM, EDS, FESEM, XRD, FTIR, UV-Vis
14	The effect of applied voltage on the corrosion resistance of MgO-C refractories	MgO-C	Composition using EDS from SEM facility
15	Experimental investigation on response and failure modes of 2D and 3D woven composites under low velocity impact	Woven composites	Olympus stereo microscope for damage impact, C-scan
16	Aging response on the stress corrosion cracking behavior of wrought precipitation-hardened magnesium alloy	Magnesium alloy	TEM
17	Flexible tuning of hole-based localized surface plasmon resonance in roxbyite Cu1.8S nanodisks via particle size, carrier density and plasmon coupling	Nanocrystals (NCs)	TEM, SEM, UV-Vis, XRD
18	Ab-initio calculations of $CaZrO_3$ (011) surfaces: systematic trends in polar (011) surface calculations of ABO_3 perovskites	Polar $CaZrO_3$	Ab-initio calculation (Simulation)
19	Tensile testing of aged flexible unidirectional composite laminates for body armor	Composite laminates	SEM
20	On the binary Sb-Sn system: ab-initio calculation and thermodynamic remodeling	Alloy of Sb-Sn	Ab-initio calculation VESTA software

APPENDIX III.   LIST OF ABBREVIATIONS, METHODS AND MEASUREMENT TECHNIQUES

Sl. no.	Abbreviations	Methods	Measurements
1	AFM / SFM	Atomic Force Microscopy / Scanning Force Microscopy	Very-high resolution type of scanning probe microscopy, force measurement, topographic imaging and manipulation
2	BET	Brunauer Emmett Teller	The theory aims to explain the physical adsorption of gas molecules on a solid surface and to measure porosity and surface specific area of nano materials.
3	CALPHAD	Theoretical method	A CALPHAD thermodynamic database allows the calculation of the equilibrium state of "real" engineering materials.
4	CT	Computed tomography	It enables a three-dimensional representation of the internal and external structure of objects with a detailed detect-ability which goes down into the micrometer range.
5	DTA	Differential thermal analysis	The material under study in an inert atmosphere is made to undergo identical thermal cycles while recording any temperature difference between sample and reference.
6	DFT (PBE-DFT)	Density Function Theory (Perdew-Burke-Ernzerh of DFT)	Computational quantum mechanical modeling method used in physics, chemistry, and materials science to investigate the electronic structure (or nuclear structure) (principally the ground state) of many-body systems, in particular atoms, molecules, and the condensed phases.
7	EA	Electrochemical analyzer	It provides trace metal analysis, trace organic analysis, computer-controlled cyclic voltammeter, and chronoamperometry techniques.
8	EDX or EDS	Energy Dispersive X-Ray (EDX) Energy Dispersive Spectroscopy (EDS)	Chemical microanalysis technique used for elemental analysis in conjunction with SEM.
9	EELS	Electron Energy Loss Spectroscopy	Material is exposed to a beam of electrons with a known kinetic energies. Some of the electrons will undergo inelastic scattering, which means that they lose energy and provides information on unoccupied energy level.
10	EIS	Electrochemical Impedance Spectroscopy	Study of doped spinal manganese cathode oxide materials synthesized for Li-ion batteries.
11	ELS Zeta potentiometer	Electrophoretic Light Scattering	In contrast, streaming potential measurements, no movement of the liquid is generated, but the movement of the particles is used to measure suspended particle size in fluids.
12	EPR / ESR	Electron Paramagnetic Resonance / Electron Spin Resonance	Spectroscopic technique that detects species that have unpaired electrons, a surprisingly large number of materials have unpaired electrons.
13	ES / OES	Emission Spectrometer / Optical Emission Spectroscopy	A rapid method for determining the elemental composition of a variety of metals and alloys. Chemical analysis labs are equipped to evaluate the properties of the material.

14	FESEM / SEM	Field Emission Scanning Electron Microscope	An advanced microscope offering increased magnification and the ability to observe very fine features at a lower voltage than the SEM.
15	FTIR	Fourier Transform Infrared Spectroscopy	An analytical technique used to identify organic (and in some cases inorganic) materials. The technique is used to obtain an infrared spectrum of absorption and emission spectra of solid, liquid, and gas.
16	Gibbs free energy	Calculated	Calculates the Thermodynamic potential of the material.
17	HRTEM /TEM	High Resolution Transmission Electron Microscopy/ Transmission Electron Microscopy	High-resolution TEM offers resolution down to the Angstrom level and gives information on the atomic packing, rather than just the morphology. Particle growth can also be studied using TEM.
18	Hybrid rheometer		Accurate measure of frequencies, material types, and experimental designs.
19	Laser diffraction particle size		Light scattering method for particle size analysis of covering a wide range from submicron to millimeter scale.
20	MST / ST	Material Shock Tube/Shock Tube	It is a device consisting of driver and driven sections separated from a metal diaphragm, used to accelerate the test gas in supersonic and hypersonic speed, upon stopping it produces high temperature and pressure used to interact with materials at the end of the shock tube
21	NMR	Nuclear magnetic resonance (NMR) spectroscopy	Used to determine the structure of organic molecules in solution and study molecular physics, crystals as well as noncrystalline materials. Also used in advanced medical imaging techniques, such as magnetic resonance imaging (MRI).
22	Optical parameter oscillator /OM	Optical microscope	The basic optical microscope, improves resolution, uses visible light, easy to develop.
23	RADIANT	RADIANT ferroelectric testing	Characterizing non-linear materials. Precision and accuracy have been the driving force behind the engineering of test equipment and thin ferroelectric film components.
24	Raman Spectra	Raman Spectroscopy	Commonly used in chemistry to provide a structural fingerprint by which molecules can be identified. The technique typically used to determine vibrational modes of molecules, although rotational and other low-frequency modes of systems may also be observed.
25	RT-MS	Room Temperature-Monochromator Spectrometer	A monochromator produces a beam of light with a very narrow bandwidth of light of single color. It is widely used for spectroscopic analysis of sample materials. The incident light from the light source can be transmitted, absorbed, or reflected through the sample.
26	SPS	Syndiotactic Polystyren	SPS techniques are refractory metals and intermetallics, oxide, and non-oxide ceramics. The particles constituting the powders before consolidation tend to decrease their surface energy by desorption of chemical species, once introduced inside the SPS chamber.
27	TCSPC	Time-correlated single-photon counting	Fluorescence lifetimes, occurring as emissive decays from singlet-state, approximated in time region from picoseconds to nanoseconds.
28	TF Analyzer	Thin-film analyzer	The most sophisticated analyzer of electro-ceramic materials and devices. The test equipment is based on a modular idea, where four different probe heads can be connected to the same basic unit. Each of the four-probe heads offers different characterization methods.
29	TG	Thermogravimetric Analysis	Thermal analysis in which the mass of a sample is measured over time as the temperature changes.
30	USAXS	Ultra-small-angle X-ray Scattering Spectrometer	SAXS and USAXS belong to a family of small angle X-ray scattering techniques that are used in the characterization of materials. This instrument can record data at smaller angle, to resolve and probe larger dimension objects.
31	UV-Vis	Ultra Violet – Visible Spectroscopy	It is absorption spectroscopy, measurement of attenuation of a beam of light after it passes through a sample or after reflection from the sample surface.
32	VSM	Value-Stream Mapping	Analyzes flow of materials.
33	XPS / ESCA	X-ray photoelectron spectroscopy or Electron Spectroscopy for Chemical Analysis	Widely used for surface analysis technique because it can be applied to a broad range of materials and provides valuable quantitative and chemical state information from the surface of the material being studied.
34	XRD	X-ray Powder Diffraction	The analytical technique primarily to identify crystal structure, unit cell, particle size and strain measurement.
35	XRF	X-ray Fluorescence Spectrometer	A non-destructive analytical technique used to determine the elemental composition of materials. XRF analyzers determine the chemistry of a sample by measuring the fluorescent (or secondary) X-ray emitted from a sample when it is excited by a primary X-ray source.
36	XRR	X-ray reflectivity	It is a analytical technique using reflected beam of x-rays from flat surface, measured for the intensity of x-rays reflected in direction to understand surface-sensitivities

Improvement of Body Movements and Stability of Blind or Visually Impaired Adults by Physical Activity using Kinect V2

Marwa Bouri[1], Ali Khalfallah[2], Med Salim Bouhlel[3]

Electrical Engineering, National School of Engineers of Carthage[1]
Industrial Computer Engineering, National School of Electronics and Telecommunications of Sfax[2]
Biomedical Engineering, Higher Institute of Biotechnology of Sfax[3]
Research Unit Sciences and technologies of image and telecommunications, Sfax, Tunisia[1, 2, 3]

Abstract—People who are blind or low vision need to follow activities routines for their mental and physical health to minimize the risk of suffering from bleeding in articulation but they have problems due to difficulties and inaccessibility of displacement. This paper introduces and evaluate a set of exercises to improve the bodily movement and stability using body tracking by Microsoft Kinect V2 and audio feedback. These exercises are composed of a sequence of different postures, has an audio feedback personalized to help people to understand each gestures and can correct them if it is not correct, and generates a summary graph to evaluate the success rate of exercises. To obtain the 3D joint coordinates from the depth sensor, we used the SDK V2.0 of the Microsoft Kinect. We use these coordinates to calculate the distances and angles between joints of interest firstly to position the user in the area field of the Kinect sensor, evaluate the different postures of movements of knees, elbows and shoulders, and detect the body balance if he is leaning and in which direction to avoid falling. These physical exercises have been evaluated to improve feasibility and feedback with persons who are blind or low vision.

Keywords—Posture; visual impaired; physical exercise; audio feedback; Kinect; body tracking; balance; falling

I. INTRODUCTION

In our daily lives, physical activity is very important; it helps us to protect our health from various medical problems. Thus, many studies indicate the importance of physical training to improve the health of people and maintain them in a better physical condition.

According to the World Health Organization (WHO), we have at least 2.2 billion people have a vision impairment or blindness [1]. These people have an inability to practice physical activities and had a problem with balance that puts them in a great possibility of falling. In addition, they cannot move without any help especially during this epidemic of the COVID-19 virus, which requires putting a distance of at least one meter between two people.

The practice of sports activities for blind person or those with reduced vision is difficult without being helped by another person because of the risk of falling and being injured due to the lack of practice of physical exercises.

The daily practice of physical activities for the blind or visually impaired person helps them to recycle posture, maintain flexibility in the body articulations and reduce the falling risk by reinforcing the body balance. In order to give people a certain autonomy by offering them a better quality of life, certain studies have been developed to study and use several tools and techniques.

Postural control is the maintenance of balance of the body either by keeping the body's center of mass or by returning it to its own base of support, which is defined as the maintenance, achievement or restoration of state of equilibrium. Some studies have shown increased risk of falls as well as impaired balance due to decreased ability to control posture [2, 3]. Thus, given the importance of fall avoidance, a method to assess postural stability that can predict the risk of falling has been developed [4, 5].

Indeed, falls present a major health problem. According to data provided by the World Health Organization (WHO) in January 2018, falls are the second-largest accidental mortality in the world causing the annual death of an estimated 646,000 people. In addition, the number of people who require medical attention because of falls has been estimated globally to be around 37,300,000 per year [6]. According to the WHO, the elderly (> 65 years) are the people most affected by fatal falls. It is also reported that approximately 28-35% of this age group (32–42% for> 70 years) fall annually [7].

Therefore, in order to assess the risk of falling, it is necessary to develop measures that characterize and determine the effects of aging on postures.

The Kinect sensor, thanks to these advantages: simplicity, affordability, reliability as well as validity, has been extensively tested and verified in numerous experimental studies devoted to measurements of the human joint center [8-11]. Several studies have focused on its measurement errors [12]. These results support the use of the Kinect sensor to assess walking and balance performance [13-17]. It is already mentioned the availability of some similar RGB depth sensors (RGB-D). Readers interested in it are referred to these alternatives to the Kinect [18].

Several recent studies regarding depth camera applications for analyzing and capturing human movements have been investigated. Among these, the Microsoft Kinect V2 sensor offers both high-resolution color images, as well as wide field of view (FOV) areas and powerful software development kits to detect the skeleton and joints. In addition to the above, the reliability and validity of the Kinect sensor has been studied and verified as much, which is why we have opted for the Kinect V2 as a measuring device for monitoring the skeleton and joints.

Several study are done on the practice of physical activities for different categories people. For example, [19] used the airflow to reinforce the motivation and the bodily movements for children who are visually impaired to increase positive interactions. In [20] the Microsoft Kinect V1 is used in a yoga exergaming dedicate for blind and low vision people. However, most of them have not been interested for blind or visually impaired person, for example in [21] they develop a simple algorithm for the automatic classification of human posture detection for body rehabilitation, and in [22] they developed a sports exercise for an athlete using the Kinect V2 for the distance and angles calculations. Another study presented in [23] presents an exergaming called HemoKinect to supervise patients with hemophilia during physical exercise. In [24] they used a rehabilitation exercise-tracking algorithm, based on a pose classification scheme combined with a trajectory recognition approach.

In this paper, in the next section (Section II) we will detail the materials and the methodology used. Then in Section III, the results will be presented with summary graphs of the joints angle trajectory and followed by a discussion in Section IV. We will finish with a conclusion in the last section (Section V).

II. MATERIALS AND METHODS

Kinect v2 sensor is a device used for human machine interaction. As shown in table I, the Kinect support three types of original data stream can be extracted via the Kinect sensors, it is indeed the depth data stream which is extracted using a depth sensor formed by an infrared CMOS camera and several infrared emitters. Also, the color video feed provided by a color camera. As well as the audio data stream delivered by the microphone.

The distance between the object and the camera in the visible range is reflected by the depth image formed by the Kinect depth data stream. In order to locate and track the joints of the human skeleton, the depth image captured by Kinect is used. The Kinect has the ability to recognize skeletal data from the entire body of six people in measurement space, as well as generate around 30 frames of skeletal data per second and provide three-dimensional coordinate information for 25 joints of the human skeleton (Fig. 1).

Indeed, an algorithm, which belongs to the background race, makes it possible to calculate the 3D depth and to carry out the follow-up of the skeleton of the users. These data are used to express the 3D positions of the joints using vectors.

However, the internal structure of the human body has not been taken into consideration. Indeed, taking into account the structure of the human body, there was an extraction of the distance characteristics of the three-dimensional coordinate data of the skeletal joints were used for the purpose of representing posture.

The Kinect sensor has a detection area ranging from 0.5m up to 4.5m. Nevertheless, for good detection and tracking of the skeleton it is preferable that the user is located in a distance of 1 to 4 m that covers the right area (sweet spot) with an angle width of 70 ° degrees (Table I). For this why the sensor is placed it at a height of 1 m (y-axis) and a Z distance of 2.5 m as shown in Fig. 2.

TABLE I. KINECT VERSION 2 CHARACTERISTICS

Feature	Kinect V2 Xbox one
Color Camera	1920 x 1080 , 30 fps
Depth Camera	512 x 424 , 30 fps Time of Flight (ToF) depth sensor IR can be used at the same time as color
Range	0.5m (1 ft) to 4.5m (14.7 ft)
Angular field of view	70° Horizontal 60° Vertical
Audio	16 bit per channel with 48 kHz sampling rate
Skeletal joints	25 joints
Skeletons traked	6
Vertical adjustment	Manual (±27°)
Latency	~50ms
USB	3.0

Spine-Base = 0
Spine-Mid = 1
Neck = 2
Head = 3
Shoulder Left = 4
Elbow Left = 5
Wrist Left = 6
Hand Left = 7
Shoulder Right = 8
Elbow Right = 9
Wrist Right = 10
Hand Right = 11
Hip Left = 12
Knee Left = 13
Ankle Left = 14
Foot Left = 15
Hip Right = 16
Knee Right = 17
Ankle Right = 18
Foot Right = 19
Spine-Shoulder = 20
Hand Tip Left = 21
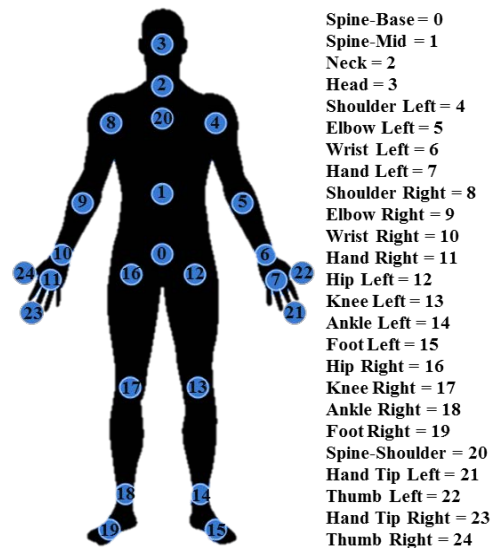Thumb Left = 22
Hand Tip Right = 23
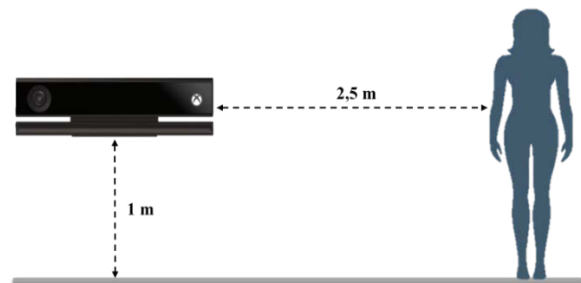Thumb Right = 24

Fig. 1. Kinect Skeleton Joints.

Fig. 2. Experimental Setup.

In order to guarantee obtaining 3D articulation positions, we benefit in our research from the precision of the Microsoft Kinect SDK v2.0 libraries in order to calculate the distances and angles between joints to set the conditions for the evaluation of the exercises and body balance. Using the random decision forest algorithms [25], skeletal data was provided from the depth images.

$$\text{distance} = \sqrt{\begin{array}{c}\left(X_{first}-X_{second}\right)^2+\left(Y_{first}-Y_{second}\right)^2\\+\left(Z_{first}-Z_{second}\right)^2\end{array}} \quad (1)$$

Where $\text{Joint}_{first} = (X_{first},Y_{first},Z_{first})$ and

$\text{Joint}_{second} = (X_{second},Y_{second},Z_{second})$

$$\text{angle} = \frac{180}{\pi}\arctan\frac{|\vec{u}\times\vec{v}|}{\vec{u}.\vec{v}} \quad (2)$$

The flowchart of the system is represented in Fig.3 as we said we are interested of the depth flow to obtain the skeleton data. After the acquisition of the 3D coordinates of the different joints of the body (x, y and z), which are used to calculate the distance and the angle between joints. The distance between the sensor and the user is calculated along the longitudinal (Z-axis) and the transverse (X-axis) direction to position the user in the good detection fields of the sensor to guarantee a wide field for the acquisition of skeletal data. After the validation of the user position, the distance and the angle associated with interest joints were calculated as indicated by equation (1) and (2), to detect the posture of the user and evaluate the gesture used to improve the physical body movement. The angle was calculated by defining a couple of vectors $(\vec{u},\vec{v}) \in R^3$ formed by the adjacent body sections of the joint and taking the angle in degrees between them. Then, a skeleton is displayed in real time on the computer screen and an audio feedback from speaker to notify the user by a voice message. Finally, summary graphs of the gestures are generated.

The exercises used in this work are composed of a sequence of different postures created to enhance the body movement (Fig. 4). The poses used for the training algorithm to enhance movement are described like this; the first pose is the neutral pose, which define the standing posture. The second, arms opened and straight with a condition at the angle of the elbow that should be almost equal to 180 ° and the wrist at the shoulder level. The third pose, right and left arm is opened and up. The forth, both hands are forward and straight. The fifth pose, the user opened her arms and make an elbow angle of 90°. The sixth pose the user joined her hands forward. The seventh pose the user make two hands on waist. The eighth, left knee up and the last pose is the right knee up with a knee angle lower than 165° for the two latest poses.

Movement exercises are described as a sequence of postures. The simplest movement is described by the start and end position (as shown in Table II). Each user was asked to perform seven exercises and repeat each of them three times.

These exercises are used for the flexibility of different joints in the body and improved the physical movement and the body balance because the visually impaired person suffer from imbalance, for this , we took into account and use a voice notification to alert the person when he is leaning and in which direction to correct his posture and avoid falling. The balance is evaluated in four directions along the X and Z plane, which are right, left, forward and backward (Fig. 5).
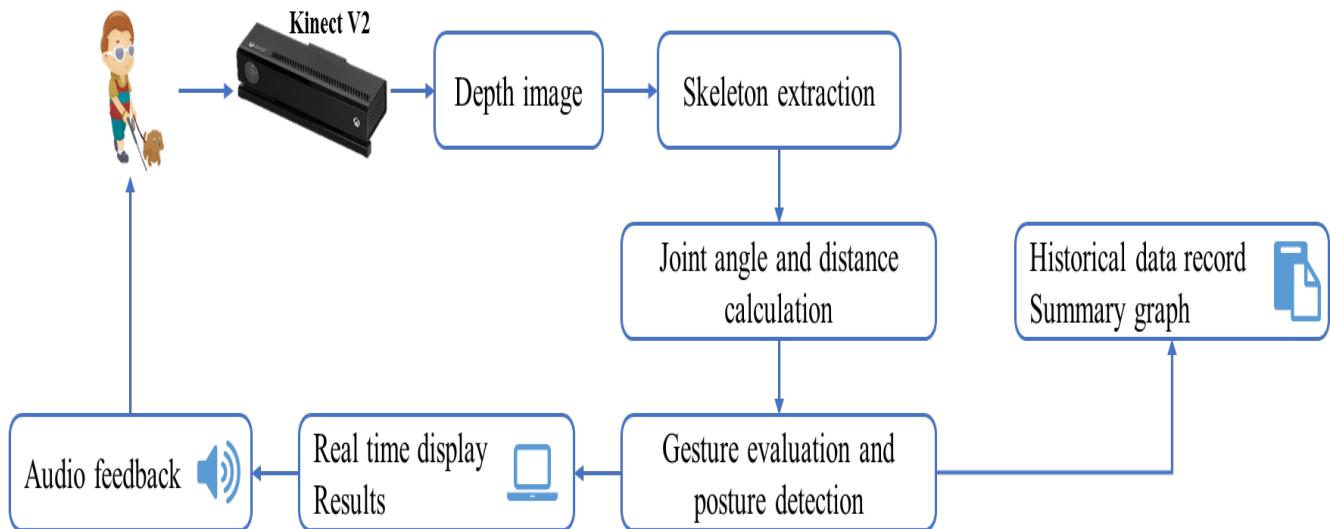

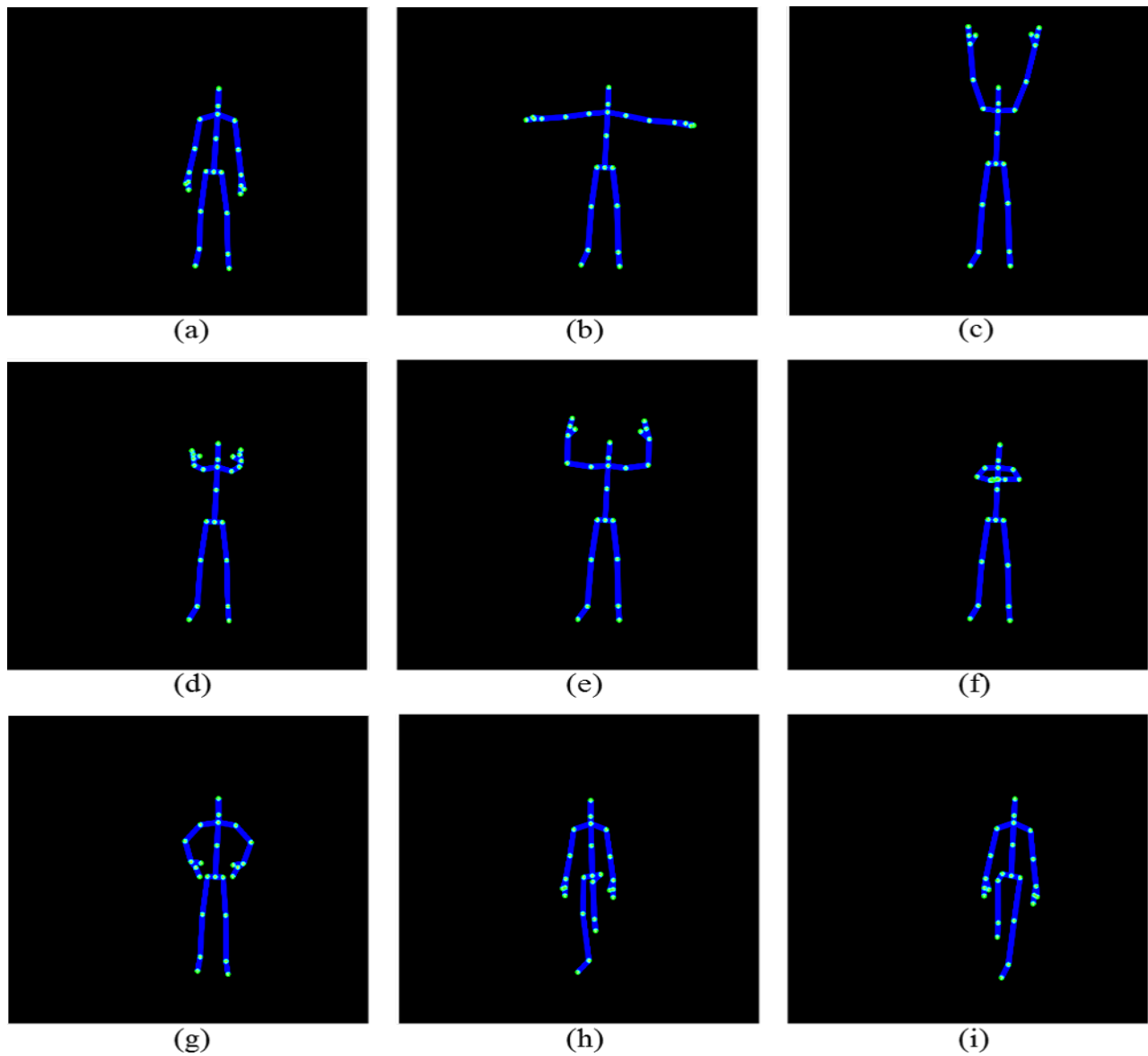Fig. 3. Flowchart Descripe the Implemented Methodology.

Fig. 4. Scrreenshot of Postures : (a) Hands Down; (b) Hands Opened; (c) Hands up; (d) Hands Forward; (e) Hands 90°
(f) Joined Hand; (g) Hands on Waist; (h) Left Knee up; (i) Right Knee up.

TABLE II. EXERCISES SEQUENCE

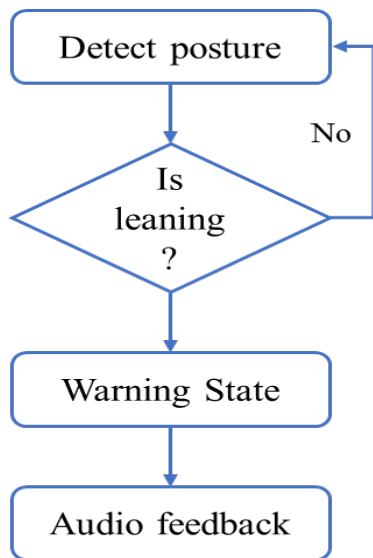| No.	Sequence of pose defining exercises
1	hands down + hands opened + hands up (a+b+c)
2	hands down + hands forward + hands up (a+d+c)
3	hands down + hands opened + hands up + hands forward (a+b+c+d)
4	hands down+hands opened+hands 90° (a+b+e)
5	hands down + hands opened + joined hands (a+b+f)
6	hands on waist + left knee up (j+h)
7	hands on waist + right knee up (j+i)

Fig. 5. Leaning Detection Flowchart.

III. RESULTS

A sample graphs for each exercise are represented in Fig. 6. For each exercise, an ensemble of rules is used to evaluate the body movement and detect the posture of the person if he is leaning or not. These conditions are imposed with taking into account the limitations of the mobility of blind or partially sighted people.

To facilitate the set of postures sequences exercises, every participant is asked to repeat each exercise three times. These exercises deal with the flexibility of the knees, elbows and shoulders. Considering the difficulty of some gestures for some

people, a slight difference is perceived in the results but this does not prevent the success of this set of different gestures.

As shown in Table III, the average of detecting the exercise achievement is more than 97%, which indicated that the proposed algorithm is feasible, through the use of equations (3),(4) and (5).

These exercises are based on instruction to verify the state of gestures (correct or not), and correct them if they are false. To inform the user, a verbal message is send through a speaker. The instructions and correction differ between postures. For difficult gestures are difficult they need more instructions than another one.
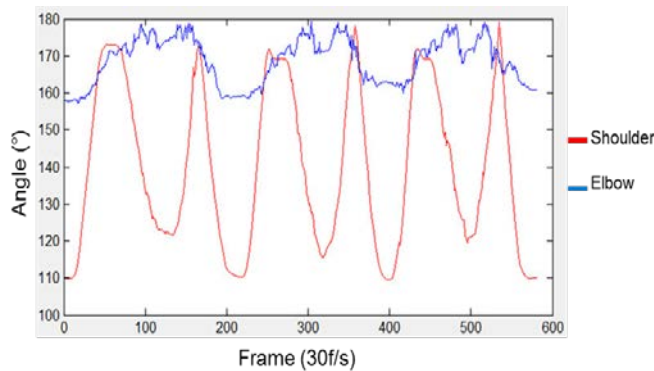
$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad (3)$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (4)$$

$$F_1\text{-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (5)$$

TABLE III. EXERCISES ACHEAVEMENT

Exercise	Recall	Precision	F1-score
1	0.962	0.98	0.97
2	0.957	0.931	0.943
3	0.923	0.9	0.911
4	0.918	0.975	0.945
5	0.942	0.942	0.942
6	0.96	1	0.979
7	0.97	1	0.984

(a)

(b)

(c)

(d)

(e)

(f)
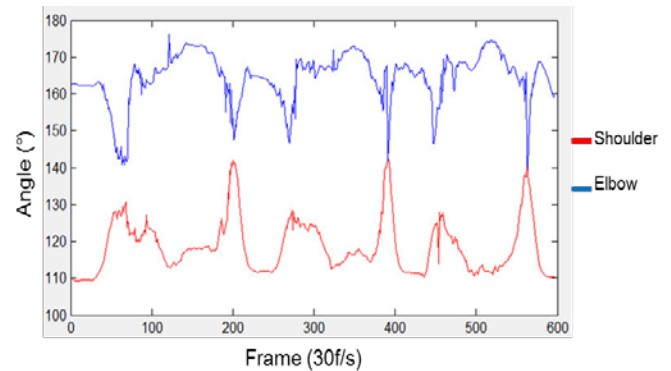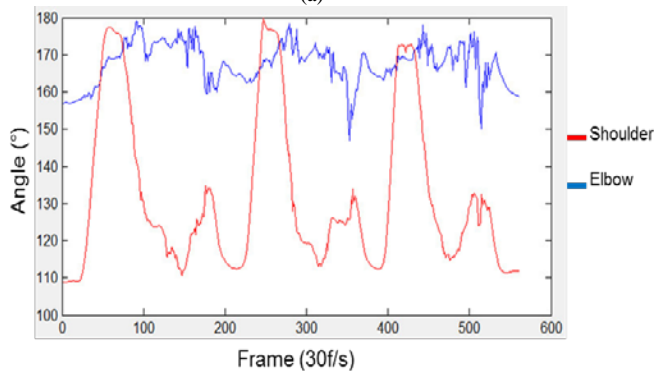
Fig. 6. Joints Angles Graphs of Exercises : (a) Exercise 1 Arms Movement; (b) Exercise 2 Arms Movement; (c) Exercise 3 Shoulder Mouvement; (d) Exercise 4 Elbow Extension and Flexion; (e) Exercise 5; (f) Exercise 6 and 7 Knees Flexions.

IV. DISCUSSION

The physical exercises chosen in this work are applied for the flexibility of the joints in order to improve the physical body movement to keep the body in balance and avoid the risk of fal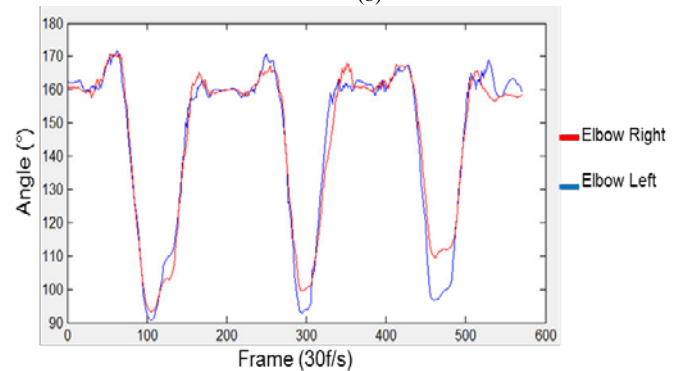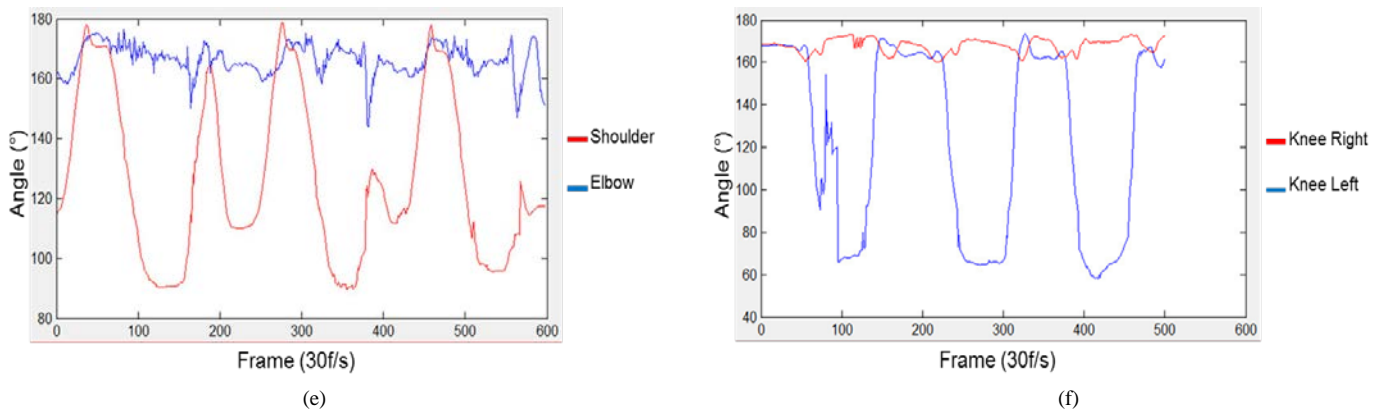ling. These movements are based on flexions and extensions gestures of the knees, elbows and shoulder with conditions on the angles and distances so that they can be easily detected in order to make the requested exercise.

For the movements of different joints which interest as in this work the success rate is 100% for most of the people concerned. The majority of the exercises were performed successfully in all tests performed. Most of the gestures are difficult to execute at the beginning of the training then they are successfully executed.

The aims in this study is to reinforce the physical movement and maintain the body in balance throw a set of different sequences of postures with an audio feedback for instructions and correction of movements facing the Kinect.

To ensure that the messages are clear and understandable enough, we have tried to detail the various instructions as much as possible. For example, for the forth exercise, the instruction is as follows: "raise your arms, open at the level of your shoulder then make an elbow angle of 90°".

Blind or visually impaired person are very careful when they are doing any movement or gestures in the daily life to avoid the risk of falling as they are suffering from imbalance. For this, they fail in some postures because they feel fear of making physical movement, which they not habituated, and the lack of flexibility in body joints.

Different commercial Motion Capture sensors exist such as RealSense R200, Orbbec Astra, ZED stereo camera, Leap Motion and Kinect sensor.

In fact, compared to alternative Motion Capture sensors, Kinect V2 has many advantages. In addition, it can be used to carry out comprehensive monitoring of the skeleton and many bodies collectively. At the same time, this allows the acquisition of data from more than one person. It also benefits from its low price as well as its support for certain varieties of toolkits and software languages. In addition, it has mature drivers, and is privileged by a well accessible and documented SDK.

The Kinect V2 is opted for a sophisticated depth sensor characterized by a higher resolution, as well as by the possibility of carrying out the following of more bodies and articulations per body. Thus, the Kinect V2 becomes a satisfactory alternative for clinical applications, as has been presented by some authors.

In our tests, Kinect V2 is able to measure the angles between the different joints of the skeleton when they are not hidden by any object or obstacle.

To guarantee a good detection and extraction of skeletal data to calculate the angles and distance of the different joints of interest, the user must not go beyond the recommended range with a uniform lighting and without any obstacle or object between the body and the Kinect. In this study, the user can position herself facing the Kinect sensor through a vocal notification.

V. CONCLUSION

To maintain our health in good condition, it is crucial to practice physical exercises. This is why in order to give more importance and dedication to the visually impaired by helping them to train at home with a Kinect sensor via audio feedback messages for the instructions.

For helping them to strengthen their balance, since the blind or visually impaired suffer greatly from balance problems. This is also crucial during this epidemic of the Covid-19 virus, since it is obligatory to keep a distance between people.

The user must position herself to do exercises in the correct fields and direction in front of the sensor used to detect the person. The Microsoft SDK of Kinect V2 is used to obtain the 3D coordinates of the joints data to position person, detect the posture to check if the person is leaning or not to avoid falling risk and evaluate the physical movement through the calculation of distance and angle between joints used for the exercises instructions.

The instruction of the exercises used in this work allow to detect successfully the flexions and extensions of the knees,

elbows and shoulders joints for the diverse postures and measure body leaning to save it in balance.

Though this work is not intended just for blind or partially sighted people, it is available for any person who want to try this treatment without watching to the screen only by receiving a customized and useful audio feedback.

Indeed, for the good acquisition of the data it is necessary to add an obstacle detection algorithm, which is a limitation in this study, to inform the person and guide him to move in order to do these exercises.

The future study will add new exercises that are more difficult and they will be distance supervised by a therapeutic doctor to improve musculoskeletal health.

REFERENCES

[1] World Health Organization. Available online: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed on july 2020).

[2] Pajala, S.; Era, P.; Koskenvuo, M.; Kaprio, J.; Törmäkangas, T.; Rantanen, T. Force platform balance measures as predictors of indoor and outdoor falls in community-dwelling women aged 63–76 years. J. Gerontol. Ser. A Biol. Sci. Med. Sci. 2008, 63, 171–178.

[3] Kurz, I.; Oddsson, L.; Melzer, I. Characteristics of balance control in older persons who fall with injury–a prospective study. J. Electromyogr. Kinesiol. 2013, 23, 814–819.

[4] Bergen, G.; Stevens, M.R.; Burns, E.R.; Centers for Disease Control and Prevention, USA. Falls and fall injuries among adults aged _65 years—United States, 2014. Morb. Mortal. Wkly. Rep. 2016, 65, 993–998.

[5] Panel on Prevention of Falls in Older Persons; American Geriatrics Society; British Geriatrics Society. Summary of the updated American Geriatrics Society/British Geriatrics Society clinical practice guideline for prevention of falls in older persons. J. Am. Geriatr. Soc. 2011, 59, 148–157. [CrossRef]

[6] World Health Organization. Available online: http://www.who.int/mediacentre/factsheets/fs344/en/ (accessed on January 2020).

[7] World Health Organization. Available online: https://www.who.int/ageing/publications/Falls_prevention7March.pdf (accessed on January 2020).

[8] Yang, Y.; Pu, F.; Li, Y.; Li, S.; Fan, Y.; Li, D. Reliability and validity of Kinect RGB-D sensor for assessing standing balance. IEEE Sens. J. 2014, 14, 1633–1638.

[9] Clark, R.A.; Pua, Y.-H.; Oliveira, C.C.; Bower, K.J.; Thilarajah, S.; McGaw, R.; Hasanki, K.; Mentiplay, B.F. Reliability and concurrent validity of the Microsoft Xbox One Kinect for assessment of standing balance and postural control. Gait Posture 2015, 42, 210–213.

[10] Otte, K.; Kayser, B.; Mansow-Model, S.; Verrel, J.; Paul, F.; Brandt, A.U.; Schmitz-Hübsch, T. Accuracy and reliability of the kinect version 2 for clinical measurement of motor function. PLoS ONE 2016, 11, e0166532.

[11] Napoli, A.; Glass, S.;Ward, C.; Tucker, C.; Obeid, I. Performance analysis of a generalized motion capture system using microsoft kinect 2.0. Biomed. Signal Process. Control 2017, 38, 265–280.

[12] Xu, X.; McGorry, R.W. The validity of the first and second generation Microsoft Kinect. for identifying joint center locations during static postures. Appl. Ergon. 2015, 49, 47–54.

[13] Eltoukhy, M.; Kuenze, C.; Oh, J.; Wooten, S.; Signorile, J. Kinect-based assessment of lower limb kinematics and dynamic postural control during the star excursion balance test. Gait Posture 2017, 58, 421–427.

[14] Eltoukhy, M.A.; Kuenze, C.; Oh, J.; Signorile, J.F. Validation of static and dynamic balance assessment using Microsoft Kinect for young and elderly populations. IEEE J. Biomed. Health Inform. 2017, 22, 147–153.

[15] Lv, Z.; Penades, V.; Blasco, S.; Chirivella, J.; Gagliardo, P. Evaluation of Kinect2 based balance measurement. Neurocomputing 2016, 208, 290–298.

[16] Hsiao, M.-Y.; Li, C.-M.; Lu, I.-S.; Lin, Y.-H.; Wang, T.-G.; Han, D.S. An investigation of the use of the Kinect system as a measure of dynamic balance and forward reach in the elderly. Clin. Rehabil. 2018, 32, 473–482.

[17] Ejupi, A.; Gschwind, Y.J.; Valenzuela, T.; Lord, S.R.; Delbaere, K. A kinect and inertial sensor-based system for the self-assessment of fall risk: A home-based study in older people. Hum.–Comput. Interact. 2016, 31, 261–293.

[18] Clark, R.A.; Mentiplay, B.F.; Hough, E.; Pua, Y.H. Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives. Gait Posture 2019, 68, 193–200.

[19] CHAO, F., CHU, H., et LEE, Liza. Enhancing Bodily Movements of the Visually Impaired Children by Airflow. Advances in Science, Technology and Engineering Systems Journal, 2019, vol. 4, no 4, p. 308-313.

[20] RECTOR, Kyle, VILARDAGA, Roger, LANSKY, Leo, et al. Design and real-world evaluation of Eyes-Free Yoga: An exergame for blind and low-vision exercise. ACM Transactions on Accessible Computing (TACCESS), 2017, vol. 9, no 4, p. 1-25.

[21] KLISHKOVSKAIA, Tatiana, AKSENOV, Andrey, SINITCA, Aleksandr, et al. Development of Classification Algorithms for the Detection of Postures Using Non-Marker-Based Motion Capture Systems. Applied Sciences, 2020, vol. 10, no 11, p. 4028.

[22] ÖRÜCÜ, Serkan et SELEK, Murat. Design and validation of rule-based expert system by using kinect V2 for real-time athlete support. Applied Sciences, 2020, vol. 10, no 2, p. 611.

[23] MATEO, Fernando, SORIA-OLIVAS, Emilio, CARRASCO, Juan J., et al. HemoKinect: a microsoft kinect V2 based exergaming software to supervise physical exercise of patients with hemophilia. Sensors, 2018, vol. 18, no 8, p. 2439.

[24] VELASCO, Fernando et NARVÁEZ, Fabián. Automatic Exercise Recognition Based on Kinect Sensor for Telerehabilitation. International Conference on Smart Technologies, Systems and Applications. Springer, Cham, 2019. p. 312-324.

[25] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images" In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Springs, CO, USA, pp. 1297–1304, 20–25 June 2011.

CAREdio: Health Screening and Heart Disease Prediction System for Rural Communities in the Philippines

Lean Karlo S. Tolentino[1] , John Erick L. Isoy[2], Kayne Adriane A. Bulawan[3], Mary Claire T. Co[4]
Caryl Faye C. Monreal[5], Ian Joshua W. Vitto[6], Maria Victoria C. Padilla[7], Jay Fel C. Quijano[8]
Romeo Jr. L. Jorda[9], Jessica S. Velasco[10]

Department of Electronics Engineering, Technological University of the Philippines, Manila, Philippines[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
University Extension Services, Technological University of the Philippines, Manila, Philippines[1]

Abstract—**Cardiovascular diseases cover a large quantity of worldwide disease load, setting it to top leading cause of death. In the Philippines, given the rapid economic advancement and urbanization, the most vulnerable sector has not been impacted by this development. Data from the Philippine Statistical Authority (PSA) in 2016 revealed that of the country's total recorded deaths, six out of ten were medically unattended and of which the largest portion are from the rural population. Consequently, medical analysis is needed to perform effectively and precisely however, most developing countries have limited resources and lack medical expert for specialized field such as cardiologists. The proponents essentially seeks to address the issues Philippine health sector specifically in rural and remote populace by executing efficient and low-cost health screening and diseases prediction system using commercially available medical devices and machine learning algorithms for the prediction of three of the most heart diseases (Hypertension, Heart Attack, Diabetes). The system is composed of CAREdio mobile app, prototype hardware consists of different health sensors and devices, and a machine learning model that is applied to determine the user's individual probability of having a specific heart disease. The machine learning models used were trained using the data gathered from Rosario Reyes Health Center and Ospital ng Sampaloc (Sampaloc Hospital), both located in Manila City, Philippines. CAREdio achieves accuracy values over 0.80 for all diseases. The system can diagnose multiple cardiovascular diseases in a single app that will benefit people rural communities.**

Keywords—Cardiovascular diseases; health screening; disease prediction; mobile application; machine learning; rural population

I. INTRODUCTION

The Philippines is an archipelago of over 7,000 islands, with an entire land scale of approximately 300,000 square kilometers, and above 100 million inhabitants, placing it to the thirteenth most populated country in the world [1], these brings a genuine task in providing efficient and effective health care particularly for rural and remote population.

In spite of the rapid economic advancement and urbanization, the most vulnerable sector has not been impacted by this development. Philippine Statistical Authority revealed that of the country's total recorded deaths in 2016, six out of ten were medically unattended. The largest portion of

medically unattended deaths are from the rural population. Only the National Capital Region (NCR) had a greater total of medically assisted deaths than unattended in general. It infers that compared to other regions; NCR has better access to quality health care [2].

Because of the maldistribution of facilities, health staff and specialists especially in rural regions, it has put the residents into a terrible situation. The limited number of health facilities relating to the growing population, and shortage of physicians add up to a low quality of health care. Likewise, based on a review by the Social Weather Station, a large portion of Filipinos, particularly the low-income households, choose to get treatment in a government-owned medical facility if a member of the family needs confinement and it is mostly because of affordability [3].

The world's leading cause of deaths are non-communicable diseases (NCDs) with 41 million individuals die every year, which equates to 71 percent of worldwide deaths. [4]. In 2016, approximately 17.9 million people died from heart illnesses, indicating the 31 percent total number of deaths. Eighty five percent of these deaths are because of heart attack and stroke. In the Philippines, NCDs are increasing quickly, seven of the ten top causes of death are non-communicable in nature. Most of the NCDs mortality cases are diabetes, chronic obstructive pulmonary disease, cardio-vascular diseases, and cancer [5].

The fast evolution of computational capacities has given present-day techniques disease prediction. Techniques that can reveal clinically important and relevant information concealed in the vast amount of data, which in this way, can support medical decision making. An artificial intelligence (AI) system can reduce medical and diagnostic mistakes, which in clinical practice is inevitable. Additionally, an AI system may collect valuable information from a huge population of patients to assist to make uninterrupted analysis for notifying health risks and calculating health findings. Machine learning algorithms are valuable for discovery complex patterns in large data.

The researchers principally pursue to address the issues in Philippine health sector specifically in rural and remote populace by executing efficient and low-cost health screening and diseases prediction system using commercially offered medical devices and machine learning algorithms. The specific

objectives of the study are: (1) to use commercially offered medical sensors and screening devices that can collect and read health parameters including heart rate (HR), body mass index (BMI), blood glucose (BG), oxygen saturation (SpO2), blood pressure (BP), uric acid level, and cholesterol; (2) to use the machine learning models such as Support Vector Machine, Naive Bayes Classifier, Random Forest, Logistic Regression, and Neural Network; and to predict the individual probabilities of a person having specific heart diseases; and (3) to evaluate the overall system in terms of accuracy and efficiency through comparison tests between the prediction engine and standard tests.

Data gathering for machine learning algorithm implementation was conducted at Rosario Reyes Health Center in San Andres, Manila, and Ospital ng Sampaloc, a District Hospital also in Manila. The collected data are from patients who have heart diseases. From this, the researchers were able to come up with three most common non-communicable diseases namely: heart attack, hypertension, and diabetes. Some other data collected were compacted into a category called as others.

The prototype was deployed at an experimental scale in Barangay Medicion II-B in Imus City, Cavite, Philippines. The prototype is composed of weight sensor, height sensor, heart rate sensor, glucometer, cholesterol meter, uric acid meter, oxygen saturation sensor, and sphygmomanometer for blood pressure, microcontroller, thermal printer, and android tablet installed with CAREdio application. The testing of patient is done in less than 10 minutes, depending on the patient. At first, the patient would have to enter his name, age, and sex in the application before proceeding to the actual testing. The patients will also have to answer several medical questionnaires integrated in the application that would help in prediction of their individual probabilities of having diseases.

II. CONCEPTUAL LITERATURE

A. Non-Communicable Dieseases

Non-Communicable Diseases (NCDs) are the premier causes of deaths in the country. NCDs such as pulmonary diseases, cardiovascular diseases, diabetes, and cancer.

Cardiovascular Disease (CVD) may be a complex class for certain diseases that can influence the heart and blood vessels. The CVD affects the structure or functions of the heart which include inherent cardiopathy, artery illness (narrowing of the arteries), irregular heart rhythms (arrhythmias), heart valve disease, cardiac failure, and muscle tissue illness. These diseases occur because of imbalance in blood pressure and pulse rate. Clear addition to the CVD disease is the increase in blood pressure, smoking, sterol (diabetes polygenic disorder polygenic disease) [6].

Some variables known to have a substantial percentage of risk for CVDs include age, gender, heart rate, level of cholesterol, level of blood pressure and BMI.

Diabetes is deemed as one of the most pervasive illness. It characterizes a group of high - glucose metabolic illnesses in the blood basically because of inappropriate ingestion of insulin. Preserving the balance between hypoglycemia (low glucose level) and hyperglycemia (high glucose level) is the job of insulin.

There are three types of diabetes: type 1 it the condition where the pancreas is not producing insulin, and type 2 is where the pancreas creates insulin but is not adequate to sustain the stability.

Heredity, obesity, hypertension, aging, unbalanced diet, lack of physical activity, and because of the nature of community are some of the factors that contribute to having type 2 diabetes. In type 2 diabetes, cardiovascular disease is critical. An additional type where the amount of glucose is increasing during pregnancy is called is called Gestational Diabetes [7].

B. Health Parameters

To get a reliable picture of one's wellbeing, predict pattern and screening of health, several health factors are being accessed. Table I shows the health parameters, values, and remarks which is to be used in arranging the data collected and ran using machine learning algorithms [5], [8 - 12].

C. Machine Learning Algorithms

The study of computer algorithms that develops naturally through repetitive experience is called machine learning algorithms (MLA). It is viewed as a subgroup of artificial intelligence. Machine learning algorithms constructs a mathematical model established on sample data, known as "training data", so as to make predictions or decisions without being openly coded to do so [13]. In this study, we utilized 5 of most common machine learning models:

1) An array of layers consists of many very interconnected parts (i.e. neurons) doing an activation is how neural network is designed. One or more concealed and the output layers are gradually generated when there is an interaction in the input layers [14].

2) Random forest is a regression and classification model that builds decision trees for every attribute; modifies the overfitting to its training set; keeps outliers of missing quantities by doing steps in analysis, pre-processing data. It's a method of creating a specific model anywhere poor models are associated. The random forest is made up of the numerous decision trees connected to the forest [15].

3) Support Vector Machine can be used for classification and regression. Though, this model is usually used in the classification applications. The SVM definition focuses on finding a dimension (straight line for 2D; plane for 3D) which offers the best path to split the groups [16].

4) Naive Bayes Classifier is a classification method that is established on Bayes' Theorem with a notion of autonomy among predictors. Naive Bayes presumes in simple terms, that the existence of an element in a class is unrelated to the presence of any other element [17].

5) Logistic regression is a model that uses logistic function to build a binary dependent variable with two possible values. It can be compatible to a dependent variable. Binary variables, 1 and 0 indicating they succeed or fail [18].

TABLE I. LIST OF HEALTH PARAMETERS

No.	Attribute	Health Parameter	Value	Remarks
1	age	Age	-	-
2	sex	Sex	0 1	Female Male
3	bmi	Body Mass Index	0 (18.5 – 22.9) 1 (23 – 24.9) 2 (25 – 29.9) 3 (=> 30)	Normal Overweight Pre-obese Obese
4	maxhr	Max Heart Rate	0 (60-90 bmp) 1 (>90 bpm)	Normal Abnormal
5	bos	Blood Oxygen Saturation	0 (95%-100%) 1 (<94%)	Normal Hypoxemia
6	fbs	Fasting Blood Sugar	0 (<100 mg/dl) 1 (101-125 mg/dl) 2 (>126 mg/dl)	Non-diabetic Pre-diabetes Diabetic
7	bpsys	Blood Pressure (systolic)	0 (<120) 1 (120-139) 2 (>140)	Normal At Risk High
8	bpdio	Blood Pressure (diastolic)	0 (<80) 1 (80-89) 2 (>90)	Normal At Risk High
9	chllvl	Cholesterol Level	0 (<200 mg/dl) 1 (201-239 mg/dl) 2 (=< 240 md/dl)	Desirable Borderline High
10	sm	Smoking	0 1	False True
11	famhis	Family History	0 1	False True
12	alchl	Alcohol	0 1	False True
13	stress	Stress	0 1	False True
14	hrtbrn	Heart Burn	0 1	False True
15	dt	Diet	0 1	False True
16	shrtnssbrth	Shortness of Breath	0 1	False True
17	exrcs	Active Exercise	0 1	False True
18	cp	Chest Pain	0 1	False True
19	num	Disease	0 1	False True

III. RELATED WORKS

Disease prediction using artificial intelligence is recently getting international acknowledgment. Presented the large scope of data mining in enhancing health care, this part discusses the methods used in determining and forecasting NCDs.

The predictive model calculates the future results based on past records recovered from database. This kind of model is used by many organizations that try that uses AI to try and understand the connections between the information characteristics. Researchers have suggested and used a variety of techniques and machine learning algorithms which they have implemented in some medical applications. Some of which are as follows:

In [19], the researchers created an Android based app that provides choices as to what patients should do when they experience a specific symptoms, may it be doing a medication at home, or going to a clinic to be checked by a doctor right away. Chaining inferences forward is the method used in this study. Clinical results of a disease can be identified with the use of application and Android phone as the medium through the consultation procedure or answering the questionnaires integrated in the system. After doctors had accessed the result of the general diagnosis symptoms, the concluded that treatment it provides are effective.

Similar study [20] projected a new ear - worn system for intermittent long - term ECG QRS duration analyzing to control the difficulties of recent wearable ECG technologies for example inconvenience of use. To enhanced wear experience, the ECG electrodes was placed at the back of the ear. The

researchers applied a heartbeat validation SVM model, a heartbeat purification method, and a regression model to extract typical chest-ECG QRS from the ear's QRS durations – ECG.

In [21], the proponents discussed all what is called a Hierarchal Health Decision Support System which fuses clinical decision support system and wearable medical sensors (WMSs). Closed - loop and multi - tier hierarchical system supported by stable machine learning classes are incorporated in the system. The proponents proposed an automated application for disease diagnosis, which could investigate various related diseases. Accuracy in the classification of major diseases was achieved as a result of physiological data gathered from WMSs: hypothyroid (95 percent), type 2 diabetes (78 percent), arrhythmia (86 percent), renal pelvis nephritis (94 percent), and urinary bladder disorder (99 percent).

In [7], blood glucose levels using the capacitance method were successfully measured by the proponents. The Artificial Rain Algorithm (ARA) was employed as a framework for the internal discovery of the finest value from the information collected via repeated examination of the capacitance value. The prototype effectively demonstrated the accuracy of 96.92% as contrasted to the laboratory test results. The researchers recommended the use of additional clustering methods to further develop the potential of ARA as an enhancing and analyzing methods.

In [22], the researchers used MLA to predict heart disease, namely: Decision Tree, SVM, and Naive Bayes. The principal component analysis was used to reduce the attributes number. After the size of the dataset has been reduced, SVM beats other two models. To analyze the accuracy of the algorithms the researchers used WEKA data mining method. The algorithms compare the accuracy of the classifier based on its time carried to construct model, ROC Area and mean absolute error. It concludes that the use of Max ROC Areas is in prediction performance was brilliant.

The study in [23] offered the evaluation and application of the given MLAs which was applied in R programming in producing CVDs prediction. Also, web-based application was also created which provides module for the attributes, test results, references, and graphs for entry. The top accuracy was obtained from logistic regression.

Artificial neural network was used in [24] for detecting circulatory diseases through analyzing fingernails. The developed model was tested on 6 patients' fingernail images. Their diagnosed diseases were all matched with the proposed prediction system.

Meanwhile, potentiometric method was implemented in [25] for detecting hormonal imbalance through the females' saliva. The results from the proposed method were compared with the conventional blood test and were all matched.

In [6], a system that produces a prediction if a patient has cardiovascular disease, in terms of YES or NO. Unless the patient is susceptible to cardiovascular disease then the outcome would be YES and vice versa. In case of good results, the patient needs to consult a doctor for more opinion. The cardiovascular disease prediction program uses the MLA

which gives the patient a predictive result which gives the patient status which leads to CAD.

IV. METHODOLOGY

This section will describe methodology followed in developing the CAREdio.

A. Block Diagram

The CAREdio system is comprised of two principal elements. The first item is the set of medical sensor and system safety parameters including and the standardized clinical questionnaires. The second half is for assessment and prediction engine.

The block diagram in Fig. 1 shows how the CAREdio works step by step. This comprises the parameter detection from the prototype kiosk which includes medical sensors and equipment, microcontrollers for the acquisition of data and the analysis, an android application where the user can enter their information and responses to the questions. Analysis of data and prediction of diseases will be done by using the MLA.

B. Machine Learning Implementation

Specific datasets of the diseases were collected to check and train the MLAs. The evaluation and selection of the accuracy and efficiency of the MLA in predicting heart disease shall be the deciding factor in evaluating which one is the best to use. The preferred MLA will then be used and be integrated to the CAREdio system for prediction of individual probability of having diseases.

C. System Flow Chart

Fig. 2 illustrates the flow chart of the CAREdio system. First, the system will prepare and wait for the patient's question to display from the application that we will be using. It will then start to gather the patient's data by filling up the information needed for new users. After logging in, medical sensors will operate to detect patient's health parameter such as height, weight, BMI, blood pressure rate and medical questions will be shown in the android application. The data gathered from the patient will be processed for detecting the probability of having a heart disease. When the results are done, it will automatically be saved in database and the patient has an option whether to print their results or not.
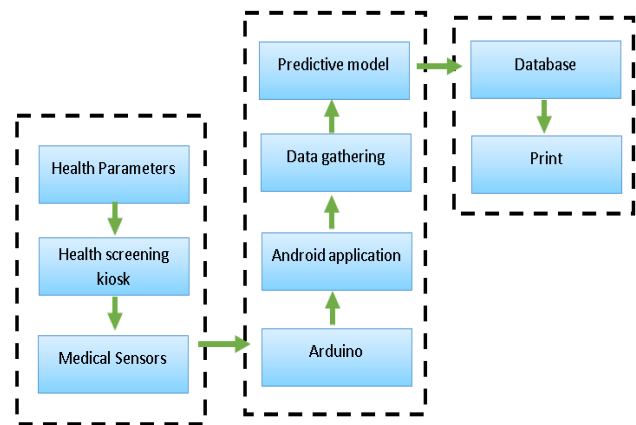

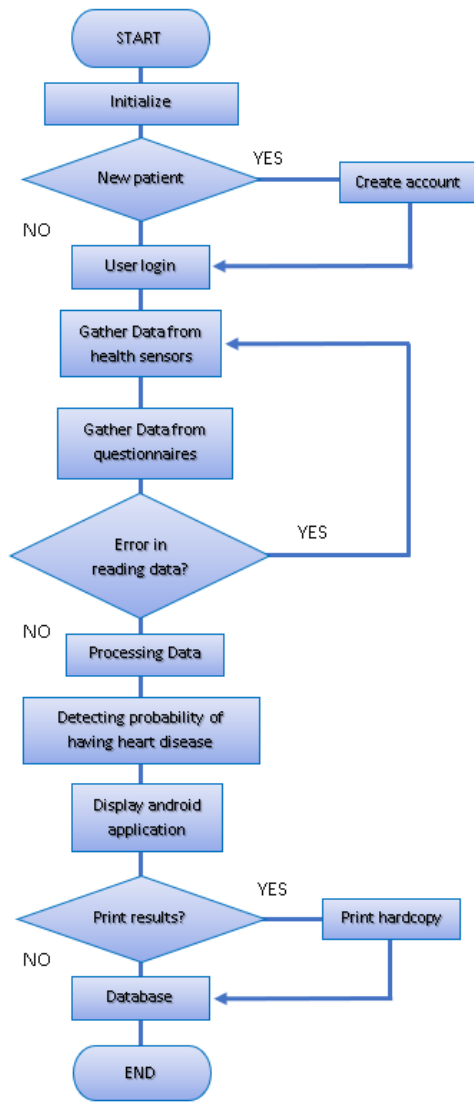
Fig. 1. Block Diagram of the System.

Fig. 2.   System Flow Chart.

Fig. 4 shows the database in android application. The data of the patient were forwarded to the database using SQLite and saved in csv file format. The database consists of timestamp, name of patients, age, health parameters and result percentage. SQLite is an open source relational database that can store and access data across multiple storage engines. It can also duplicate data and sectioning tables for better durability and performance.
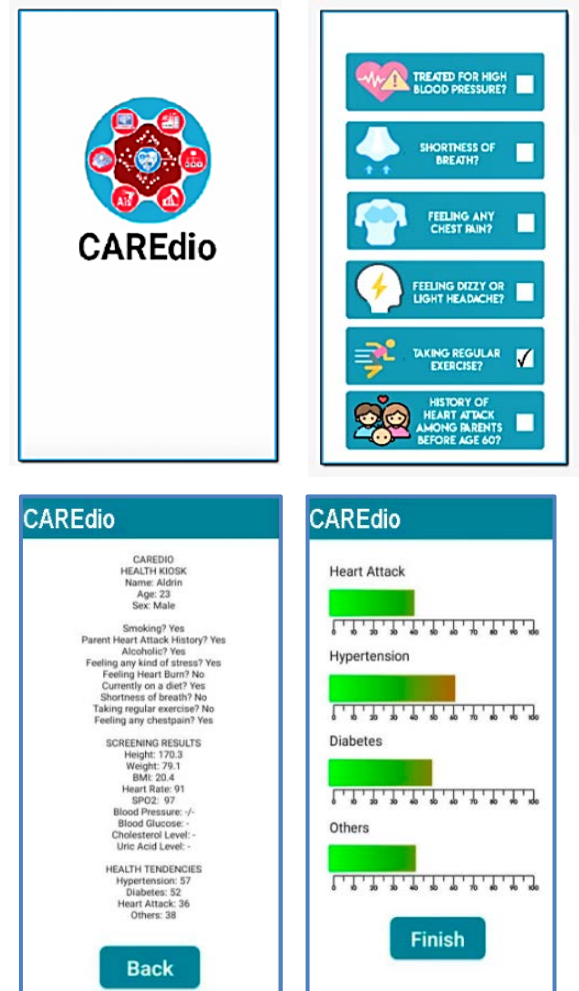

Fig. 3.   From Left, Screenshots of CAREdio Application, (a) Home Page, (b) Input Parameter, (c) Medical Questionnaire.

D. Mobile Application Design

This section presents all of the tabs included in the Android application. Every tab view has a different function for the system which include manual input of patient's information, obtaining of parameters from the prototype, questionnaires, patient and record information, and display of results.

The mobile application design is shown in Fig. 3. On the app interfaces, the user will enter personal data like name, age and gender; after that, they will have to use the prototype for health screening to gather parameters, answer the standard clinical questionnaires. The patients can view the data collected included the prediction percentage, they will also be given also get a hardcopy result. The section for admin is where the data base can be viewed and extracted in .CSV file.

It shows the display of patients result in android application. The results obtained from medical devices and sensors, responses in the questionnaire, and patient information are shown. The probability of having those diseases are also displayed by percentage.

Fig. 4.   Database in Android Application Sample.

E. Hardware Prototype

Fig. 5 shows the proposed hardware structure for the health screening kiosk. It is composed of microcontrollers, medical sensors and equipment, Android tablet, and printer. While the actual prototype is shown in Fig. 6. The height is 4 feet (which be expanded to 6 feet max for the height sensor), while the width spans 1.5 feet. Height sensor (HC-SR04) is positioned at the top of the prototype kiosk. For the middle left, shows the additional sensor, which is MAX30105 for heart rate and blood oxygen saturation. The drawer in the middle is the storage for devices like digital blood pressure; the 3-in-1 test kit for blood glucose, cholesterol level, and acid saturation; Arduino Mega and Raspberry Pi microcontrollers. The device has an aluminum door for a weighing scale located at the lower part of prototype. Thermal printer and 7 inches Android tablet were also installed.
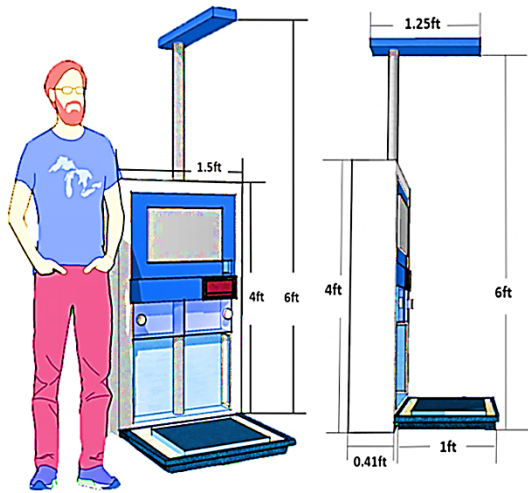

Fig. 5.   Proposed Prototype.

Fig. 6.   Actual Prototype.

Health parameter scores from health sensors and devices, and the answers from medical survey serves as input for the system. The patient will stand on the platform of the prototype for the detection its height and weight, and they are to place their index finger in the pulse oximeter sensor for the blood oxygen saturation and heart rate reading. The blood pressure, cholesterol level, uric acid level, and blood glucose level, will be collected from medical devices and will be manually recorded in the mobile application along with answering the medical questionnaire on it. When all inputs are entered in the system, the prediction of cardiovascular diseases will take place.

V.   ANALYSIS AND RESULTS

The CAREdio is health screening and disease prediction system that utilizes MLA to automate prediction of cardiovascular diseases, specifically (1) diabetes, (2) heart attack, (3) hypertension, and (4) others, a separate group labeled as for other cardiovascular diseases that is currently not included by our system. The system can provide the individual probability of having cardiovascular diseases by displaying each percentage output. CAREdio is projected to meet the standards of health associations, hospitals, and clinic tests with convenience and efficient cost.

The system is not fully automated, some medical devices are not integrated in microcontroller when getting health parameters, it also needs a manual encoding to generate in mobile application. It was created as a health screener for the stated health parameters and disease prediction, it does not give certain treatment, but it will give individual probability of having heart diseases based on the data gathered.

The Random Forest model is chosen for the calculation of the probability of having specific cardiovascular diseases, among the five machine algorithms including Logistic Regression Support Vector Machine, Naive-Bayes Classifier, and Neural Network, because it formed the highest percentage of accuracy when the data collected from hospitals was trained and tested.

A. Machine Learning Accuracy

Table II shows the summary of results of each algorithm. The highest percentage for machine learning is Random forest with average of 79.3% accuracy. Shows the summary of machine learning accuracy percentage for each disease. Random Forest got the highest accuracy rate for hypertension and heart attack.

Table III shows the accuracy score, precision, recall, f1-score, and mean absolute error. The model does well in most heart diseases with accuracy scores above 0.90, except for category "others", which has an accuracy score of 0.87. Precision recall are also above 0.90 for most groups, and f1-score is above 0.88 for all groups. The "others" category has low sensitivity which is to be expected because of the range of diseases that make up this group.

The following table numbers 4.7, 4.8, 4.9, and 4.10, below shows the project device's output comparing it to collected data from Rosario Reyes Health Center, and *Ospital ng*

Sampaloc [Sampaloc Hospital], both located in Manila City, Philippines.

The proponents tested the mobile application against our dataset by randomly choosing 20 patient's data as the "test set" and operating to this it in system to calculate the probability of diseases and the accuracy of the system.

Table IV shows that 20 out of 20 (100%) data of person with diabetes matched. The device coordinates with the results from the dataset collected which produced the individual probability of having diabetes.

TABLE II.　Machine Learning Performace Per Disease

	Logic Regression	Naive Bayes	Neural Network	Random Forest	Support Vector Machine (SVM)
Diabetes	83..33%	81.25%	77.08%	91.67%	87.50%
Heart Attack	80.25%	89.89%	82.72%	90.12%	80.25%
Hypertension	97.53%	90.12%	85.19%	98.77%	98.77%
Others	84.26%	82.41%	80.56%	87.96%	87.96%

TABLE III.　Accuracy Score, Precision, Recall, F1-score, and mean absolute error

	Accuracy score	Precision	Recall	F1-score	Mean Absolute Error
Diabetes	0.98	0.98	0.98	0.98	0.95
Heart Attack	0.91	0.91	0.90	0.89	0.78
Hypertension	0.98	0.99	0.99	0.99	0.99
Others	0.87	0.88	0.88	0.88	0.88

TABLE IV.　Trial For Patient's with Diabetes

Patient No.	With Diabetes	CAREdio Result	Remarks
1	Yes	70% (Yes)	Matched
2	Yes	94% (Yes)	Matched
3	Yes	89% (Yes)	Matched
4	Yes	98% (Yes)	Matched
5	Yes	91% (Yes)	Matched
6	Yes	94% (Yes)	Matched
7	Yes	80% (Yes)	Matched
8	Yes	96% (Yes)	Matched
9	Yes	98% (Yes)	Matched
10	Yes	86% (Yes)	Matched
11	Yes	96% (Yes)	Matched
12	Yes	93% (Yes)	Matched
13	Yes	51% (Yes)	Matched
14	Yes	97% (Yes)	Matched
15	Yes	78% (Yes)	Matched
16	No	10% (No)	Matched
17	No	2% (No)	Matched
18	No	7% (No)	Matched
19	No	18% (No)	Matched
20	No	12% (No)	Matched

Table V shows that 20 out of 20 (100%) data of person with heart attack matched. The device coordinates with the results from the dataset collected which produced the individual probability of having heart attack.

Table VI shows that 20 out of 20 (100%) data person with hypertension matched. The device coordinates with the results from the dataset collected which produced the individual probability of having diabetes.

TABLE V.　Trial for Patient's with Heart Attack

Patient No.	With Heart Attack	CAREdio Result	Remarks
1	Yes	88% (Yes)	Matched
2	Yes	60% (Yes)	Matched
3	Yes	72% (Yes)	Matched
4	Yes	84% (Yes)	Matched
5	Yes	83% (Yes)	Matched
6	Yes	80% (Yes)	Matched
7	Yes	82% (Yes)	Matched
8	Yes	83% (Yes)	Matched
9	Yes	91% (Yes)	Matched
10	Yes	91% (Yes)	Matched
11	Yes	82% (Yes)	Matched
12	Yes	86% (Yes)	Matched
13	Yes	87% (Yes)	Matched
14	Yes	92% (Yes)	Matched
15	Yes	84% (Yes)	Matched
16	No	08% (No)	Matched
17	No	19% (No)	Matched
18	No	15% (No)	Matched
19	No	07% (No)	Matched
20	No	22% (No)	Matched

TABLE VI.　Trial for Patient's with Hypertension

Patient No.	With Hypertension	CAREdio Result	Remarks
1	Yes	97% (Yes)	Matched
2	Yes	96% (Yes)	Matched
3	Yes	78% (Yes)	Matched
4	Yes	91% (Yes)	Matched
5	Yes	90% (Yes)	Matched
6	Yes	99% (Yes)	Matched
7	Yes	88% (Yes)	Matched
8	Yes	91% (Yes)	Matched
9	Yes	90% (Yes)	Matched
10	Yes	90% (Yes)	Matched
11	Yes	86% (Yes)	Matched
12	Yes	88% (Yes)	Matched
13	Yes	89% (Yes)	Matched
14	Yes	87% (Yes)	Matched
15	Yes	89% (Yes)	Matched
16	No	31% (No)	Matched
17	No	71% (Yes)	Mismatched
18	No	10% (No)	Matched
19	No	13% (No)	Matched
20	No	18% (No)	Matched

Table VII shows that 19 out of 20 (95%) data in "others" category matched. The device coordinates with the results from the dataset collected which produced the individual probability of having hypertension.

TABLE VII.   TRIAL FOR PATIENT'S WITH OTHER CVD

Patient No.	With Other CVD	CAREdio Result	Remarks
1	Yes	80% (Yes)	Matched
2	Yes	56% (Yes)	Matched
3	Yes	59% (Yes)	Matched
4	Yes	95% (Yes)	Matched
5	Yes	86% (Yes)	Matched
6	Yes	88% (Yes)	Matched
7	Yes	51% (Yes)	Matched
8	Yes	100% (Yes)	Matched
9	Yes	79% (Yes)	Matched
10	Yes	95% (Yes)	Matched
11	Yes	91% (Yes)	Matched
12	Yes	95% (Yes)	Matched
13	Yes	55% (Yes)	Matched
14	Yes	86% (Yes)	Matched
15	Yes	98% (Yes)	Matched
16	No	4% (No)	Matched
17	No	0% (No)	Matched
18	No	2% (No)	Matched
19	No	9% (No)	Matched
20	No	4% (No)	Matched

VI. CONCLUSION AND FUTURE WORK

The development of health screening and heart disease prediction was successfully executed for the three of the most prevalent heart diseases in the Philippines. In addition to this, the improvement and accuracy of the system was well implemented with Random Forest Classifier.

The system is screening the health parameters with the CAREdio prototype composed of different health sensors and devices and analyzing the data collected from the patients to provide the individual probability of having diseases.

CAREdio has good accuracy score, precision, recall, f1-score and AUC, and can easily be operated by a community health worker to screen health and predict diseases, and also help them to make decisions towards patient health risks. The system is low cost, as it only involves a CAREdio application and the hardware protype. Screening and predicting system such as the CAREdio can provide a significant role for community health workers and doctors to prevent disease and disability in remote and low resource areas. It produces results that make it closer to the real-life situations.

This study affirms that health centers or even government offices in rural areas need a program or application of software that can predict NCDs. It was necessary to have a huge amount of raw data to have proper classification rules which lead to appropriate prediction of NCDs. The forecast is helpful not only for physicians but also for neighborhood patients to test their health on constant basis. The system built must be constantly reviewed, and the programming should be optimized so that the technology can also be applied through several health centers and clinics.

For this study, the proponents would like to make the following recommendations to further improve the project. First is the use of additional medical sensors, this will increase the accuracy and sensitivity of the device. It is also recommended to use additional types of diseases for a wider scope which will make the project more useful. Thirdly, the researchers recommend that future researchers develop an application that will easily gather the readings of sensors and saving the file with .csv file format without the help of terminal emulation application. And lastly, the proponents would like to recommend increasing the number of samples to be validated since it affects the percentage accuracy of the device and the laboratory testing.

REFERENCES

[1] "Population of the Philippines (2019 and historical)," 2019. Available: https://www.worldometers.info/world-population/philippines-population/.

[2] P. L. Bersales, Deaths in the Philippines, 2016, Available: https://psa.gov.ph/content/deaths-philippines-2016.

[3] World Health Organization, "Non-communicable Diseases," 2018. Available: https://www.who.int/news-room/fact-sheets/detail/noncom municable-diseases.

[4] World Health Organization, "Cardiovascular Diseases (CVDs)", 2017. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovasc ular-diseases-(cvds).

[5] A. Mata and G. Jasul Jr, "Prevalence of Metabolic Syndrome and its Individual Features Across Different (Normal, Overweight, Pre-Obese and Obese) Body Mass Index (BMI) Categories in a Tertiary Hospital in the Philippines," Journal of the ASEAN Federation of Endocrine Societies, vol. 32, no. 2, pp. 117-117, 2017.

[6] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1275-1278, 2018.

[7] A. Cuello, K. Dizon, M. L. Fortu, J. G. Paguia, A. V. Tañola, J. Velasco, A. Aquino, and I. Valenzuela, "Blood Glucose Analysis Utilizing Artificial Raindrop Algorithm and Capacitance Method," in 2017 International Conference on Computer and Applications (ICCA), pp. 135-139, 2017.

[8] A. A. Jambora, "For Filipinos, high blood pressure guideline remains 140/90, not the US' 130/80," 2018. Available: https://lifestyle.inquirer.net/295738/filipinos-high-blood-pressure-guideline-remains-140-90-not-us-130-80/.

[9] Marcin, J., and Holland, K. (2017). Is My Blood Oxygen Level Normal? . [online] Available at: http://www.healthline.com/health/normal-blood-oxygen-level#oxygen-levels3.

[10] M. Ong, "Health numbers you need to know," 2016. Available: https://www.philstar.com/lifestyle/health-and-family/2016/11/22/1644 371/health-numbers-you-need-know.

[11] World Health Organization, Technical Report #727, 1985; NCEP Report, 1998.

[12] C. Jimeno, "A Summary of the Philippines UNITE for Diabetes Clinical Practice Guidelines for the Diagnosis and Management of Diabetes (Part I: Screening and Diagnosis of DM)," 2011.

[13] T. Chakravorty, "How Machine Learning Works: An Overview," 2016.

[14] D. Team, "Introduction to Artificial Neural Network Model," 2018.

[15] D. Storey, "Random Forests, Decision Trees, and Ensemble Methods Explained," N.D.

[16] N. Bambrick, "Support Vector Machines: A Simple Explanation," 2016

[17] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, no. 01, 2017.

[18] G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine et al., "An evaluation of machine-learning methods for predicting pneumonia mortality,"Artificial intelligence in medicine, vol. 9, no. 2, pp. 107-138, 1997.

[19] I. M. Shofi, L. K. Wardhani, and G. Anisa, "Android application for diagnosing general symptoms of disease using forward chaining method," in 2016 4th International Conference on Cyber and IT Service Management, pp. 1-7, 2016.

[20] Q. Zhang, D. Zhou, and X. Zeng, "Hear the heart: Daily cardiac health monitoring using Ear-ECG and machine learning," in 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), pp. 448-451, 2017.

[21] H. Yin and N. K. Jha, "A health decision support system for disease diagnosis based on wearable medical sensors and machine learning ensembles," IEEE Transactions on Multi-Scale Computing Systems, vol. 3, no. 4, pp. 228-241, 2017.

[22] B. D. Kanchan and M. M. Kishor, "Study of machine learning algorithms for special disease prediction using principal of component analysis," in 2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC), pp. 5-10, 2016.

[23] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of cardiovascular disease using machine learning algorithms," in 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1-7, 2018.

[24] L. K. Tolentino, R. M. Aragon, W. R. Tibayan, A. Alvisor, P. G. Palisoc, and G. Terte, "Detection of Circulatory Diseases Through Fingernails Using Artificial Neural Network," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 1-4, pp. 181-188, 2018.

[25] R. J. Jorda, K. E. Bartolome, C. Buban, M. A. De Los Reyes, A. Maingat, M. J. Oledan, and L. K. Tolentino, "Ovulation hormonal imbalance recognition system using saliva's conductivity analysis based on potentiometric method," Indonesian Journal of Electrical Engineering and Computer Science, vol. 16, no. 1, pp. 176-183, 2019.

Lean IT Transformation Plan for Information Systems Development

Muhammad K. A. Kiram[1]
[1] TNB Research Sdn. Bhd.
43000 Kajang, Selangor, Malaysia

Maryati Mohd Yusof[2]
[2]Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia
Bangi, Malaysia

Abstract—**Information systems development (ISD) is prone to failure, which can be defined as a time-consuming and costly phenomenon that provides value that is not directly appealing to clients. While ISD can be enhanced using various tools, models, and frameworks, failures related to ISD remain to be evident and costly. These failures are related to human, organizational, and technological factors and waste in ISD. This study identifies the information system (IS) success criteria and factors that contribute to ISD waste. A qualitative case study was conducted for an ICT research unit by using interview, observation, and document analysis techniques as a means of analyzing the IS success criteria, leanness level, and waste. Findings show that lean IT approaches and IS success criteria can be combined to develop a holistic transformation plan for organizational ISD. This transformation plan can potentially assist IS developers deliver high-value IS while driving organizational growth towards the fourth industrial revolution.**

Keywords—Failure; lean IT; information systems; information systems development; socio-technical; waste

I. Introduction

Information systems development (ISD) has emerged with cutting edge technologies and dynamic business processes as a means of supporting organizational performance. However, an ISD is prone to failures in terms of project overruns, over-budgeting, and unfulfilled user requirements [1]. An ISD is more likely to generate waste if it is not planned and managed appropriately [2]. In addition, the collaborative, complex, and subjective nature of ISD renders it vulnerable to producing waste [3,4], such as partially done work, unnecessary processes, extra features, handoffs, delays, and motions and errors, despite the introduction of various methodologies and tools [5]. Therefore, evaluating the IS success factors and identifying their root cause of waste in ISD by using the quality improvement approach, such as lean IT, are important in enhancing the quality of the IS. ISD interacts with three socio-technical factors (organization, technology, and human) in producing IS [6]. Despite the potential of IS to increase organizational effectiveness and efficiency, IS failure remains to be a widespread concern [7] in terms of the technical, project management, quality, and human aspects [4].

IS failure is commonly associated with various socio-technical factors and ISD waste. Inaccurate development of product features, poor work management, rework, unnecessary problem solving, cognitive over-burden, psychological stress, delays, relearning, and poor communication are examples of waste identified in ISD [4]. The longer the delay of the waste removal, the higher the cost to be incurred by an organization [8]. Consequently, this study attempts to identify the root cause of ISD waste in the quest of producing high-quality IS, in which the lean IT method and an evaluation are both carried out to achieve IS success/efficiency from the socio-technical perspective.

The paper is organized as follows. The following section provides an overview of IS success criteria and the Lean IT approach. Section 3 describes the case study method, followed by the case study findings that are presented in Section 4 based on ISD success criteria, Leanness level, waste, and document analysis. Section 5 discusses the findings based on their subsections and the Transformation Plan for lean IT in ISD organization. The final section concludes the study.

II. Related Works

A. IS Success Criteria

Despite its benefits, IS is also costly and risky. Therefore, IS developers need to understand the critical success factors of IS [9]. From the technical perspective, unclear and complex requirement specifications can affect the planning phase, rippling further through the ISD activities [10,11] in software development lifecycle (SDLC) phases [12]. Different perceptions and mental models among clients and IS developers during requirement elicitation and design can create conflict and subsequently weaken the teamwork spirit [13,14]. Limited technology expertise can also contribute to failure, in which the IS cannot meet the required features [4,13]. Furthermore, limited time and focus on IS testing can affect quality and generate errors during early system operation [8]. Poor system quality has been reported in terms of inflexibility [15], low functionality [16], irrelevance with requirements [4], and low usefulness in supporting user-centered design [17].

ISD failure in project management is associated with slow progress, limited funds to hire experts [15], scope creep and poor risk control [10], poor communication, knowledge management, and implementation methodology [4], and poor talent management [18,19]. Scope creep can lead to rework, subsequent delays, and additional cost [4]. An IS that is halted upon completion may result in exorbitant waste [16]. Many human factors, including low productivity [3], limited technical expertise [15], emotional problems [20], low teamwork spirit [13], and low user involvement [21], lead to ISD failure. IS developers are frequently blamed even if the problems are

caused by the user's unclear requirement specifications and business processes [22]. Human limitations, such as inability to work concurrently and overburdened mental and psychological ability, also hinder ISD success [4].

An evaluation framework known as human, organization, and technology fit (HOT-fit) can be used to specify IS success/efficiency based on the abovementioned three factors and their dimensions, namely, system quality, information quality, service quality, system development, system use, user satisfaction, organizational structure and environment, and net benefits [6,23]. A stable organizational structure ensures that IS project development can be completed on time without drastic changes. Organizational settings and culture, such as good political situation and the absence of bureaucracy, also contribute to IS success.

B. Lean IT Approach

Lean evolution in IT, known as lean IT, encompasses the following five principles similar to those in lean manufacturing: value identification, map value flow, construct flow, supply based on demand, and pursuit for perfection [24]. Lean IT is defined as a holistic management approach in IT organizations, in which continuous improvement is attained by eliminating waste and unevenness by using the lean philosophy, principle, and tools [25,26]. The literature on lean IT is scarce despite its similar principles with manufacturing [25-28]. Moreover, the implementation of lean IT remains to be at its infancy compared with those in other industries [28,29], particularly in IT management [26]. The lean IT approach is also vague due to missing principles or suitable lean activities [30].

Waste in Lean IT. Waste can be defined as any activity that does not add value to products or services or does not fulfill customer requirements [24]. Waste is categorized as overproduction, delay, defect, transportation, inventory, motion, and over-processing [22]. Waste in the lean IT domain has been adopted from the original waste definition in the context of IT management or IT support activities. Most literature on IT waste is related to ISD [4,5].

Lean IT Tools. Various lean tools, including just-in-time, automation (Jidoka), leveling (Heijunka), continuous improvement (Kaizen) and standardized work, 5S, Kanban, Takt Time, Andon, and 5-whys, can be used at different lean transformation stages [19]. Common lean IT tools usually entail virtual Kanban, automation, Kaizen events, and scrum meetings in the ISD context.

Lean IT in ISD. Waste in ISD is inevitable, and it affects cost, schedule, scope, and project survival. The literature reports various waste types, including those in the ISD context [4,5,31]. Waste in ISD includes excessive IS features, relearning due to lost knowledge, partially done work, handoffs, task switching, delays, and defects [32]. Table I illustrates the relationship between the ISD failure issue and potential waste. The seven principles of lean software development (LSD), namely, waste elimination, build-in quality, knowledge creation, commitment deferment, fast delivery, respect for people, and optimize-the-whole approach [33], can be extended to organizational ISD.

TABLE I.   RELATIONSHIP BETWEEN ISD FAILURE AND WASTE IN SDLC

SDLC phases	ISD failure issue	Waste types
Requirement elicitation	Lack of involvement among IS developer, user and stakeholders	Extra features, delay
	Poor scope control and constantly changed requirements	Extra features, partially done work, delay
	Difficulty in understanding tacit requirement meaning	
	Poor documentation management	
Feasibility study	Poor evaluation of project resources availability such as financial and expertise	Partially done work, relearning, delay, defect
	Risk control and change management are not implemented during the early phase of development	Extra features, partially done work, delay, defect
Development	Slow project progress due to lack of technology expertise	Delay, relearning
	Loss of expertise with no replacement	Relearning, task switching, handoff
	Recurring rework in the debugging process	Defect, delay
	Unnecessary feature development	Extra features, partially done work, delay, defect
	Complex and irrelevant solution	Extra features, delay
	Pressure and burden on physical and mental ability	Partially done work, task switching, handoff
	Excessive emotional and psychological stress	
	Inability to implement concurrent work	
	Poor morale and teamwork	
	Poor knowledge and communication sharing	Relearning, handoff, task switching
Implementation	Incompatibility between IS and IT infrastructure	Partially done work, defect
	Difficulty in integrating new and current IS (or legacy system)	
	Lack of involvement during training session (transfer of technology)	Defect
Maintenance	Poor customer service management	Partially done work, defect
	No maintenance manual	

III. METHOD

We conducted a qualitative exploratory case study among IS developers from different organizations to analyze the root cause of waste in ISD. We employed interview, observation, and document analysis techniques to collect data. The informants were selected via purposeful sampling. Most informants are full-stack developers who are highly involved in system development activities and skillful in database, network, and system management. The informants were also selected on the basis of their ability to answer the research questions. One-on-one and group semi-structured interviews were conducted with four- and three-pair informants among the ten IS developers with different backgrounds in organizational,

project, designation, role, and work experience (Table II). The document analysis included the use of physical documents, such as meeting minutes, reports, software requirement specifications, user acceptance tests (UATs), change request forms, and user manuals. Digital documents included e-mails, system architectures and designs, storyboards and source codes.

TABLE II.   INFORMANT LIST

Code	Designation	Experience (Year)	Organization	Status Project
R1	System analyst	2	Research	Successful
R2	System analyst	2	Research	Successful
R3	IS developer	3	System development	Fail
R4	IS developer	3	System development	Fail
R5	System analyst	5	Research	Successful
R6	IS developer	4	Research	Successful
R7	IS developer	5	Research	Successful
R8	IS developer	3	Research	Successful
R9	IS developer	2	Research	Successful
R10	IS developer	2	Research	Successful

We employed thematic and inductive analyses [34]. The data were analyzed to discover and highlight common phrases or keywords and subsequently identify patterns or themes. Upon identifying certain themes, the coded data were compared with the non-coded ones to verify their accurate meanings and their correlation with the research question. The final themes were refined based on the discovered discrepancies or new themes. First, we analyzed the ISD success criteria from the IS developers' perspective based on the HOT-fit framework [6,22] to understand their holistic views on successful IS. Second, we analyzed the ISD leanness level to evaluate the extent in which IS developers apply basic lean principles in an organization. Third, we analyzed waste in ISD based on the waste category [32]. The analyses were then combined, and a unique transformation plan for ISD was recommended for the case.

IV. RESULTS

The case study was conducted at an ICT Research Unit (IRU) of a Malaysian electricity company. The IRU works dynamically in various IT projects, including system development, management of IT infrastructure and networks, multimedia assignments, digital arts, IT asset acquisition, and policy monitoring. Several large-scale projects used to require IRU to hire third-party experts among vendors, interns, or freelance workers. The ISD managed by the IRU can be categorized as R&D project support, support service, and laboratory services. The IRU, as a non-profit organization, develops IT infrastructures and ISs by using open-source platform resources, in which internal expertise is utilized, to save operational cost. IS developer group used two main ISD methods, namely Agile and SDLC traditional phases.

A. ISD Success Criteria

The ISD success factors in this study were identified from the IS developer perspective (Table III) based on the nine main components of the HOT-fit framework. Fig. 1 illustrates the ISD success criteria factors and their relationships at IRU.

System quality. IS success is influenced by the fulfillment of the user requirement, user interface design, and third-party feedback. The fulfillment of user requirements refers to the extent in which the system can successfully meet the user's finalized business process that is free of bugs upon testing. A system is considered successful if it can be completed according to plan and subsequently used by the user (R1). Moreover, the design should entail an interface that can be easily adopted by users of all age ranges and learned independently given its minimalist and attractive features (R10). System quality can also be measured on the basis of the feedback from a third party independent from the IS team. "Positive feedback and recommendations from an external party who test the system are considered for system improvement" (R3). In general, most informants attest to the importance of system quality; they acknowledge it as a critical success factor, and errors and usability issues should be continuously addressed.

Information quality is associated with secure data storage, information processing accuracy and efficiency, and data security. Data security requires data aggregation and analysis to protect the system from any threat (R3). Service quality is focused on change request and errors. IS success is defined by the level of service support (1) for the user upon IS completion and 2) for fulfilling the requirements of system and business changes. Support provision to address errors is related to finding solutions to semantic or syntax errors in all aspects of system use. IS developers must be prepared to provide support to user request (R6).

TABLE III.   ISD SUCCESS CRITERIA AT IRU

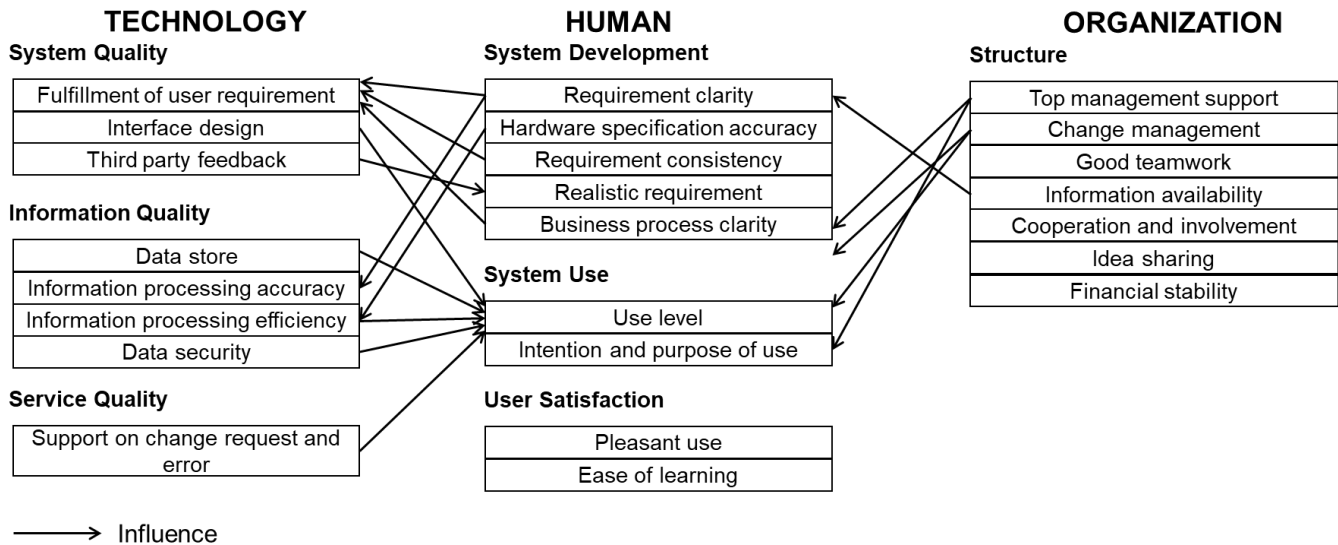| Success factor	Success criteria
Technology	
System quality	fulfillment of user requirement, user interface design, and third-party feedback
Information quality	data storage; information processing accuracy and efficiency; data security
Service quality	support on change request and error
human	
System development	requirement clarity; accuracy of hardware specification; requirement consistency; realistic requirement; clarity of business process
System use	use level; intention and purpose of use
User satisfaction	pleasant use; ease of learning
organization	
Organizational structure	management support; change management; teamwork information availability; cooperation and involvement idea sharing; financial stability
Net benefits	support vision and mission alignment support operational planning

TECHNOLOGY   **HUMAN**   **ORGANIZATION**

System Quality

| Fulfillment of user requirement |
| Interface design |
| Third party feedback |

Information Quality

| Data store |
| Information processing accuracy |
| Information processing efficiency |
| Data security |

Service Quality

| Support on change request and error |

System Development

| Requirement clarity |
| Hardware specification accuracy |
| Requirement consistency |
| Realistic requirement |
| Business process clarity |

System Use

| Use level |
| Intention and purpose of use |

User Satisfaction

| Pleasant use |
| Ease of learning |

Structure

| Top management support |
| Change management |
| Good teamwork |
| Information availability |
| Cooperation and involvement |
| Idea sharing |
| Financial stability |

⟶ Influence

Fig 1.   ISD Success Criteria Factors and their Relationships at IRU.

System development factors that influence ISD include requirement clarity and accuracy, hardware specification requirements, realistic requirements, and business process clarity. Clarity is important in ensuring that the IS will be developed according to user requirements in terms of the functional, nonfunctional, and user experience aspects. "Basic function, actual system purpose, accurate definition of main functions, and additional functions to support system efficiency must also be defined" (R2). Consistency refers to maintained, instead of constantly changed requirements that are aligned with the original ISD project scope (R8).

The accuracy of hardware specification is also important for system development in terms of cost-saving targets, compatibility with the developed IS platform, and avoidance of waste, such as the development of a sophisticated but unnecessary hardware for a simple system (R9). The realistic requirement refers to feasible requirements that can be implemented using available technologies, cost, time, and IS developers' skills. This realistic requirement depends on finalizing a business process instead of simply relying on the expectation and imagination of certain stakeholders (R4). In addition, a clear and finalized business process that is commonly understood by all involved units is critical in avoiding confusion during system development (R3).

System use measures the level, intention, and purpose of use. A high usage level refers to the optimized use of system modules and features (R2). Intention and purpose are also important; without them, ISD becomes a waste (R5). Meanwhile, user satisfaction can be measured on the basis of user experience, including system pleasantness and ease of learning. A successful system is also associated with its attractive, intuitive, and interactive features.

The factors related to organizational structure were highly cited among informants. Management support plays an important role in ensuring the enterprise-wide use of IS and its fit with the organization's mission and vision, thus facilitating stakeholder involvement, planning, and implementation as a

means of managing changes (R4) and expediting processes that require approval, such as process redesign (R5). A conducive organizational setting and the freedom to share ideas can yield the much needed harmonious ambience among IS developers. Moreover, the ISD team members need to ensure cooperation and support one another to successfully attain their objectives. "Each team member needs to play their respective roles and facilitate one another's task to achieve the [system] objective" (R10). Financial stability is also critical in ensuring ISD survival; that is, without being affected by adjustment risks or budget limitations, as these concerns affect overhead expenses or hardware acquisition (R2). The organization must constantly be apprised of the information required by IS developers throughout the project. Net benefits refer to the overall benefits obtained by organization. The unit strategic plan should be aligned with that of organization, and the ISD should be able to facilitate the management task of organization during operational planning (R6).

B. Leanness Level of ISD

The analysis of the leanness level or lean compliance in the ISD context is imperative in examining the current business process. The current practices, which are contrasted with the LSD principles for benchmarking purposes [1], are summarized in Table IV and discussed further as follows.

Waste elimination. The root cause of waste is identified and removed using lean tools. IS developers remove waste based on the understanding of the process value and user problems. Focus is given on system functionality to optimize resources. "Problems and requirements are specified before system development; problems that genuinely require a system [as a solution] indicate a valuable task" (R1). The current practice strives to constantly comply and agree with the original project scope and planned schedule (R4) to avoid scope creep and rework due to uncontrollable user requirements. Organizational priorities pull instead of push production, in which ISD is only initiated on the basis of demand (R1). The IRU unit organizes case or scenario presentations related to system use to increase

the team members' understanding and facilitate the acquisition requirement process. Waste in task changes can be minimized by postponing the ad hoc task until all current tasks have been completed. The *build-in quality* principle refers to the IS developers' assurance regarding the absence of system flaws that can affect the system's functionality. Some of the organizational practice for improving system quality is to always understand and fulfill user expectations by obtaining their feedback during system operation, such as "asking users to describe their experience or expectation from the system" (R2). In a workgroup environment, the use of work collaboration software, such as *Sourcetree*, may enhance ISD quality (R3).

Knowledge creation involves the gathering of knowledge in the course of project development for reference purposes. Some common practices include studying problems and mastering problem-solving skills (R1), generating and disseminating knowledge in meetings, mentoring sessions, and weekly or biweekly discussions (R2), and writing code snippet notes that can be reused and shared (R7). When IS developers are stuck with work and cannot generate new ideas, knowledge can be generated by searching the Internet or consulting other experts (R5). Knowledge is also generated from online learning portals, such as Udemy and Coursera, subscribed by the IT manager for subordinates (R6).

Deferred commitment means postponing decision making to the last minute; in this manner, certain ISD modules can be initiated without having to finalize the system requirements. IS developers can apply this principle by synthesizing information and drafting solutions until the actual solution is discovered (R2).

TABLE IV.   LEANNESS LEVEL OF ISD AT IRU

LSD principles	Organizational practices
Eliminate waste	Understand value and problem statement
	Ensure adherence to the original scope and agreement with it
	Develop system based on requirement
	Present use-case during requirement elicitation
	Postpone ad-hoc task until current tasks are completed
Build quality in	Understand and fulfill end user expectation
	Use work collaboration software
Create knowledge	Revise issue and be skillful in problem solving
	Share knowledge during meeting, mentoring, and discussion sessions
	Create and share code snippet
	Refer to experts
	Subscribe online training
Defer commitment	Postpone decision making until solution is discovered
Immediate result	Prioritize and quickly completed functional requirement
Respect for people	Avoid arguments and confrontation
	Establish effective communication
	Support one another
	Understand team members' limitations
Optimize the whole	Understand overall problem

Fast delivery refers to the development of the IS as soon as possible without compromising its quality and cost based on the ability of the IS developers. This principle adheres to the prioritization and completion of the functional requirement in a quick manner. "My to-do list is based on priority and completion targets" (R8).

Respect for others is emphasized in communication and activities among IS developers. Arguments and confrontation, despite the different specializations and backgrounds, should be avoided. For example, developers can describe their task complexity to other members who are unfamiliar with specific ISD processes (R2). Respect is also cultivated through effective communication, such as providing opportunities for IS developers to provide suggestions and ideas independently to users during discussions (R4). Helping one another based on abilities, regardless of role and position, in each ISD phase indicate respect. Developers can divide tasks based on the requirement analysis to ensure that they work in the same direction. "When we obtain our module, we divide the task, aid in unfinished job, and ensure that other programmers' codes are in line with ours" (R6). Moreover, the recognition of strengths and weaknesses of each team member can lead to common understanding and a harmonious working ambience (R2).

The *optimize-the-whole* principle views the process holistically. On this basis, IS developers can understand problem as a whole. Moreover, the developers can propose the best solution based on team capabilities and enhance the algorithm and system design system. "We seek to view from wider and different angles in consideration of the solution alternatives" (R2).

C. Waste in ISD

Waste in ISD, which has been analyzed prior the identification of the root cause based on the waste categories [1], is listed Table V. *Extra features* include those items beyond the project scope and are not specified by the user, "but they could have overlooked them" (R1), which do not add value to the IS developer team. These wastes are also caused by inappropriate technology selection, hardware and software acquisition strategies (R6), and unclear business processes, such as the absence of a valid purpose for process improvement or automation. Unclear requirement specification is a common scenario attributed to user difficulty in the finalization of the system requirements (R4).

Relearning refers to the acquisition of the same or previously acquired knowledge. Knowledge loss are costly in terms of skill, expense, and time. This type of waste is also due to the lack of experts or error reference sources online for guidance purposes (R1) and prolonged abandonment or non-use of knowledge. The knowledge obtained from previous training and projects were only stored but not referred as they are not needed. However, stored knowledge may later become irrelevant to the current requirements due to technology advancements. New bug and error discoveries can be categorized as new knowledge, but no specific platform can be used to store and disseminate them, resulting in limited reference for encountered errors. IS developers who focus on specific ISD tasks over long durations tend to forget about their

other existing knowledge. "I have been focusing on development activities for so long and have forgotten how to solve the same encountered issue [in other aspects]" (R4).

Partially done work refers to work that is postponed and "in progress" for too long due to certain reasons. This waste is attributed to external disruptions affecting the IS developers who focus on the ISD (R7) and deadlocks or lack of ideas in solving problems, such as an encountered new bug and the low commitment to execute concurrent tasks. "[Solving] a new bug or problem consumes time and delays my work (R1). As a result, IS developers usually focus on easier tasks, leaving the unsolved task, a partially work done.

TABLE V.   WASTE ROOT CAUSE IN ISD PROJECT AT IRU

Waste type	Waste root cause
Extra features	Features beyond scope and not required
	Inappropriate selection of technology, hardware and software acquisition
	Unclear requirement specification
	Unclear business process – no actual need for process improvement
Relearning	Unavailable error reference
	Prolonged unused knowledge
Partially done work	External disruption
	Stuck or out of idea
Handoff	Transfer of IT infrastructure to other party
	Sudden resignation
	Module based outsourcing
	Sudden termination of IS developer
Task switching	Significant focus disruption
Delay	Time needed to regain focus
	Waiting for endorsement or approval
	Waiting for commitment or user involvement
	Lengthy meeting
	Fixing newly discovered error
	Time consumed to understand source code written by other developer
	Emotional disruption
Defect	Rework due to uncontrolled change request

Handoff refers to a handover work. In IRU, this waste stems from handing over IT infrastructure to external parties. Handoff occurs immediately upon module completion. An appointed vendor is expected to hand over the task during a technology transfer session with a IS developer. However, the developer is likely unable to completely understand the session, requiring him or her to reexamine the vendor's work. The process requires a compliance to and an understanding of the company's standard operational procedure, terms, and conditions. This approach is time consuming compared with the direct handing over of infrastructure to the internal IT management (R9). Waste can also be caused by the sudden resignation or termination of members of the IS developer team

but without leaving adequate handoff documentation. "When a member suddenly resigns, I assume that at least half of the module had been prepared, but it was [actually] less than that, so I had to [re]do it from scratch" (R3). Handoff waste can thus be caused by incomplete outsourced modules that had not been developed on the basis of the specified requirements (R3). The immediate resignation of team members results in stress and additional workload to other members due to unavailable handoff documentation (R10). Transferring IT infrastructure to another party also results in a handoff. IS developers should provide comprehensive documentation based on specific standards to ensure that the system can function well in the new IT infrastructure.

Task switching refers to changing one task to another, causing an IS developer to lose the critical focus required in the ISD. Disrupted focus affects the quality of a module-in-progress due to the time needed to regain focus and emotional stability and address possible confusion and lapses. "System development requires a hundred percent focus. A high-quality system needs to be developed [only] one at a time" (R1).

Delay, the most commonly cited type of waste, requires waiting for a process to be completed before proceeding to the next one. Delay is caused by the time needed to regain the disrupted focus of the IS developer due to his or her disorganized thinking and task switching (R2). Delay can also be attributed to the time wasted resulting from authorization or approval from another party, such as those related to infrastructure or work orders (R7). Unavailable instruction from the top management also delays commitment, cooperation, and user involvement in the requirement elicitation, testing, and system use phases. "My biggest problem is to obtain user commitment to use the system" (R2). In addition, some automation protocols in the business process requires changes in standard documentation, such as ISO, which are controlled by the top management. Lengthy meetings also waste time and cause delay (R1). Encountering a new error results in delays, as the IS developer usually takes time to understand the cause of error (R9). However, commonly encountered errors can be fixed faster than other types of errors through IS developer experience. In the handoff scenario, delays occur when the IS developer takes too much time to understand the source codes developed by another individual (R10). This waste may also stem from emotional disruption, which is closely related to task switching (R1).

Defect refers to resources that are repeatedly used to improve a developed IS. In this case, defect is caused by additional requirements that are significantly diverted from the original scope (R2). Therefore, the developed features need to be removed in spite of the consumed cost and time. This waste may also stem from uncontrolled and improper management of change, as requested by the project manager. "A user requests for a design during the first meeting. In the second meeting, the user requests for a requirement change. It is a major requirement, so I feel that the first design is totally wasted" (R1).

D. Document Analysis

Table VI summarizes the results of our document analysis. Most projects have properly documented the agreed system

specification requirements for the sake of the IS developer and the user. The different waste types, including constantly changed business processes, ambiguous requirements, lack of details and confusing documentation, inaccurate business process modeling, inability to consider user satisfaction, absence of risk planning and change management plan, and poor management of document version, have been identified.

In particular, UAT documents can help identify weaknesses in the documentation of user feedback obtained during UAT sessions. Moreover, IS developers may be affected by the limited time to perform unit testing. The analysis of change request forms can present weaknesses in terms of managing the change request, sophisticated requirement, and unrealistic time to address changes. Poorly written user manuals may manifest other weaknesses that are unable to guide in solving issues.

TABLE VI.   DOCUMENT ANALYSIS IN ISD AT IRU INFORMANT LIST

Document type	Findings
System requirement specification	Constantly changed business process
	Overly vague requirement specification
	Lack of details in producing document requirements
	Inaccurate business process modelling
	Not considering user satisfaction
	Unavailable risk planning
	Unavailable change management
	Weak management of document version
User acceptance test	Poor user feedback documentation
	Lack of time to perform unit testing
Change request form	Poor management of change request
	Sophisticated change requirement
	Unrealistic time duration to address request
Manual User	Unhelpful in solving issue
	Poorly written

V. DISCUSSION

A. ISD Success Criteria

The IS developer team is also committed in delivering the best results in each ISD phase, proving that they are enthusiastic in producing quality IS that can add value to the organization. Meanwhile, the accurate processing of information is highly emphasized, as most in-house developed ISs require their own logic and formula. Inaccurate information can jeopardize organizational credibility in generating reports, which are eventually shared to the third party. As for service quality, support service, which includes warranty, service level agreement, or change request, may be changed by formally submitting a request form via manual, e-mail, phone, and during meetings to ensure that the IS can be continuously developed according to the organization's requirement. In system development, the clarity requirement in the IRU context refers to the current business process that needs to be automated through IS. A clearly defined business process can minimize waste risk in terms of rework or development of

unnecessary extra features. This requirement should also be maintained by the developers while finalizing the work scope with the users even before development phase is implemented. Inconsistent requirements are expected to challenge the original project plan, particularly the management of the work schedule.

The system's usage level refers to the extent in which the IS can be used to support the business process. In this study, the factors that discourage system use include system sophistication and poor user interface design. We learned that some ISs were only developed to fulfill departmental aspirations in the organizational digital transformation plan but with no real intention to use the system. A user can be regarded highly satisfied if he or she can use the system pleasantly, proving that ISs should possess high usefulness and able to attract user interest to learn and understand the system. Many interface frameworks, such as bootstraps or material design, can be used to design ISs according to industrial best practices.

Top management support plays an important role in determining IS success, given the authority in decision making that they provide for resource planning and influencing the sponsorship of the ISD. Improper management decisions developers. The management should also ensure that the ISD team remains unchanged during system development to facilitate smooth communication, cooperation, and information acquisition and ensure that work synchronization can be managed optimally. The collaborative nature of ISD projects requires a good team, high commitment among team members, and alignment between the IS developer and the user. IS developers are responsible in ensuring the fulfillment of user requirements; by contrast, users are accountable in finalizing the information, business process, and requirements for the IS developers. Most informants have encountered problems in adhering to project schedule due to the constantly changing user requirements, which result in waste, such as delays and rework. In addition, a user may encounter problems in providing accurate information for the system requirement due to poor information availability and inefficient idea sharing. Nonetheless, many informants argue that a well-controlled and managed change management can lead to IS success. A conducive work environment can be established by encouraging good rapport among team members.

Organizations can move towards the fourth industrial revolution if they focus on adapting IT in their business models and daily operations. Thus, developing an IS in line with the organizational vision and mission can add value and benefits, and the organization can manage its resources effectively and efficiently. A successful IS also supports operational planning. Technologies, such as big data analytics, can assist the organization predict organizational resources accurately and support the management in making decisions effectively.

B. Leanness Level of ISD

Leanness level or compliance to lean IT principles is crucial in understanding the current ISD situation. Although most informants have no specific knowledge on LSD principles, their work practices are aligned with most of the abovementioned principles. In waste elimination, although the developers have attempted to maintain the original project

scope, non-finalized and constantly changing business process can result in scope creep. However, the IS developers become highly motivated with their work upon understanding the values that can be achieved through their respective ISD specializations, thus indirectly eliminating waste in terms of partially done work and delay. From the build-in quality aspect, feedback on IS improvement are gathered not only through users but also from all parties who use the system. This effort is encouraged by offering system updates or improvements to users. Work collaboration software, such as Sourcetree and Jira, are also used to expedite source code generation, ease the work planning, and ensure the creation of non-conflicting codes by multiple IS developers. In terms of knowledge creation, despite the good sharing practices, some of the knowledge are lost upon acquisition due to the lack of proper documentation or the absence of a knowledge management system.

Postponing decisions enable users and developers to gather more information, including risks, to avoid rework resulting from requirement changes. In terms of fast delivery, prioritization can be achieved by considering the system module, functional (instead of non-functional) requirement, and user experience. Some functional requirements consume time due to their logic or calculation complexity. Respect for others can cultivate a harmonious group and a conducive work environment that encourage freedom of speech in sharing ideas and constructive comments, motivate the IS developer team, and avoid stress that affect work productivity. Finally, the optimize-the-whole practice enables an enterprise to seek wide buy-in solutions.

C. Waste in ISD

Waste in ISD must be identified prior the finalization of the transformation plan. Fig. 2 summarizes the root cause of ISD IS developers can apply the *Plan, Do, Check,* and *Act* (PDCA) cycle throughout the ISD to enhance its quality. In the *Plan* phase, the problem is examined to identify the best method to fulfill the user requirement. Shortlisted methods are tested in small scale to measure their effectiveness during the *Do* phase. In the *Check* phase, the obtained information is analyzed to determine the feasibility of the selected method in solving the problem. If successful, the tasks will be implemented in the *Act* phase. Moreover, the quality of the whole IS function can be enhanced using *Poka-Yoke*, which means to prevent mistakes and identify and avoid errors. Applying these elements during the ISD can ensure IS quality, subsequently avoiding various inefficiencies, such as errors and rework.at IRU. Delay and defect are the most significant waste categories at IRU. Table VII summarizes the relationship between ISD failure at IRU and waste type [2].

IS developers can choose different methods to obtain knowledge, but this knowledge may not be managed, stored, and disseminated properly. An online- and cross-boundary knowledge sharing [2] portal enables knowledge to be stored in various forms and shared among IS developers intra and inter organization. In disseminating knowledge, the experienced and trained IS developers can provide training or tutorial sessions based on their expertise. A project manager may seek to postpone his or her decision making until a solution is discovered, while an IS developer should not make a decision hastily without scrutinizing the existing resources.
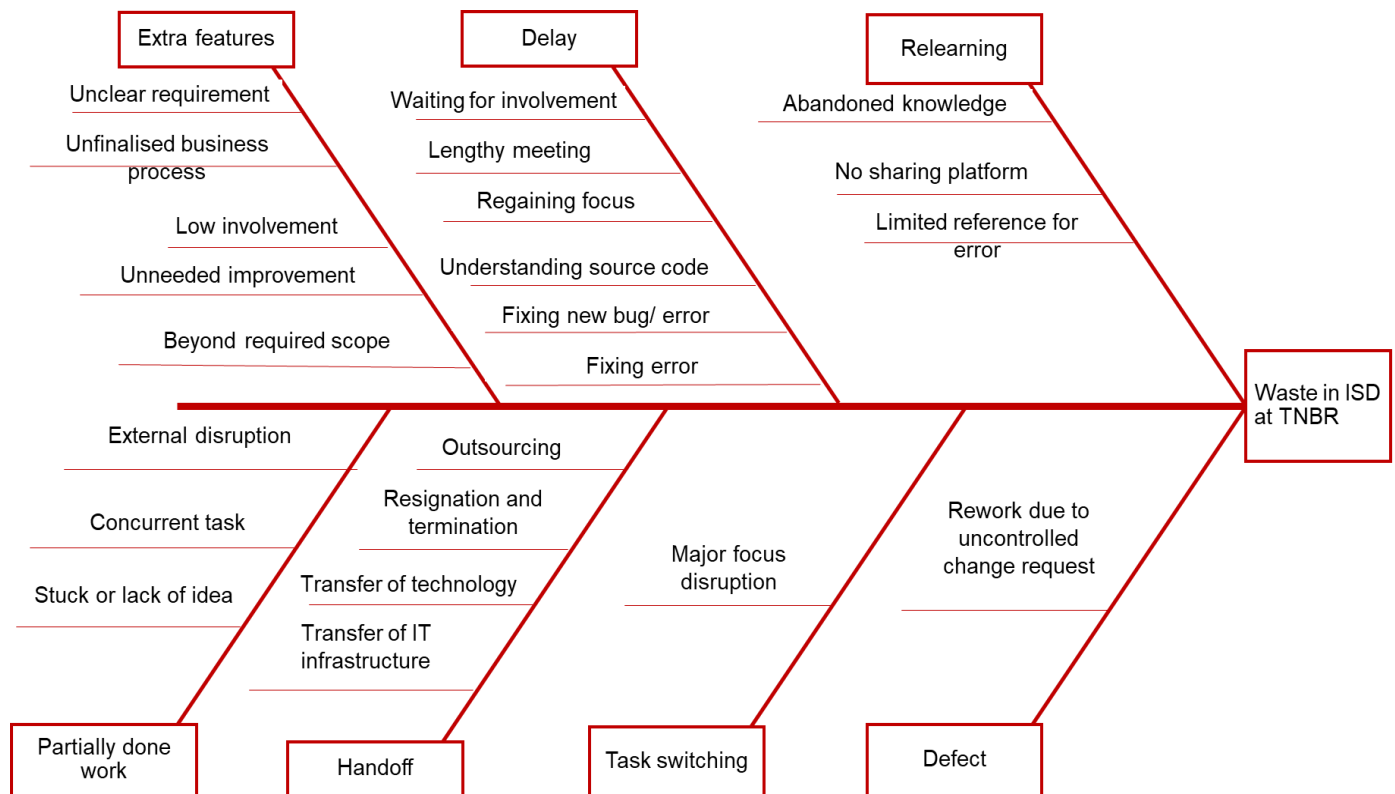

Fig 2.   Fishbone Diagram for Waste Root Cause in ISD at IRU.

TABLE VII.   RELATIONSHIP BETWEEN ISD FAILURE AND WASTE TYPE IN
SDLC PHASES AT IRU

SDLC phases	Failure issues of ISD at IRU	Waste types
Requirement elicitation	Unclear requirement specification	Extra features
	Unable to finalize business process	
Analysis	Vague business process and no valid reason for improvement or automation	Extra features
Development	Develop features that are out of scope, extra, and excessive	
	Knowledge that is abandoned and stored but not referred to	Relearning
	Unavailable specific plaform for storing and disseminating knowledge	
	Limited reference to encountered error	
	External factor disruption	Partially done work
	Concurrent work	
	Out or lack of idea	
	Module based outsourcing	Handoff
	Sudden and unplanned resignation and termination of the IS developer team	
	Time needed to regain lost focus	Delay
	Rework due to uncontrolled change request	Defect
	Significant focus disruption	Task switching
Testing	Time to understand third-party source code	Delay
Implementation	Transfer of technology upon a module completion	Handoff
	Transfer of IT infrastructure to another party	
Maintenance	Fixing new encountered error	Delay
	Recurring defect	

D. Transformation Plan for Lean IT in ISD Organization

The lean IT approach is correlated with the organizational IS success factors. The IS developer team must understand what comprises a successful IS criteria set, besides practicing lean IT principles based on the HOT-fit framework and applying lean tools to manage ISD optimally. Fig. 3 illustrates the transformation plan based on the findings derived from this study. An ISD can be improved using various lean tools, such as the fishbone diagram, to understand a problem and its root cause and subsequently identify alternatives to the best

solution. Scrum meeting sessions can be held at the early or late part of the week among IS developers to expedite tasks. In the elicitation phase, the identified business process can be optimized through automation to significantly eliminate waste and increase operational efficiency. Smooth value flow can be ensured by properly planning the task assignments, thus avoiding unevenness and overburden.

In expediting deliverables, IS developers can easily prioritize certain tasks by using a *Kanban* board. This lean tool is extremely useful in visualizing project progress. Tasks are usually drawn on a white board or depicted using online application tools, such as *Asana* or *Jira*, in which case the *To-Do*, *In Progress*, and *Done* tasks can be easily viewed. In realizing the respect-for-others principle, teamwork spirit, instead of individualistic attitude, must be cultivated because IS deliverables require different types of expertise from the team. As for overall optimization, the IS developers can apply value stream mapping (VSM) to understand the value flow, allowing them to easily identify and eliminate waste. VSM can also improve ISD management and reduce unevenness and over-burden of IS developers. Table VIII summarizes our recommendations for ISD improvement at IRU.

The framework discussed above ensures system quality with high user fulfillment, good interface design, and positive feedback from the third party. Information quality depends on the data storage method, the accuracy and efficiency of information processing, and the data security level. Service quality needs to be offered to the users in the form of change request or error mitigation. IS developers must ensure that the requirements are clear, consistent, and realistic to subsequently warrant smooth system development. Business processes must also be finalized to avoid uncontrollable changes. The accurate selection of hardware specification is also crucial in optimizing the IS developer specialization.

Moreover, IS developers need to learn the level of IS use to identify its value and contribution towards the ISD. User intention and IS purpose should be identified to ensure that they will be revisited upon project completion. IS developers need to focus on pleasant system use and ease of learning to achieve user satisfaction. The IS will likely be successful if it can gain the support of the top management. Good rapport among team members is important in cultivating the respect-for-others attitude. IS developers should also be given access to sufficient information to ensure the fulfillment of user requirements. Cooperation and involvement among IS developers and stakeholders are both critical in obtaining feedback and ideas. The project manager must play his or her role in ensuring the best project financial state. IS developers must contribute towards supporting both the organizational vision and mission and the operational planning.
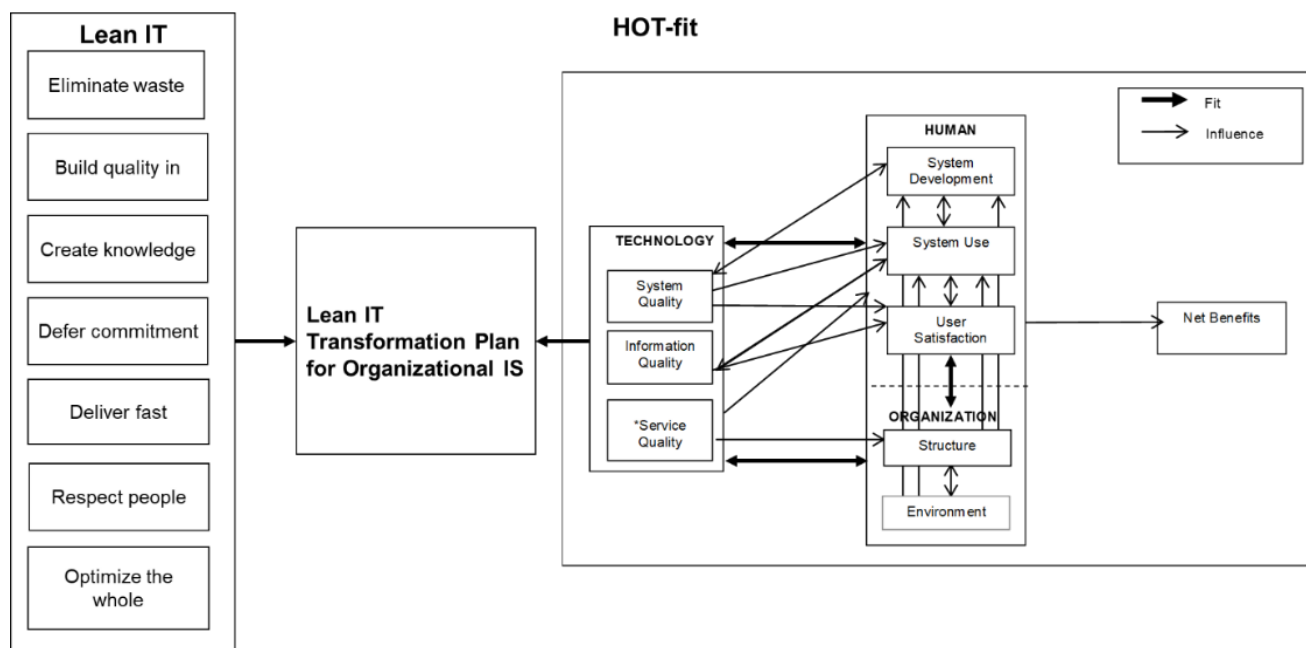
Fig 3. Lean IT Transformation Plan in IS Organization.

TABLE VIII. RECOMMENDATIONS FOR ISD IMPROVEMENT AT IRU

Principle	Recommendations for ISD improvement
Waste elimination	Use root cause analysis to understand problems Organize weekly scrum meeting Optimize business process through automation (*Jidoka*) Avoid uneven workload assignment
Build quality in	Use the PDCA methodology to improve IS Use *Poka-Yoke* in all IS functions
Create knowledge	Establish portal for online knowledge Held specific training and tutorial sessions
Defer commitment	Postpone decision until solution is discovered
Deliver fast	Sort and determine priority using Kanban board
Respect for others	Enhanced teamwork spirit among IS developers
Optimize the whole	Understand the value flow by using VSM

VI. CONCLUSION

The lack of understanding in developing a successful IS and avoiding waste in the ISD context triggers a chain reaction that hinders the achievement of a lean and efficient ISD. Therefore, IS developers need to understand the IS success factors and practice lean IT principles. This study contributes in the identification of the ISD success criteria, waste categories in ISD, and levels of organizational compliance to lean IT principles. More importantly, this study proposes a lean IT transformation plan to evaluate ISD success and waste in the ISD context based on the HOT-fit framework and a lean IT approach to guide IS developers in producing successful systems. The plan features a holistic, systematic, and structured approach in ISD and supports IT and business alignment in the organization.

The study scope is limited to the IS developer perspective only, and access to document analysis has been constrained by data confidentiality. However, the study is applicable to other ISD endeavors with the same development environment and project scale. Future work can be carried out to refine and test the proposed ISD transformation plan in terms of measures, implementation method, breadth of scope, and inclusion of other stakeholders' perspective, including end-users or the management. The transformation plan may guide IS developers in delivering high-quality IS to support the organizational vision and mission and in progressing towards the fourth industrial revolution.

REFERENCES

[1] R. Marques, G. Costa, M. Mira da Silva, and P. Gonçalves, "A survey of failures in the software development process. 25th Eur Conf Inform Syst", pp. 2445-2459, 2017.

[2] V. Rastogi, "Software development life cycle models- comparison, consequences" Int J Comp Sc IT vol. 6 (1), pp.168-172, 2015.

[3] Y. Lindsjørn, D.I.K. Sjøberg, T. Dingsøyr, G.R. Bergersen, and T. Dybå, "Teamwork quality and project success in software development: A survey of agile development teams", J Syst Softw vol. 122, pp. 274-286, 2016.

[4] T. Sedano, P. Ralph, and C. Peraire, "Software development waste". IEEE/ACM 39th Int Conf Softw Eng (ICSE), pp.130-140, 2017.

[5] H. Alahyari, T. Gorschek, and S.R. Berntsson, "An exploratory study of waste in software development organizations using agile or lean approaches: A multiple case study at 14 organizations", Inform Softw Tech vol. 105, pp. 78-94, August 2018.

[6] M.M. Yusof, J. Kuljis, A. Papazafeiropoulou, and L.K. Stergioulas, "An evaluation framework for Health Information Systems: human,

organization and technology-fit factors (HOT-fit)", Int J Med Inf vol. 77 (6), pp. 386-398, 2008.

[7] S.A. Ebad, "Influencing Factors for IT software project failures in developing countries — a critical literature survey", J Soft vol. 11 (11), pp. 1145-1153, 2016.

[8] M.S. Hossain, "Challenges of software quality assurance and testing", Int J Softw Eng Com Syst vol. 4 (1), 133-144, 2018.

[9] K.C. Laudon and J.P. Laudon, Management Information Systems: Managing the Digital Firm. New York:Pearson, 2019.

[10] A. Alami, "Why do information technology projects fail?", Procedia Com Sc vol. 100, pp. 62-71, 2016.

[11] M.Z. Yusoff, M. Mahmuddin, and M. Ahmad, "A knowledge work productivity conceptual model for software development process in SME", ARPN J Eng App Sc vol. 10 (3), pp. 1123-1130, 2015.

[12] M.K. Sharma, "A study of SDLC to develop well engineered software", Int J Adv Res Comp Sc vol. 8 (3), pp. 520-523, 2017.

[13] M. Shameem, B. Chandra, C. Kumar, and A.A. Khan, "Understanding the relationships between requirements uncertainty and nature of conflicts: a study of software development team effectiveness", Arab J Sc Eng vol. 43 (12), pp. 8223-8238, 2018.

[14] S.Z. Salam, and M.M. Yusof, "Knowledge integration in determining user requirements" The 9th Malaysian Softw Eng Conf, Seri Kembangan, Malaysia, pp. 112-116, 2015.

[15] S. Aitzaz, G. Samdani, M. Ali, and M. Kamran, "A Comparative Analysis of In-house and Outsourced Development in Software Industry", Int J Com App vol. 141 (3), pp.18-22. 2016.

[16] L. Mieritz, Gartner survey shows why projects fail. This is what good look like. 2013.

[17] B.S. Masood, "Usability evaluation method for agile software development", Int J Com Sys Soft Eng vol. 1 (1), pp. 29-40, 2015.

[18] V. Blijleven, K. Koelemeijer, and M. Jaspers, "Identifying and eliminating inefficiencies in information system usage: A lean perspective", Int J Med Inform 107:40-47, 2017.

[19] J.K. Liker, The Toyota way: 14 management principles from the world's greatest manufacturer, New York: Mc Graw Hill. 2004.

[20] T. Taipalus, V. Seppänen, and M. Pirhonen, "Coping with uncertainty in an agile system development course", J Inform Syst Edu vol. 29 (2),

pp.117-126, 2018.

[21] A. Janes, "A guide to lean software development in action", IEEE 8th Int Conf Softw Testing, Verification and Validation Workshops, ICSTW pp.1-2, 2015.

[22] S. Bell and M. Orzen. Lean IT: enabling and sustaining your lean transformation, Boca Raton: CRC Press, 2011.

[23] M.M. Yusof, "A case study evaluation of a Critical Care Information System adoption using the socio-technical and fit approach", Int J Med Inf vol. 77 (6), pp. 377-385, 2015.

[24] J.P.Womack and D.T. Jones, "Lean thinking−banish waste and create wealth in your corporation", J Oper Res Soc vol. 48 (11), pp. 1148-1148, 1997.

[25] J. Kobus and M. Westner, "Lean management of IT organizations: a perspective of IT slack theory", Int Conf on Inform Syst, pp 1-12, 2016.

[26] J. Kobus, M. Westner, S. Strahringer, and D. Strode, "Enabling digitization by implementing Lean IT: lessons learned", TQM J vol. 30 (6), pp. 764-778, 2018.

[27] A.N.A. Wahab, M. Mukhtar, and R. Sulaiman, "A conceptual model of lean manufacturing dimensions", Procedia Technology vol. 11, pp. 1292-1298, 2013.

[28] N.A. Kalong and M.M. Yusof, "Waste in health information systems: a systematic review", Int J Health Care Q Assur vol. 30(4), pp. 341-357, 2017.

[29] J. Wanitwattanakosol and S. Noamna S, "Action research framework in Lean Information Technology", Int Conf Syst Sci Eng (ICSSE), pp. 1-3, 2018.

[30] R.K. Yadav, M.L. Mittal, and R. Jain, R, "Adoption of lean principles in software development projects", Int J Lean Six Sigma, vol. 11(2), pp. 285-308, 2020.

[31] M. Poppendieck and M.A. Cusumano, "Lean software development: A tutorial", IEEE Softw vol. 29 (5), pp. 26-32, 2012.

[32] M. Poppendieck and T. Poppendieck, Implementing Lean software development: from concept to cash. Boston: Addison-Wesley, 2006.

[33] M. Poppendieck and T. Poppendieck T, Lean software development: an agile toolkit, Upper Sadle River: Pearson, 2003.

[34] D.R. Thomas, "A general inductive approach for analyzing qualitative evaluation data. Am J Eval vol. 27 (2), pp. 237-246, 2006.

Automatic Hate Speech Detection using Machine Learning: A Comparative Study

Sindhu Abro[1], Sarang Shaikh[2], Zafar Ali[4]
Sajid Khan[5], Ghulam Mujtaba[6]
Center for Excellence for Robotics, Artificial Intelligence
and Blockchain, Department of Computer Science
Sukkur IBA University, Sukkur, Pakistan

Zahid Hussain Khand[3]
Department of Computer Science
Sukkur IBA University
Sukkur, Pakistan

Abstract—The increasing use of social media and information sharing has given major benefits to humanity. However, this has also given rise to a variety of challenges including the spreading and sharing of hate speech messages. Thus, to solve this emerging issue in social media sites, recent studies employed a variety of feature engineering techniques and machine learning algorithms to automatically detect the hate speech messages on different datasets. However, to the best of our knowledge, there is no study to compare the variety of feature engineering techniques and machine learning algorithms to evaluate which feature engineering technique and machine learning algorithm outperform on a standard publicly available dataset. Hence, the aim of this paper is to compare the performance of three feature engineering techniques and eight machine learning algorithms to evaluate their performance on a publicly available dataset having three distinct classes. The experimental results showed that the bigram features when used with the support vector machine algorithm best performed with 79% off overall accuracy. Our study holds practical implication and can be used as a baseline study in the area of detecting automatic hate speech messages. Moreover, the output of different comparisons will be used as state-of-art techniques to compare future researches for existing automated text classification techniques.

Keywords—Hate speech; online social networks; natural language processing; text classification; machine learning

I. INTRODUCTION

In recent years, hate speech has been increasing in-person and online communication. The social media as well as other online platforms are playing an extensive role in the breeding and spread of hateful content – eventually which leads to hate crime. For example, according to recent surveys, the rise in online hate speech content has resulted in hate crimes including Trump's election in the US [2], the Manchester and London attacks in the UK [3], and terror attacks in New Zealand [4]. To tackle these harmful consequences of hate speech, different steps including legislation have been taken by the European Union Commission. Recently, the European Union Commission also enforced social media networks to sign an EU hate speech code to remove hate speech content

within 24 hours [1]. However, the manual process to identify and remove hate speech content is labor-intensive and time-consuming. Due to these concerns and widespread hate speech content on the internet, there is a strong motivation for automatic hate speech detection.

The automatic detection of hate speech is a challenging task due to disagreements on different hate speech definitions. Therefore, some content might be hateful to some individuals and not to others, based on their concerned definitions. According to [5], hate speech is:

"the content that promotes violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, veteran status, and sexual orientation/gender identity".

Despite these different definitions, some recent studies claimed favorable results to detect automatic hate speech in the text [21-32]. The proposed solutions employed the different feature engineering techniques and ML algorithms to classify content as hate speech. Regardless of this extensive amount of work, it remains difficult to compare the performance of these approaches to classify hate speech content. To the best of our knowledge, the existing studies lack the comparative analysis of different feature engineering techniques and ML algorithms.

Therefore, this study contributes to solving this problem by comparing three feature engineering and eight ML classifiers on standard hate speech datasets. Table I shows major concepts related to automatic text classification along with their explanations and references. This study holds practical importance and served as a reference for new researchers in the domain of automatic hate speech detection.

This rest of the paper is organized as: Section II highlights the related works. Section III discusses the methodology. Sections IV, V, and VI explain the experimental settings, results, and discussion. Finally, Section VII discusses the limitation, future work, and conclusion as well.

TABLE I.   TEXT CLASSIFICATION (KEY CONCEPTS)

S. No.	Concept	Acronym	Definition	References
1	Feature Extraction	FE	It is mapping from text data to real-valued vectors.	[6]
2	Bigram	-	It's a feature engineering technique which represents two adjacent words in a single numeric feature while creating master feature vectors for words.	[7]
3	Term Frequency - Inverse Document Frequency	TFIDF	It's a feature representation technique that represents "word importance" is to a document in the document set. It works in a combination of the frequency of word appearance in a document with no. of documents containing that word.	[8]
4	Word2vec	-	It is a technique used to learn vector representation of words, which can further be used to train machine learning models.	[9]
5	Doc2vec	-	It is an unsupervised technique to learn document representations in fixed-length vectors. It is the same as word2vec, but the only difference is that it is unique among all documents.	[10]
6	Machine Learning Classifiers	ML Classifiers	These are applied to numeric features vector to build the predictive model which can be used for prediction class labels.	[11]
7	Naïve Bayes	NB	It's a probabilistic based classification algorithm, which uses the "Bayes theorem" to predict the class. It works on conditional independence among features.	[12]
8	Random Forest	RF	It's a type of ensemble classifier consisting of many decision trees. It classifies an instance based on voting decision of each decision trees class predictions.	[13]
9	Support Vector Machines	SVM	It's a supervised classification algorithm which constructs an optimal hyperplane by learning from training data which separates the categories while classifying new data.	[14]
10	K Nearest Neighbor	KNN	It's a simple text classification algorithm, which categorize the new data using some similarity measure by comparing it with all available data.	[15]
11	Decision Tree	DT	It is a supervised algorithm. It generates the classification rules in the tree-shaped form, where each internal node denotes attribute conditions, each branch denotes conditions for outcome and leaf node represents the class label.	[16]
12	Adaptive Boosting	AdaBoost	It is one of the best-boosting algorithms, which strengthens the weak learning algorithms.	[17]
13	Multilayer Perceptron	MLP	It is a feedforward artificial neural network. It produces a set of outputs using a set of inputs	[18]
14	Logistic Regression	LR	It is a predictive analysis. It uses a sigmoid function to explain the relationship between one independent variable and one or more independent variables	[19]

II. RELATED WORKS

These days, hate speech is very common on social media. Therefore, in previous years, some of the researchers have applied a supervised ML-based text classification approach to classify hate speech content. Different researchers have employed different variety of feature representation techniques namely, dictionary-based [21-23], Bag-of-words-based [24-26], N-grams-based [27-29], TFIDF-based [30, 31] and Deep-Learning-based [31].

Peter Burnap et al. [20] employed a dictionary-based approach to identify cyber hate on Twitter. In this research, they employed an N-gram feature engineering technique to generate the numeric vectors from the predefined dictionary of hateful words. The authors fed the generated numeric vector to ML classifier namely, SVM and obtained a maximum of 67% F-score. Stéphan Tulkens et al. [22] also used a dictionary-based approach for the automatic detection of racism in Dutch Social Media. In this study, the authors used the distribution of words over three dictionaries as features. They fed the generated features to the SVM classifier. Their experimental results obtained 0.46 F-Score. Njagi Dennis et al. [21] used

ML-based classifier to classify hate speech in web forums and blogs. The authors employed a dictionary-based approach to generate a master feature vector. The features were based on sentiment expressions using semantic and subjectivity features with an orientation to hate speech. Afterward, the authors fed the masters feature vector to a rule-based classifier. In the experimental settings, the authors evaluated their classifier by using a precision performance metric and obtained 73% precision.

Nonetheless, the combination of dictionary-based and ML approaches showed a good result. However, the major disadvantage of such type of approach is that it requires a dictionary, based on the large corpus to look for domain words. To overcome this drawback, many of the researchers have used a BOW-based approach which is similar to a dictionary-based approach but the word features are obtained from training data and not from the predefined dictionaries.

Edel Greevy et al. [23] used the supervised ML approach to classify the racist text. To convert the raw text into numeric vectors, the authors employed a bigram feature extraction technique. The authors used bigram features, with the BOW feature representation technique. They used the SVM

classifier to perform experimental results. In their results, they achieved 87% accuracy. Irene Kwok et al. [24] employed an ML-based approach to the automatic detection of racism against black in the twitter community. In their research, they employed unigram with the BOW-based technique to generate the numeric vectors. The authors fed the generated numeric vector to the Naïve Bayes classifier. Their experimental results obtained a maximum of 76% accuracy. Sanjana Sharma et al. [25] classified hate speech on twitter. In their research, they employed BOW features. The authors fed the generated numeric vector to the Naïve Bayes classifier. Their experimental results showed a maximum of 73% accuracy.

Nevertheless, BOW showed better accuracy in social network text classification. However, the major disadvantage of this technique is, the word-order is ignored and causes misclassification as different words are used in different contexts. To overcome this limitation, researchers have proposed an N-grams-based approach [7].

Zeerak Waseem et al. [28] classify the hate speech on twitter. In their research, they employed character Ngrams feature engineering techniques to generate the numeric vectors. The authors fed the generated numeric vector to the LR classifier and obtained overall 73% F-score. Chikashi Nobata et al. [27] used the ML-based approach to detect the abusive language in online user content. In their research authors employed character Ngrams feature representation technique to represent the features. The authors fed the features to the SVM classifier. The results showed that the classifier obtained overall 77% F-score. Shervin Malmasi et al [26] used an ML-based approach to classify hate speech in social media. In their research, the authors employed 4grams with character grams feature engineering techniques to generate numeric features. The authors fed the generated numeric features to the SVM classifier. The authors reported maximum of 78% accuracy.

In recent years, few researchers employed ML approaches to detect automatic hate speech. For example, Karthik Dinakar et al. [29] classified sensitive topics from social media comments or posts. In their research, they employed unigram with the TFIDF feature representation technique to generate the numeric feature vectors. The authors fed the generated features to four ML classifiers namely Naïve Bayes, rule-based, J48, and SVM. Their experimental results showed that the rule-based classifier outperformed NB, J48 and SVM classifiers by obtaining 73% accuracy. Shuhua Liu et al. [30] performed classification on web content pages into hatred or violence categories. In their study, they used trigram features, represented using TFIDF. The authors used the Naïve Bayes classifier. In their experimental settings, the Naïve Bayes classifier obtained highest accuracy of 68%.

The N-gram-based approach gives better results than the BOW-based approach but it has two major limitations. First, the related words may be at a high distance in sentence and finally increasing the N value, results in slow processing speed [32].

In recent years, authors employed deep learning-based NLP techniques to classify hate speech messages. Sebastian Köffer et al. [31] employed word2vec features and SVM

classifiers to classify German texts hate speech messages and obtained 67% F-score. The word2Vec showed the lowest results because such approaches need enormous data to learn complex word semantics.

Recently, there has been a good attempt to construction and detection of hate speech as well as offensive language in other languages (i-e: Danish). An important research study [45] in 2019 worked on the construction of Danish dataset for hate speech and offensive language detection. The dataset contained comments from Reddit and Facebook. It also contained the various types and targets of the offensive language. The authors achieved the highest F1 score of 0.74 by using deep learning models with different features sets.

Schmidt et al. [46] conducted a survey on hate speech detection using natural language processing in 2017. The authors discussed in detail studies regarding various feature engineering techniques to be used for supervised classification of hate speech messages. The major drawback of this survey is that there were no experimental results for those mentioned techniques.

Previous studies showed that a variety of researchers from across the globe are working on hate speech recognition written in different languages such as German, Dutch and English. However, according to our information, no study provides a comparative study of various features and ML algorithms on the standard dataset that can serve as a baseline study for future researchers in the field of hate speech recognition. Hence, in this study, we compared three feature engineering and eight ML classifiers to evaluate which one best works on hate speech datasets (discussed in Section III).

III. METHODOLOGY

This section explains the proposed system which we have employed to classify tweets into three different classes namely, "hate speech, offensive but not hate speech, and neither hate speech nor offensive speech". Fig. 1 shows the complete research methodology. As shown in this figure, the research methodology is contained of six key steps namely, data collection, data preprocessing, feature engineering, data splitting, classification model construction, and classification model evaluation. Each of the step is discussed in detail in the subsequent sections.
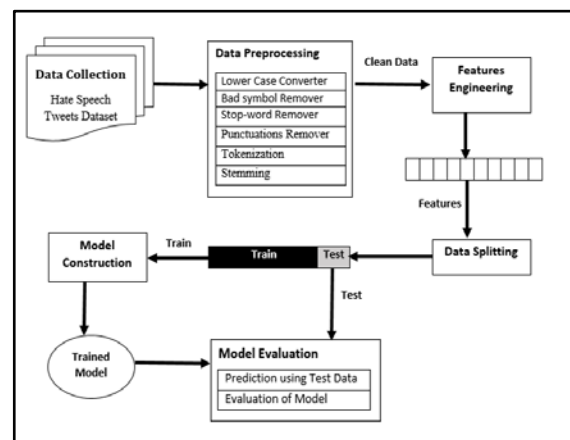

Fig. 1.   System Overview.

A. Data Collection

In this research study, we collected publicly available hate speech tweets dataset. This dataset is compiled and labeled by CrowdFlower. In this dataset, the tweets are labeled into three distinct classes, namely, hate speech, not offensive, and offensive but not hate speech. This dataset has 14509 number of tweets. Of these, 16% of tweets belong to class hate speech. In addition, 50% of tweets belong to not offensive class and the remaining 33% tweets are offensive but not hate speech class. The details of this distribution are also shown in Fig. 2.

B. Text Preprocessing

Several research studies have explained that using text preprocessing makes better classification results [33]. So, in our dataset, we applied different preprocessing-techniques to filter noisy and non-informative features from the tweets. In preprocessing, we changed the tweets into lower case. Also, we removed all the URLs, usernames, white spaces, hashtags, punctuations and stop-words using pattern matching techniques from the collected tweets. Besides this, we have also performed tokenization and stemming from preprocessed tweets. The tokenization, converts each single tweet into tokens or words, then the porter stemmer converts words to their root forms, such as offended to offend using porter stemmer.

C. Feature Engineering

The ML algorithms cannot understand the classification rules from the raw text. These algorithms need numerical features to understand classification rules. Hence, in text-classification one of the key steps is feature engineering. This step is used for extracting the key features from raw text and representing the extracted features in numerical form. In this study, we have performed three different features engineering techniques, namely, *n*-gram with TFIDF [8], Word2vec [9] and Doc2vec [10].

D. Data Splitting

Table II shows the class-wise distribution of the overall dataset as well as data set after splitting (i.e. Training set and Test set). We have used the 80-20 ratio to split the preprocessed data (i.e. 80% for Training Data and 20% for Test Data). The training data is used to train the classification model to learn classification rules. Moreover, the test data is further used to evaluate the classification model.
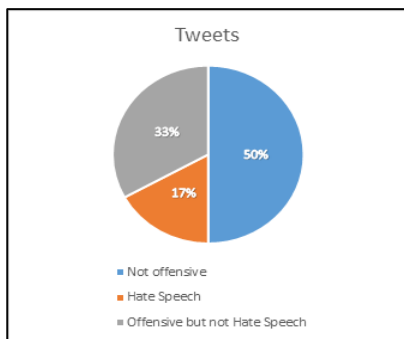
TABLE II.   DETAILS OF DATA SPLIT

	Class	Total Instances	Training instances	Testing instances
0	Hate Speech	2399	1909	490
1	Not offensive	7274	5815	1459
2	Offensive but not Hate Speech	4836	3883	953
	Total	**14509**	**1607**	**2902**

E. Machine Learning Models

According to "no free lunch theorem" [34], there is no any single classifier which best performs on all kinds of datasets. Therefore, it is recommended to apply several different classifiers on a master feature vector to observe which one reaches to the better results. Hence, we selected *eight* different classifiers NB [12], SVM [14], KNN [15], DT [16], RF [13], AdaBoost [17], MLP [18] and LR [19].

F. Classifier Evaluation

In this step, the constructed classifier predicts the class of unlabeled text (i.e. "hate speech, offensive but not hate speech, neither hate speech nor offensive speech") using test set. The classifier performance is evaluated by calculating true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP). These four numbers constitute a confusion matrix as in Fig. 3. Different performance metrics are used to assess the performance of the constructed classifier. Some common performance measures in text categorization are discussed briefly below. The more details of performance metrics can be found in [35].

1) Precision: Precision is also known as the positive predicted value. It is the proportion of predictive positives which are actually positive. Refer to "(1)".

$$Precision = \frac{TP}{(TP+FP)} \tag{1}$$

2) Recall: It is the proportion of actual positives which are predicted positive. Refer to "(2)".

$$Recall = \frac{TP}{(TP+FN)} \tag{2}$$

3) F-Measure: It is the harmonic mean of precision and recall (as shown in Equation 3). The standard F-measure (F1) gives equal importance to precision and recall. Refer to "(3)".

$$F - measure = \frac{2 \times (precision \times recall)}{(precision + recall)} \tag{3}$$

4) Accuracy: It is the number of correctly classified instances (true positives and true negatives). Refer to "(4)".

$$Accuracy = \frac{(TP+TN)}{TP+FP+TN+FN} \tag{4}$$

Fig. 2.   Class wise Data Distribution.

	Predicted No	Predicted Yes
Actual No	TN	FP
Actual Yes	FN	TP

Fig. 3.   Confusion Matrix.

IV. Experimental Settings

As mentioned in section C, we used three types of features namely *n*-gram (bigram) with TFIDF, Word2vec and Doc2vec. Hence, we have a total of three different master feature representations. In addition, eight different ML algorithms were applied to the created three master feature vectors. Hence, overall 24 analyses (3 master feature vectors x 8 ML algorithms) were evaluated to check the effectiveness of classification models.

V. Results

This section explains the overall results of 24 analyses. Tables III to Table VI shows the precision, recall, F-measure and accuracy of all 24 analyses, respectively. The bold values represented are the maximum and minimum result values. All the tables are showing performance for different features representation and classification techniques applied in experimental settings. In all 24 analyses, the lowest precision (0.58), recall (0.57), accuracy (57%) and F-measure (0.47) found in MLP and KNN classifier using TFIDF features representation with bigram features. Moreover, the highest recall (0.79), precision (0.77), accuracy (79%) and F-measure (0.77) were obtained by SVM using TFIDF features representation with bigram features. In feature representation, bigram features with TFIDF obtained the best performance as compared to Word2vec and Doc2vec. However, there was a fringe difference between the result observed in bigram, and Doc2vec. In text-classification models, the SVM classifier best performed among all the eight classifiers. However, the AdaBoost and RF classifiers results were lesser than SVM results and were better than LR, DT, NB, KNN, and MLP results.

Furthermore, Fig. 4 and Fig. 5 show the confusion matrix of best-performing analyses. Fig. 4 shows the SVM classifiers' confusion matrix using bigram with TFIDF features. As shown here, out of 490 tweets belonging to hate speech class, only 155 were correctly classified. However, the 335 instances were incorrectly classified. Of these 335

instances, 54 were falsely classified as not offensive and 281 were falsely classified as Offensive but not Hate Speech. The 1459 instances belong to the second class, the 1427 tweets were correctly classified as not offensive speech. The remaining 32 instances were misclassified, 5 were incorrectly classified as hate speech and 27 were falsely classified as an offensive language but not hate speech. The remaining 953 instances out of 2902 test set belonging to offensive language but not hate speech class. Here, the SVM classifier correctly classified the 698 tweets as an offensive language but not hate speech. The 122 and 133 instances were misclassified into hate speech and not offensive speech, respectively.

However, Fig. 5 shows the confusion matrix of the Adaboost classifier using bigram with TFIDF features. As shown here, the overall performance of the Adaboost classifier is lower than the SVM classifier while using bigram with TFIDF features. The Adaboost only performed well in offensive language but not hate speech class.
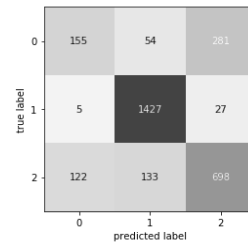
Fig. 4. Confusion Matrix (Features: Bigram (TFIDF), Classifier: SVM).
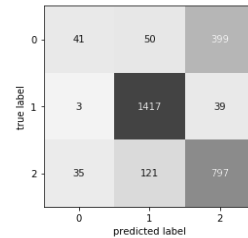
Fig. 5. Confusion Matrix (Features: Bigram (TFIDF), Classifier: ADABOOST).

TABLE III. Precision of all 24 Analysis

Features	LR	NB	RF	SVM	KNN	DT	AdaBoost	MLP
Bigram	0.72	0.71	0.73	**0.77**	0.61	0.71	**0.75**	**0.58**
Word2vec	0.69	0.66	0.66	0.70	0.64	**0.62**	0.65	0.69
Doc2vec	0.70	0.65	0.65	0.70	0.69	**0.61**	0.66	0.71

The bold marked values represented are the higher and lower result values.

TABLE IV. Recall of all 24 Analysis

Features	LR	NB	RF	SVM	KNN	DT	AdaBoost	MLP
Bigram	0.75	0.73	0.75	**0.79**	**0.57**	0.73	**0.78**	0.70
Word2vec	0.72	0.67	0.68	0.73	**0.61**	0.63	0.68	0.71
Doc2vec	0.72	**0.62**	0.67	0.72	0.65	0.63	0.67	0.71

The bold marked values represented are the higher and lower result values.

TABLE V.   F-MEASURE OF ALL 24 ANALYSIS

Features	LR	NB	RF	SVM	KNN	DT	AdaBoost	MLP
Bigram	0.72	0.68	0.74	**0.77**	**0.47**	0.71	0.73	0.63
Word2vec	0.69	0.66	0.66	0.70	0.61	**0.60**	0.65	0.65
Doc2vec	0.70	0.63	0.66	0.72	0.65	**0.61**	0.66	0.66

The bold marked values represented are the higher and lower result values.

TABLE VI.   ACCURACY OF ALL 24 ANALYSIS

Features	LR	NB	RF	SVM	KNN	DT	AdaBoost	MLP
Bigram	0.75	0.73	0.75	**0.79**	**0.57**	0.73	0.78	0.70
Word2vec	0.72	0.67	0.68	0.73	**0.61**	0.63	0.68	0.71
Doc2vec	0.72	**0.62**	0.67	0.72	0.65	0.63	0.67	0.71

The bold marked values represented are the higher and lower result values.

VI. DISCUSSION

In the experimental work, we have evaluated eight classifiers over three different feature engineering techniques, giving 24 different analyses over hate speech dataset containing three classes. Our experimental results showed that the SVM algorithm with the combination of bigram with TFIDF FE techniques showed the best results. The theoretical analysis is discussed in subsequent sections.

A. Feature Engineering

The selection of feature engineering is important in text classification. In this study, we compared three distinct feature extraction techniques namely, Bigram with TFIDF, word2vec and doc2vec. The experimental results exhibited that from these three techniques, bigram with TFIDF outperformed. Conversely, the Word2vec and Doc2vec showed lower results. The possible reason for the outperformance of bigram and TFIDF is that bigram maintains the sequence of words compared to word2Vec and doc2vec [36]. Moreover, several studies showed that the TFIDF representation technique is better than the binary and term frequency representation [6].

The possible reason for the lower performance of Word2vec is because it is unable to handle OOV (out-of-vocabulary) words specially in the domain of Twitter data. Moreover, Word2Vec requires a huge amount of training set to learn the complex relationship between the words [37]. However, as shown in Table I (data collection table), our dataset has approximately 15000 tweets, which might be not enough to train effectively to word2vec for eliciting the complex word relationship.

In our experimental results, Doc2Vec also showed lower performance. This might be because it performs low in case of very short length documents [38] and the tweets which we used in our dataset often having 280 character length.

B. Machine Learning Classifier

Several studies proved that no single ML algorithm performed better on all kinds of data. Therefore, the comparison of various ML algorithms is required to discover which one is best performing on the given dataset. Hence, on our dataset, we used eight different ML algorithms as discussed in Section 3.E i.e. ML Models.

The experimental results proved that SVM and AdaBoost classifiers achieved the best performance possibly because SVM uses threshold functions to separate the data, not the number of features based on margin. This shows that SVM is independent upon the presence of the number of features in the data [7, 15]. In addition, SVM has the capability to best perform on non-linear data apart from the linear data because of its kernel functions. The possible reasons behind the outperformance of AdaBoost are that it uses adaptive algorithms to learn the classification rules iteratively [39] and it focuses on the reduction of the training error. The results obtained with RF and LR classifiers are a little lower than SVM and AdaBoost results but are somewhat higher than the results of NB, DT, KNN, and MLP. The low performance of RF might be due to the unavailability of informative features which leads to incorrect predictions [40]. It is possible that the performance of LR might be lower because its decision surface is linear in nature and cannot handle nonlinear data adequately [41].

The lowest performance was obtained amongst the NB, DT, MLP and KNN classifiers. The NB classifier works on conditional independence among features. Thus, the performance of the NB classifier is negatively affected as the conditional dependence becomes more complicated due to the increase in the number of features [12]. The DT showed lower performance in predicting hate speech because the features inside the master features vector are represented as continuous data points that make it difficult to find the ideal threshold values that are required to build a decision tree [42]. The reason behind the poor performance of the MLP classifier is due to not having enough training data that's why it is considered as complex "black box" [43]. The KNN had the worst performance due to laziness of the learning algorithm and it does not work adequately for noisy data [44]. Hence the KNN is not suitable for detecting hate speech tweets.

C. Classwise Performance

As discussed in Section 3.A we have three classes name "hate speech", "offensive but not hate speech" and "neither hate speech nor offensive speech". The results show that all

features and classifiers performed well for two classes (i.e. offensive but not hate speech, and neither hate speech nor offensive speech). Our experimental results showed that the 24 combinations performed lowest for class hate speech. According to Table I, the class "Hate Speech" has the lowest training instances as compared to other classes, but the major reason for misclassification of class "Hate Speech" (as shown in Fig. 3 and Fig. 4) might be overlapping of different bigram words with higher frequency in other classes than hate speech class. For example, bigrams like "lame nigga, white trash, bitch made" are more frequently appearing in class "Offensive but not Hate Speech" as compared to class "Hate Speech". Hence, it might be possible that the classifier learned weak learning rules.

VII. CONCLUSION

This study employed automated text classification techniques to detect hate speech messages. Moreover, this study compared three feature engineering techniques and eight ML algorithms to classify hate speech messages. The experimental results exhibited that the bigram features, when represented through TFIDF, showed better performance as compared to word2Vec and Doc2Vec features engineering techniques. Moreover, SVM and RF algorithms showed better results compared to LR, NB, KNN, DT, AdaBoost, and MLP. The lowest performance was observed in *KNN*. The outcomes from this research study hold practical importance because this will be used as a baseline study to compare upcoming researches within different automatic text classification methods for automatic hate speech detection. Furthermore, this study also holds a scientific value because this study presents experimental results in form of more than one scientific measures used for automatic text classification. Our work has two important limitations. First, the proposed ML model is inefficient in terms of real-time predictions accuracy for the data. Finally, it only classifies the hate speech message in three different classes and is not capable enough to identify the severity of the message. Hence, in the future, the objective is to improve the proposed ML model which can be used to predict the severity of the hate speech message as well. Moreover, to improve the proposed model's classification performance two approaches will be used. First, the lexicon-based techniques will be explored and assessed by comparing with other current state-of-the-art results. Secondly, more data instances will be collected, to be used for learning the classification rules efficiently.

REFERENCES

[1] Hern, A., Facebook, YouTube, Twitter, and Microsoft sign the EU hate speech code. The Guardian, 2016. 31.

[2] Rosa, J., and Y. Bonilla, Deprovincializing Trump, decolonizing diversity, and unsettling anthropology. American Ethnologist, 2017. 44(2): p. 201-208.

[3] Travis, A., Anti-Muslim hate crime surges after Manchester and London Bridge attacks. The Guardian, 2017.

[4] MacAvaney, S., et al., Hate speech detection: Challenges and solutions. PloS one, 2019. 14(8): p. e0221152.

[5] Fortuna, P. and S. Nunes, A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 2018. 51(4): p. 85.

[6] Mujtaba, G., et al., Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. Journal of forensic and legal medicine, 2018. 57: p. 41-50.

[7] Cavnar, W.B. and J.M. Trenkle. N-gram-based text categorization. in Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval. 1994. Citeseer.

[8] Ramos, J. Using tf-idf to determine word relevance in document queries. in Proceedings of the first instructional conference on machine learning. 2003. Piscataway, NJ.

[9] Mikolov, T., et al. Distributed representations of words and phrases and their compositionality. in Advances in neural information processing systems. 2013.

[10] Le, Q. and T. Mikolov. Distributed representations of sentences and documents. in International conference on machine learning. 2014.

[11] Kotsiantis, S.B., I.D. Zaharakis, and P.E. Pintelas, Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 2006. 26(3): p. 159-190.

[12] Lewis, D.D. Naive (Bayes) at forty: The independence assumption in information retrieval. in European conference on machine learning. 1998. Springer.

[13] Xu, B., et al., An Improved Random Forest Classifier for Text Categorization. JCP, 2012. 7(12): p. 2913-2920.

[14] Joachims, T. Text categorization with support vector machines: Learning with many relevant features. in European conference on machine learning. 1998. Springer.

[15] Zhang, M.-L. and Z.-H. Zhou, A k-nearest neighbor based algorithm for multi-label classification. GrC, 2005. 5: p. 718-721.

[16] Abacha, A.B., et al., Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification. Journal of biomedical informatics, 2015. 58: p. 122-132.

[17] Ying, C., et al., Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica, 2013. 39(6): p. 745-758.

[18] Gardner, M.W. and S. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment, 1998. 32(14-15): p. 2627-2636.

[19] Wenando, F.A., T.B. Adji, and I. Ardiyanto, Text classification to detect student level of understanding in prior knowledge activation process. Advanced Science Letters, 2017. 23(3): p. 2285-2287.

[20] Burnap, P. and M.L. Williams, Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Science, 2016. 5(1): p. 11.

[21] Gitari, N.D., et al., A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 2015. 10(4): p. 215-230.

[22] Tulkens, S., et al., A dictionary-based approach to racism detection in dutch social media. arXiv preprint arXiv:1608.08738, 2016.

[23] Greevy, E. and A.F. Smeaton. Classifying racist texts using a support vector machine. in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004. ACM.

[24] Kwok, I. and Y. Wang. Locate the hate: Detecting tweets against blacks. in Twenty-seventh AAAI conference on artificial intelligence. 2013.

[25] Sharma, S., S. Agrawal, and M. Shrivastava, Degree based classification of harmful speech using twitter data. arXiv preprint arXiv:1806.04197, 2018.

[26] Malmasi, S. and M. Zampieri, Detecting hate speech in social media. arXiv preprint arXiv:1712.06427, 2017.

[27] Nobata, C., et al. Abusive language detection in online user content. in Proceedings of the 25th international conference on world wide web. 2016. International World Wide Web Conferences Steering Committee.

[28] Waseem, Z. and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. in Proceedings of the NAACL student research workshop. 2016.

[29] Dinakar, K., R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. in fifth international AAAI conference on weblogs and social media. 2011.

[30] Liu, S. and T. Forss. Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification. in KDIR. 2014.

[31] Köffer, S., et al., Discussing the value of automatic hate speech detection in online debates. Multikonferenz Wirtschaftsinformatik (MKWI 2018): Data Driven X-Turning Data in Value, Leuphana, Germany, 2018.

[32] Chen, Y., Detecting offensive language in social medias for protection of adolescent online safety. 2011.

[33] Shaikh, S. and S.M. Doudpotta, Aspects Based Opinion Mining for Teacher and Course Evaluation. Sukkur IBA Journal of Computing and Mathematical Sciences, 2019. 3(1): p. 34-43.

[34] Ho, Y.-C. and D.L. Pepyne, Simple explanation of the no-free-lunch theorem and its implications. Journal of optimization theory and applications, 2002. 115(3): p. 549-570.

[35] Seliya, N., T.M. Khoshgoftaar, and J. Van Hulse. A study on the relationships of classifier performance metrics. in 2009 21st IEEE international conference on tools with artificial intelligence. 2009. IEEE.

[36] Chaudhari, U.V. and M. Picheny, Matching criteria for vocabulary-independent search. IEEE Transactions on Audio, Speech, and Language Processing, 2012. 20(5): p. 1633-1643.

[37] Li, Y. and T. Yang, Word embedding for understanding natural language: a survey, in Guide to Big Data Applications. 2018, Springer. p. 83-104.

[38] Wang, Y., et al. Comparisons and selections of features and classifiers for short text classification. in IOP Conference Series: Materials Science and Engineering. 2017. IOP Publishing.

[39] Schapire, R.E., The boosting approach to machine learning: An overview, in Nonlinear estimation and classification. 2003, Springer. p. 149-171.

[40] Xu, B., Y. Ye, and L. Nie. An improved random forest classifier for image classification. in 2012 IEEE International Conference on Information and Automation. 2012. IEEE.

[41] Eftekhar, B., et al., Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC medical informatics and decision making, 2005. 5(1): p. 3.

[42] Dreiseitl, S., et al., A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of biomedical informatics, 2001. 34(1): p. 28-36.

[43] Singh, P.K. and M.S. Husain, Methodological study of opinion mining and sentiment analysis techniques. International Journal on Soft Computing, 2014. 5(1): p. 11.

[44] Bhatia, N., Survey of nearest neighbor techniques. arXiv preprint arXiv:1007.0085, 2010.

[45] Sigurbergsson, G. I., & Derczynski, L. (2019). Offensive language and hate speech detection for Danish. arXiv preprint arXiv:1908.04531.

[46] Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International workshop on natural language processing for social media (pp. 1-10).

Study on Dominant Factor for Academic Performance Prediction using Feature Selection Methods

Phauk Sokkhey[1]

Graduate School of Engineering and Science
University of the Ryukyus
Senbaru, Nishihara, Okinawa, 903-0123, Japan
Institute of Technology of Cambodia, Phnom Penh
Cambodia

Takeo Okazaki[2]

Department of Computer Science and Intelligent Systems
University of the Ryukyus
1 Senbaru, Nishihara,
Okinawa, 903-0123,
Japan

Abstract—All educational institutions always try to investigate the learning behaviors of students and give early prediction toward student's outcomes for interventing and improving their learning performance. Educational data mining (EDM) offers various effective prediction models to predict student performance. Simultaneously, feature selection (FS) is a method of EDM that is utilized to determine the dominant factors that are needed and sufficient for the target concept. FS method extracts high-quality data that reduce the complexity of the prediction task that can increase the robustness of decision rule. In this paper, we provide a comparative study of feature selection methods for determining dominant factors that highly affect classification performance and improve the performance of prediction models. A new feature selection CHIMI based on ranked vector score is proposed. Analysis of feature sets of each FS method to get the dominant set is executed. The experimental results show that by using the dominant set of the proposed CHIMI method, the classification performance of the proposed models is significantly improved.

Keywords—Educational data mining; dominant factors; feature selection methods; prediction models; student performance

I. Introduction

The development of developing countries mainly relies on the potential education system that can produce human resources. The success of human resource development depends on long-term investment in education from primary schools, secondary schools, and higher educations. Student performance in high school plays an important role that maximizes or minimizes student success in the secondary school national exam, higher education, and their future careers [1]. Student academic performance can be measured and monitored effectively by using methods of educational data mining (EDM).

Various EDM techniques are effectively used to predict student performance, identify their learning behaviors and progress, and many more [2-5]. The results of these tasks are helpful for students themselves, academic institutions, and related individuals to follow up academic performance, improve student performance, and use as information for planning and scheduling in education systems. In EDM, feature selection (FS) is used in many research work [6-8].

The prediction of student performance highly depends on the choice of selection of most relevant variables. In the educational domain, several factors were concerned to influence academic performance, mainly considers school environment factors, domestics environment factors, demographic background, attitudinal factors, and academic records. Various factors lead to have higher dimensions. Hence, many studies have focused on determining the related factors that affect student performance and predict their academic progress by using FS methods and applying predictive models of EDM [11-16]. In educational research, FS methods aim at determining important factors that are in need and sufficient to report the academic performance; we call it as *dominant factors*.

The dominant set was considered for two main contributtions. Interm of gving intervention to poor-performing students, dominant set is known as the set of important factors that affect student performance. Another most common contribution is to raise the performance of prediction models. FS methods are categorized into 3 classes: filter-based methods, wrapper-based methods, and embedded/hybrid methods [5]. wrapper method and hybrid-based features selection methods are effective, yet computational expesnive to detect the optimal sets in big data content [9]. Filter-based is a simple FS method, yet effective for all types of datasets. In addition, Filter is independence of classifiers and more scalable comparing to other FS methods [10]. The main objective of FS is to select optimal subsets consisting of relevant and informative features.

II. Literature Review

A. Feature Selection Methods and Prediction Models in EDM

This part presents a brief of previous works that have used FS techniques to enhance the performance of the prediction models of EDM.

Estrera et al. [11] gave analysis on high school record of student enrolled for a university. The analysis of the study proposed decision tree (DT), naive Bayes (NB), and k-nearest neighbor (KNN). The decision tree algorithm generates affective results in this classification and prediction problem. Several FS methods were utilized to improved the performance of proposed models and to detect student learning patterns. The proposed FS methods are: Chi-square Statistics (CHS) test, Information Gain (IG) test, and Information Gain Ration (IGR)

test in the study. Experimental results indicated the decision tree produced the most satisfied accuracy.

Ramaswami et al. [12] developed a comparative study of six filter-based feature selection methods for improving academic performance prediction. The used algorithms are correlation-based feature evaluator (CB), Chi-square feature evaluator (CH), gain ration feature evaluator (GR), information gain feature evaluator (IG), Relief (RF), and Symmetric Uncertainty (SU). The results indicated an increase in prediction performance and reduced time consumption.

Febro [13] utilized feature selection methods to improve prediction models and extract important features that affect student retention in higher education. Three filter-based selection methods (Information Gain Ratio (IGR), Correlation-based Feature Selection (CFS), and Chi-square (CHS)) are introduced. The optimal subset of 14 features was extracted from an original set of 29 features. The accuracy result jumps to 92.09%.

Zaffar et al. [14] proposed a study of feature selection techniques to enhance the prediction performance of academic performance. The FS techniques utilized in the study are: correlation based feature evaluator (CFS), Chi-squared test, the fiiltered, gain ration (GR), principal component aanlysis (PC), and Relief method. To comfirm the performance of the proposed FS techniques, the study utilized fifteen prediction models and make comparision of the models on each FS methods. The experiment indicates the improvement of accuracy when applying feature selection.

Alhassan et al. [15] proposed a study of analyzing student learning behaviors and predicting their academic performance in web-based learning management systems (LMS). The study observed the student learning patterns on the online study platform using five machine learning classifiers: J48 of decision tree, random forest (RF), the logistic regression (LR), sequential minimal optimization (SMO), and multilayer perceptron (MLP). Analysis of all feature sets and subsets of features are conducted by using six feature selection methods: Correlation Attribute Evlaition, Information Gain, CFS Subset, Wrapper-J48, Wrapper-NB, Wrapper-IBK. The RF algorithm combined with the feature selection methods outperformed the rest models.

Mythili et al. [16] proposed an analysis of student performance by applying data mining algorithms. The various classification algorithms used in the study are J48, Random Forest (RF), Multilayer Perceptron (MLP), Artificial Neural Network (ANN), IBI of the nearest neighbor classifier, and Decision Table. The experiment was conducted by filtering the important features using Information Gain (IG), and Gain Ratio (GR). With the merit of selecting only high ranking important features, it is discovered that RF performance is the best than that of different algorithms employed in the study.

B. The Current Study

Even if several works of literature have studied using FS algorithms in EDM, however, a lot of attention and consiseration are needed to build academic performance prediction model with the analysis and help of FS methods. The primary purpose of our study is to present an analysis of

feature selection methods to extract the *dominant factors* that are necessary and sufficient to evaluate the success of students' performance. The primary purpose of this study is to introduce a study of analysis on feature selection methods on a set of classifiers and then determine the performance of each algorithm on each classifier. The study proposed a novel FS algorithm to improve the performance of predictive models. We search for an optimal and effective subset that improves the classification performance of classifiers. Consequently, we can obtain the potential prediction model and the dominant factors that maximize and control the evaluation of student performance. The proposed framework of this study is shown in Fig. 1.
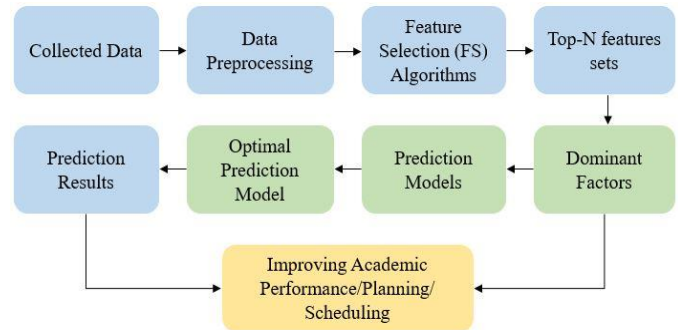
Fig. 1. The Proposed Framework of this Study.

III. METHODOLOGY

A. Participants and Data

The target of the study is to utilized related factors to predict student performace in high school.The proposed data in the study was collected from serval high schools in Cambodia. The questionnaire concerning with any related factors that affect (weak or strong) the student performance and it was designed into five main parts. The first part consist of six questions conecerning with student performanal information. The second part related to domestics or home factors, which has 17 questions. The third part has 15 questions including any information related with student learning behaviors and materials in study. The forth part consists of 14 questions in total concerning with school factors. The last part is the student score for output variables. The describption is illustrated in Table I.

B. Data Preprocessing

Preprocessing task is treated as a necessary step in every modeling. The utilized data mining models usually require data cleansing, data encoding, and data transformation to convert the data into an executable format and enhance model performance. The tool that is used in preprocessing and experiment in this study is R, a powerful tool for machine learning and statistical computation.

C. Model Evaluation Metrics

Evaluating model is a core part of EDM work, two standard model evaluation metrics are utilized in this study. Accuracy and root mean square error are the two commonly used metrics evaluating predction models.

- Accuracy (ACC): ACC is a common model evaluation metric used to evlauce the performance of prediction model by computing the percentage of coreectly prediction [15]. It is calculated as in (1).

$$ACC = \frac{\text{Correctly predicted values}}{\text{Total values}} \tag{1}$$

- Root Mean Square Error (RMSE): RMSE is a standard evaluation metric used to evaluate prediction error by computing the error or difference between actual output and predicted output [15]. It can be calculated using (2).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^a - y_i^p)^2} \tag{2}$$

D. Feature Selection Methods and Dominant Set

Analysis of student's information, their learning behavior, and factors affecting students' academic performance is still a challenging task in educational institutes [13]. Many cognitive and non-cognitive factors affect the academic performance of children and adolescents. Several related domains weakly or highly influence on results and achievements of high school students. Various factors lead to have higher dimensions. Hence, this study focus on determining the related factors that affect student performance and improved the proposed EDM with the merit of data set determining by FS techniques. This feature set is called the dominant set.

TABLE I.   FEATURES AFFECTING STUDENTS PERFORMANCE

Factors	ID	Predictors (number of questions)	Data types
		Student personal information (6)	
Domestic	PEDU	Parents' educational levels (2)	Nominal
	POCC	Parents' occupational status (2)	Nominal
	PSES	Parents' socioeconomic levels (3)	Ordinal
	PI	Parents' involvement (4)	Ordinal
	PS	Parenting styles (4)	Ordinal
	DE	Domestic environment (2)	Ordinal
Student	SELD	Self-regulation on study (5)	Ordinal
	SIM	Interest and motivation (4)	Ordinal
	ANXI	Students' anxiety toward their classes and exams (3)	Ordinal
	POSS	Possession materials for study (3)	Nominal
School	CENV	School and class environment (1)	Ordinal
	CU	Curriculum (2)	Nominal
	TMP	Teaching methods and practices (4)	Ordinal
	TAC	Teachers' attribute & characteristics (4)	Ordinal
	ARES	Academic resource (3)	Nominal
SCORE	PL	Student's performance level based on their mark or score	Ordinal

The dominant set was considered for two main contributions. Interm of gving intervention to poor-performing students, dominant set is known as the set of important factors that affect student performance. Another most common contribution is to improve the performance of prediction mod. From the literature reviews, filter-based feature selection methods is the most popular method in the research of educational domain. Filter-based is a simple FS method, yet effective for all types of datasets. In addition, Filter is independence of classifiers and more scalable comparing to other FS methods [17]. In this study, we propose comparative approach of experiement of a proposed feature selection method to three existing baseline methods.

1) Information Gain (IG): IG is one of the popularly used feature selection methods in data mining. IG utilized the entropy-based method to capture the importance of features [17]. Entropy class C prior to feature F is expressed as:

$$H(C) = -\sum p(c) \times \log_2 p(c), \tag{3}$$

where $p(c)$ is a marginal function of density probability for class C. The conditional entropy of class C for a feature F is denoted as:

$$H(C \mid F) = \sum_{i=1}^{m} \frac{|C_i|}{|C|} H(C_i), \tag{4}$$

where $C = \{C_1, C_2, ..., C_m\}$ is the m partition of class C. The IG of class C as acquired from a feature F is given by the following equation:

$$IG(C, F) = H(C) - H(C \mid F). \tag{5}$$

2) Symmetric Uncertainty (SU): SU is one of the leading feature selection techniques [17]. SU determines the correlation between a feature and target variable using entropy and information gain theory as in equation (6):

$$SU(A, D) = \frac{IG(D, A)}{H(D) + H(A)}, \tag{6}$$

where $H(D)$ and $H(A)$ are entropies of based on probability of class associated with the example set D and attribute A, respectively; $IG(D, A)$ is the information gain as shown in equation (5).

3) Mutual information (MI): MI of two random variables or features is a measure quantifies the dependence measurement between those variables [18]. It is asymmetric measurement such that $I(X, Y) = I(Y, X)$ that can recognize non-linear relationships between variables. MI of two discrete variables X and Y can be described as:

$$I(X, Y) = \sum_{x, y} p_{xy}(x, y) \log \frac{p_{xy}(x, y)}{p_x(x) \times p_y(y)}, \tag{7}$$

where $p_x(x)$ and $p_y(y)$ are marginal probability such that

$$p_x(x) = \sum_y p_{xy}(x, y), \ p_y(y) = \sum_x p_{xy}(x, y).$$

(8)

If the variables X and Y are independent, then the joint probability $p_{xy}(x, y) = p_x(x) \times p_y(y)$ and $I(x, y) = 0$.

4) Chi-square (CHI): CHI is a statistical method that is utilized for measuring the dependency of each input feature to the target class. The technique utilized the feature score from Chi-square test to get the rank list of all input features [9]. The list of an informative feature set can be computed using the equation below:

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(n_{ij} - \varepsilon_{ij})^2}{\varepsilon_{ij}},$$

(9)

where l denotes the number of classes or determined intervals of a particular feature; c represents the number of classes in target variable; n_{ij} is the observed (actual) frequency of sample of ith interval and j^{th} class; ε_{ij} indicate the expected frequency of n_{ij}.

5) The Proposed FS Method (CHIMI): Many studies have confirmed the effectiveness of information gain (IG), symetric uncertainty (SU), and mutual infromation (MI) in many applications. The techniques use the concepts of information theory [18]. These techniques are information-based methods. Chi-square is considered as one of the top methods utilized in many applications and works best with categircal data type [18]. CHI and MI are known as the two outstanding methods [19]; however, many studies have confrimed that working on combined-FS methods is better than trusting on a single method. The CHIMI: CHIMI is a proposed combined-FS method which is the combination of CHI and MI methods by computing the new feature scores.

CHI and MI methods have demonstated its merits of FS methods. Hence, this study come up with the concepts of calucating a new feature scores based on the score of CHI and MI algorithms. Since, it is broadly known that different FS methods compute scores of feature sets differently, therefore, produced different scales. Hence, before computing the new feature scores, the original scores of CHI and MI are first normalized. The normalization of MI and CHI scores can be done as in (10).

$$\overline{CHI} = \frac{CHI_i \ - \ CHI_{min}}{CHI_{max} \ - \ CHI_{min}}, \ \overline{MI} = \frac{MI_i \ - \ MI_{min}}{MI_{max} \ - \ MI_{min}}$$

(10)

Once, the normaliuzation of two score vectors are done, then it is passed to combined a new vector of feature score as formulated in (11).

$$CHIMI = \begin{pmatrix} \overline{CHI} \\ \overline{MI} \end{pmatrix}$$

(11)

The score vector indicated in (11) store the information of feature score of CHI and MI in form of vector. To get the absolute value or magnitude of the combined-FS method, Euclidean norm need to be computed. Hence, the new feature score of the combined-FS method is calculated using (12).

$$|CHIMI_i| = \sqrt{\left(\overline{CHI_i}\right)^2 + \left(\overline{MI_i}\right)^2}$$

(12)

To filter the redundant feature, Correlation Feature Selection (CFS) [13] is then introduced to filter the features of CHIMI. The CFS method evaluates the performance of feature subsets by evaluating the predictive ability of individual feature along with the degree of redundancy between input features.

This implies that the score of a feature in the CHIMI method containing the score vector generated by the CHI and MI algorithms with the different predictive ability of each feature. The new feature rearranges the order of importance of feature, feature with with bigger value of $|CHIMI_i|$ will be ranked higher. Unlike other previous methods of combining scores from different techniques such as AND and OR, our proposed approach gives a true metric on the space for score vector [20]. Some experimental results in earlier works reported a minor improvement or no improvement in classification performance when more than three feature selection methods were combined [21]. This method conducts a mathematical structure for examing the vector space of combined scores.

The normalization of CHI and MI scores is to introduce a new rank of input features based on the computed scores. This method may place the input features within their true rank and improves the higher possibility of certain significant features to being identified for selecting the dominant feature set.

E. Classification Algorithms

Several EDM techniques from many works of literature [11]-[16] were considered. The comparative study of prediction models on predicting student performance was conducted in [23]. The improvement version of the comparative study was conducted in [24]. The experimental results of both works indicated that k-nearest neighbor (KNN), two tree-based models: C5.0 and random forest (RF) are the optimal models. The developed EDM classifiers were proposed in earlier works [25][26][27]. The study of this work utilized the four prediction models as follows:

1) K-nearest neighbor (KNN): KNN KNN is known as an popular non-parametric EDM models utilized in many classicaiton problems. The KNN is confirmed to be a succesful classifier in our classifcation problem as proposed in the previous work [24]. Similarly to other classifiers, the KNN is noise-sensitive classifier. Its performance highly depends on the quality of the training data. The noise of data and mislabeled data, outliers, and overlaps regions between the data of different classes or targets lead to inaccurate classification [22].

2) Hybrid C5.0 and Hybrid RF: Hybrid C5.0 and Hybrid RF are the developed models that were studied in our earlier work [25]. The study gave the development and improvement

version of [23] [24] for prediction academic performance. Four baseline models, naïve Bayes (NB), support vector machine (SVM), C5.0, and random forrest (RF) were utilized. The concept of principal component analysis (PCA) and k-fold cross validation (10-fold CV) were appied to baseline models. The Hybrid C5.0 (C5.0 + PCA + 10-fold CV) and Hybrid RF (RF + PCA + 10-fold CV) are the two standout models.

3) Improved Deep Belief Networks (IDBN): The IDBN is the optimization version of deep belief network (DBN) model. In our previous work, we gave a study of an optimization approach of DBN concerning (i) feature selection method, (ii) optimization of hyper-parameter, and (ii) regularization method [26]. The proposed IDBN successfully achieves the high prediction performance when applying with larger datasets.

IV. Experimental Results of Prediction Models

This section reported the performance evaluation of feature selection methods in selecting the dominant factors for predicting academic performance. We executed the proposed optimal classifiers using subsets that were obtained from each FS method. After applying the FS algorithms to the original datasets, each algorithm captures a subset of top N features. The FS algorithm selects the relevant factors to the target variables, then we rank the feature weight denoting the importance of features from sets selected by each FS algorithm decreasingly. We defined the dominant set as a set of input features containing top-n features that provide the highest prediction performance. The framework of the study is illustrated in Fig. 2.

The experiment was carried out with two phases. The first experiment was executed with the dataset ADS comprised of 1204 samples. The second experiment was with dataset GDS4 comprises 10000 samples. The second experiment was carried out with the context of a larger dataset to confirm the performance of IDBN and other proposed models. The experiment was made a minimal subset of five features to a fully selected set. To evaluate and compare the performance of prediction models, ACC and RMSE are measured. Recall that the higher value of ACC, the better model is. In contrast to ACC, the smaller value of RMSE, the better model is.

A. Experimental Results with ADS Dataset

This section illustrate the experiemenal results of the proposed FS methods with the developed classfiers using ADS dataset. Table II describe the experimental results of the proposed method regarding with orginal dataset. Table III to Table VI illustrate the computational results of each classifier on subsets selected by IG, SU, CHI, MI, and the proposed CHIMI. The experiement aim to detect the dominant of each FS methods.

Table II illustrate the staitistcal results of the two metrics of KNN, Hybrid C5.0, Hybrid RF, and IDBN. The average of ACC and RMSE from serveral iteration run are recorded. The two developed tree-based models, Hybrid C5.0 and Hybrid RF generate the highest ACC and lowest RMSE.

From Fig. 3 and Fig. 4, we can obtain the results of ACC and RMSE of KNN models using a selected set of each FS method, and it is summarized in Table III.

From Fig. 5 and Fig. 6, we can obtain the results of ACC and RMSE of the Hybrid C5.0 model using a selected set of each FS method, and it is summarized in Table IV.

From Fig. 7 and Fig. 8, we can obtain the results of ACC and RMSE of the Hybrid RF model using a selected set of each FS method, and it is summarized in Table V.
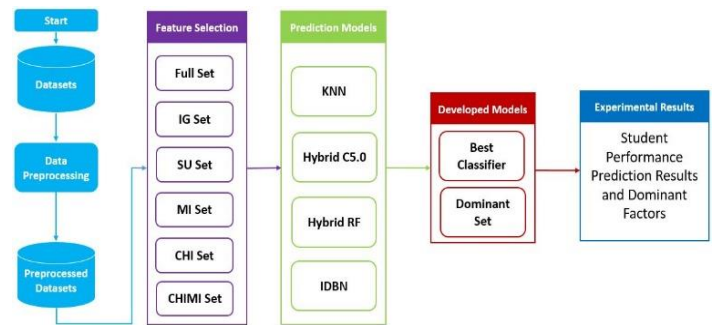
Fig. 2. Flowchart for the Experiment in this Study.

TABLE II. Results of Proposed Models on Original Datasets

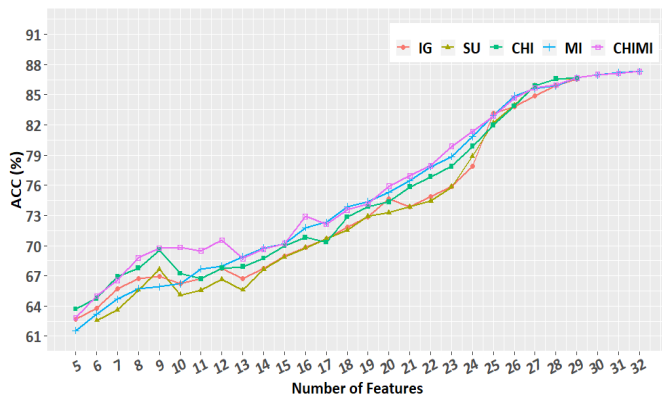| Proposed Models	KNN	Hybrid C5.0	Hybrid RF	IDBN
ACC (%)	94.95	99.25	99.72	83.14
Std. of ACC	0.801	0.601	0.357	0.640
RMSE	0.261	0.073	0.041	0.759
Std. of RMSE	0.041	0.045	0.029	0.031
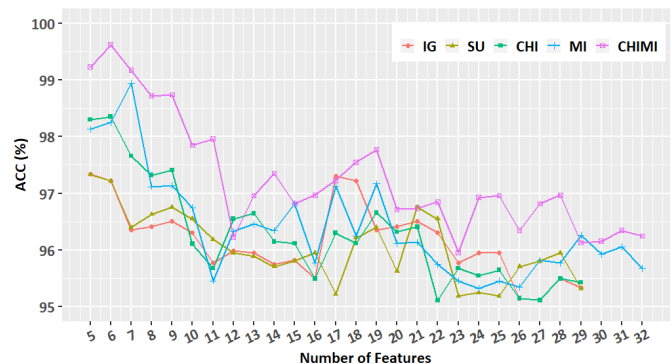
Fig. 3. ACC of KNN using ADS Dataset.
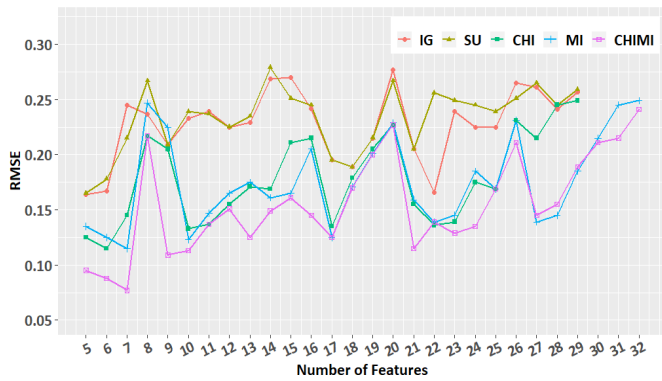
Fig. 4.   RMSE of KNN using ADS Dataset.

TABLE III.   THE PERFORMANCE EVALUATION ON KNN MODEL

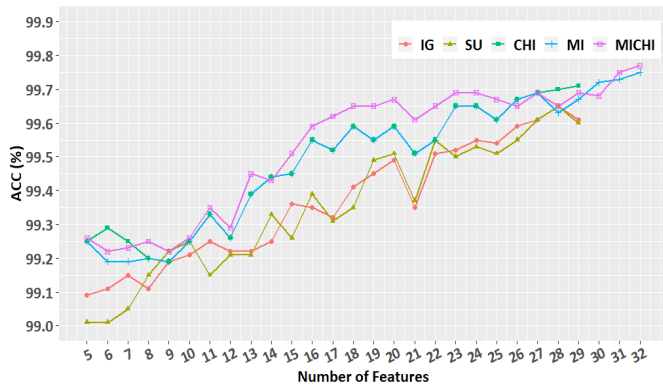| Models	Selected set			Dominant set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	95.35	0.257	5	97.35	0.163
SU	29	95.32	0.259	5	97.33	0.164
CHI	29	95.43	0.249	6	98.35	0.115
MI	32	95.683	0.241	7	98.94	0.077
CHIMI	32	96.25	0.179	6	99.62	0.063
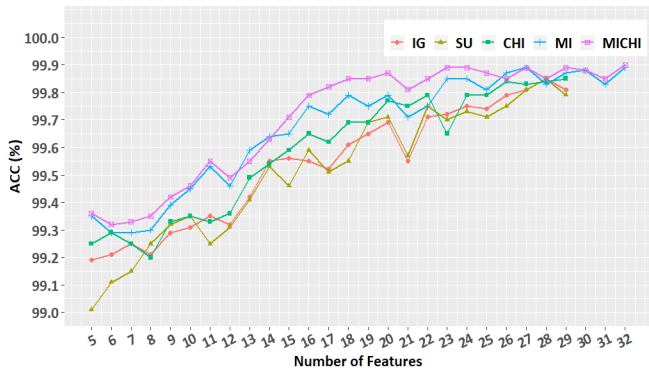
Fig. 5.   ACC of Hybrid C5.0 using ADS Dataset.
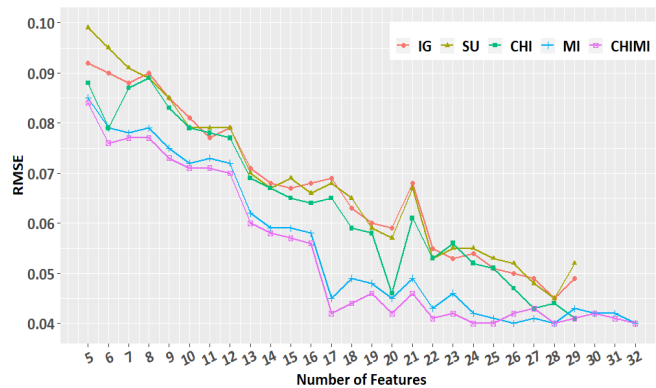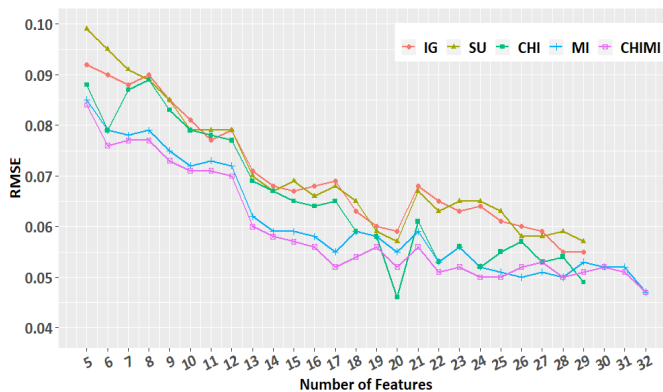

Fig. 6.   RMSE of Hybrid C5.0 using ADS Dataset.

TABLE IV.   THE PERFORMANCE EVALUATION OF HYBRID C5.0 MODEL

Models	Selected set			Dominant set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.81	0.049	28	99.85	0.045
SU	29	99.79	0.051	28	99.85	0.045
CHI	29	99.85	0.041	29	99.85	0.041
MI	32	99.89	0.035	32	99.89	0.035
CHIMI	32	99.90	0.033	32	99.90	0.033
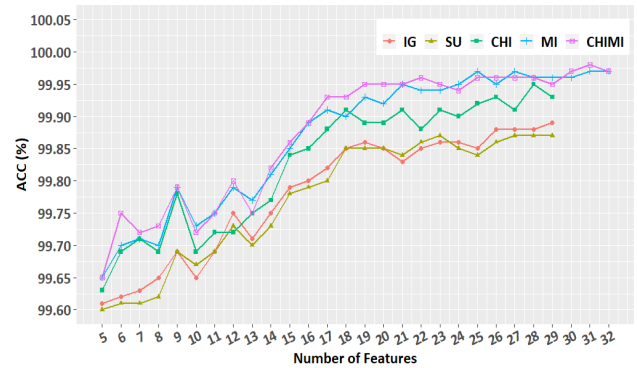
Fig. 7.   ACC of Hybrid RF using ADS Dataset.
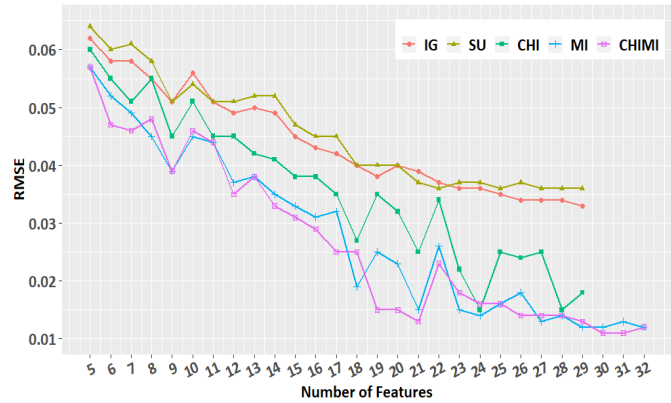
Fig. 8.   RMSE of Hybrid RF using ADS Dataset.

TABLE V.   THE PERFORMANCE EVALUATION OF THE HYBRID RF MODEL

Models	Selected set			Dominant set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.89	0.033	28	99.89	0.033
SU	29	99.87	0.036	28	99.87	0.036
CHI	29	99.95	0.015	29	99.95	0.015
MI	32	99.97	0.012	32	99.97	0.012
CHIMI	32	99.97	0.012	31	99.98	0.011

From Fig. 9 and Fig. 10, we can obtain the results of ACC and RMSE of the IDBN model using a selected set of each FS method, and it is summarized in Table VI.

The results presented in Table III demonstrate the performance of the KNN model on feature sets and dominant sets selected by IG, SU, CHI, MI, and CHIMI methods. The models work best with the low dimension of the most important features. The KNN model with the dominant set of the proposed CHIMI achieves the highest performance. The ACC and RMSE are improved to 99.62 and 0.063, respectively.

From Table IV, the performance of the Hybrid C5.0 is significantly improved when using the dominant set of the CHIMI. The ACC and RMSE of the Hybrid C5.0 are improved to 99.90 and 0.033, respectively.

The results of ACC and RMSE of the Hybrid RF are shown in Table V. The proposed CHIMI method outperforms the rest FS methods in achieving the highest ACC and lowest RMSE. The ACC and RMSE of the Hybrid RF are improved to 99.98 and 0.011, respectively.

Table VI demonstrates the performance of the developed IDBN classifier with the input feature sets selected by the five FS methods. The ACC and RMSE of the IDBN are improved to 87.32 and 0.514, respectively.

TABLE VI.   THE PERFORMANCE EVALUATION OF THE IDBN MODEL

Models	Selected set			Dominant set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	86.55	0.571	28	86.55	0.571
SU	29	86.54	0.575	28	86.54	0.575
CHI	29	86.67	0.563	29	86.67	0.563
MI	32	87.11	0.545	32	87.11	0.545
CHIMI	32	87.32	0.514	32	87.32	0.514

B. Experimental Results with GDS4 Dataset

In this subsection, we do our experiment with an artificial dataset of a larger size of 10000 samples, dataset GDS4. We want to compare the performance of IDBN with other models using subsets of FS algorithms. The performance of the proposed models with the original dataset is shown in Table VII.

Table VII illustrate the staititscal results of the two metrics of KNN, Hybrid C5.0, Hybrid RF, and IDBN. The average of ACC and RMSE from serveral iteration run using GS4 dataset are recorded. The two developed tree-based models, Hybrid C5.0 and Hybrid RF generate the highest ACC and lowest RMSE, follow by the IDBN model.

From Fig. 11 and Fig. 12, we can obtain the results of ACC and RMSE of KNN models using a selected set of each FS method, and it is summarized in Table VIII.

From Fig. 13 and Fig. 14, we can obtain the results of ACC and RMSE of the Hybrid C5.0 model using a selected set of each FS method, and it is summarized in Table IX.

TABLE VII.   RESULTS OF PROPOSED MODELS ON ORIGINAL DATASETS

Proposed Models	KNN	Hybrid C5.0	Hybrid RF	IDBN
ACC (%)	95.12	98.55	98.88	97.01
Std. of ACC	0.942	0.578	0.312	0.666
RMSE	0.193	0.163	0.161	0.195
Std. of RMSE	0.016	0.049	0.028	0.015

Fig. 9.   ACC of IDBN using ADS Dataset.
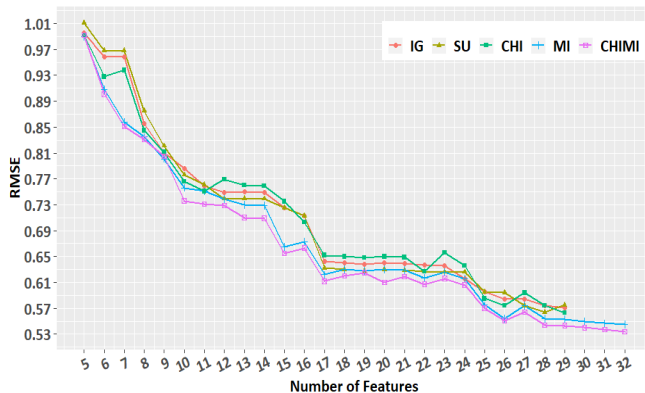
Fig. 10. RMSE of IDBN using ADS Dataset.
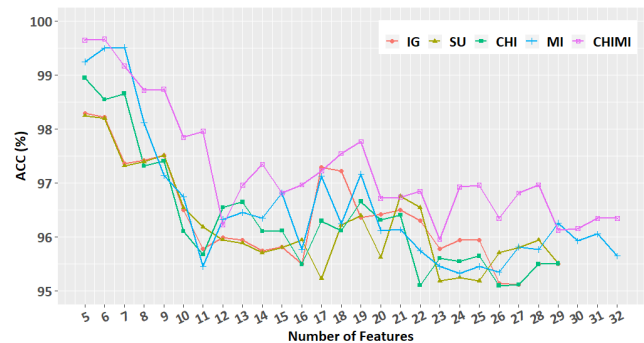
Fig. 11. ACC of KNN using GDS4 Dataset.
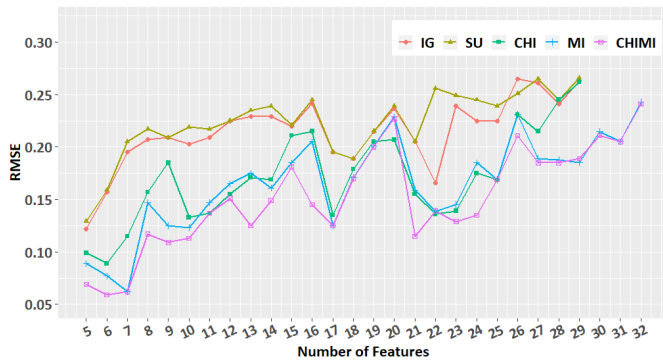
Fig. 12. RMSE of KNN using GDS4 Dataset.

TABLE VIII. THE PERFORMANCE EVALUATION OF THE KNN MODEL

Models	Selected set			Dominant set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	95.50	0.263	5	98.30	0.122
SU	29	95.52	0.266	5	97.25	0.129
CHI	29	95.51	0.262	6	98.95	0.089
MI	32	95.65	0.243	7	99.52	0.062
CHIMI	32	96.35	0.175	6	99.67	0.059

Fig. 13. ACC of Hybrid C5.0 using GDS4 Dataset.

Fig. 14. RMSE of Hybrid C5.0 using GDS4 Dataset.

TABLE IX. THE PERFORMANCE EVALUATION OF HYBRID C5.0 MODEL

Models	Selected set			Dominant set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.61	0.059	28	99.65	0.055
SU	29	99.60	0.058	28	99.63	0.057
CHI	29	99.71	0.047	29	99.71	0.051
MI	32	99.75	0.045	32	99.75	0.047
CHIMI	32	99.75	0.045	32	99.75	0.045

From Fig. 15 and Fig. 16, we can obtain the results of ACC and RMSE of the Hybrid RF model using a selected set of each FS method, and it is summarized in Table X.

From Fig. 17 and Fig. 18, we can obtain the results of ACC and RMSE of the IDBN model using a selected set of each FS method, and it is summarized in Table XI.

The results presented in Table VIII demonstrate the performance of the KNN models on feature sets and dominant sets selected by IG, SU, CHI, MI, and CHIMI methods on the GDS4 dataset. The KNN model with the dominant set of the proposed CHIMI achieves the highest performance. The ACC and RMSE of the KNN are improved to 99.67 and 0.059.
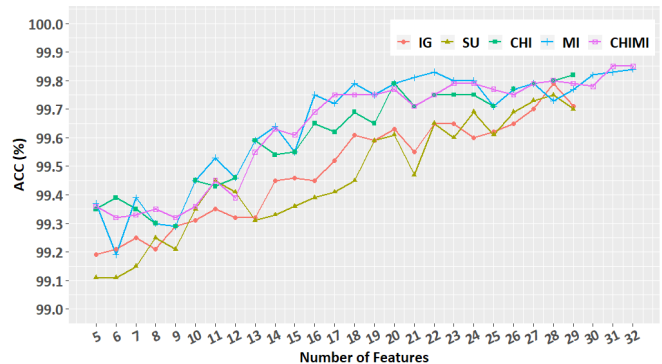
Fig. 15. ACC of Hybrid RF using GDS4 Dataset.
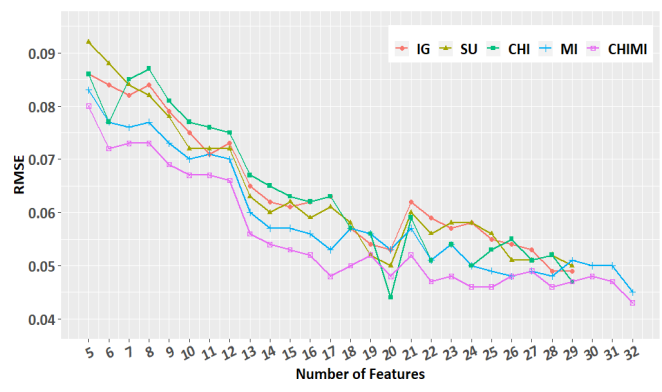
Fig. 16. RMSE of Hybrid RF using GDS4 Dataset.

TABLE X.   THE PERFORMANCE EVALUATION OF THE HYBRID RF MODEL

Models	Selected set			Dominant set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.73	0.051	28	99.79	0.049
SU	29	99.73	0.052	28	99.75	0.050
CHI	29	99.82	0.047	29	99.82	0.047
MI	32	99.84	0.045	32	99.84	0.045
CHIMI	32	99.85	0.043	32	99.85	0.043

TABLE XI.   THE PERFORMANCE EVALUATION OF THE IDBN MODEL

Models	Selected set			Dominant set		
	N	ACC	RMSE	N	ACC	RMSE
IG	29	99.65	0.052	28	99.67	0.050
SU	29	99.65	0.052	28	99.67	0.050
CHI	29	99.77	0.047	29	99.77	0.047
MI	32	99.81	0.045	32	99.81	0.045
CHIMI	32	99.82	0.044	32	99.82	0.044
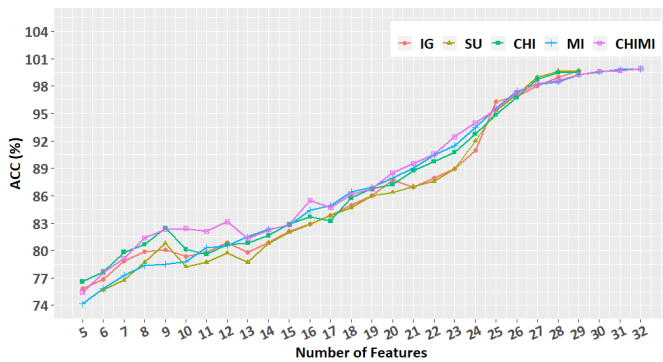
Fig. 17. ACC of IDBN using GDS4 dataset
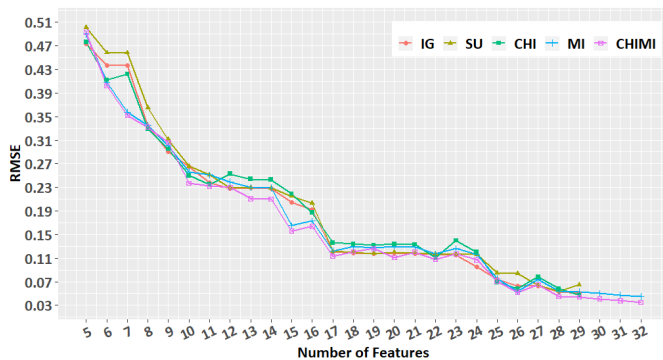
Fig. 18. RMSE of IDBN using GDS4 dataset

From Table IX, the performance of the Hybrid C5.0 is significantly improved when using the dominant set of the CHIMI. The ACC and RMSE of the Hybrid C5.0 are improved to 99.75 and 0.045, respectively.

The results of ACC and RMSE of the Hybrid RF are shown in Table X. The proposed CHIMI method outperforms the rest FS methods in achieving the highest ACC and lowest RMSE. The ACC and RMSE of the Hybrid RF are improved to 99.85 and 0.043, respectively.

Table XI demonstrates the performance of the developed IDBN classifier with the input feature sets selected by the five FS methods. The ACC and RMSE of the IDBN are improved to 99.82 and 0.044, respectively.

C. Summary and Discussion

This study aims to boost up the performance of the proposed classifiers to reach the most classification results. Hence, the optimal models are then combined with dominant sets, which is belive to significantly improve the performance of prediction models and selected the highly influencing factors in academic performance.

From Table II to Table XI, we can summarize the performance of KNN, Hybrid C50, Hybrid RF, and IDBN on dominant sets of IG, SU, CHI, MI, and CHIMI methods. Fig. 19 and Fig. 20 summarize the value of ACC and RMSE of each classifier on each FS method on data ADS and GDS4, respectively.

Fig. 19 represents the values of ACC and RMSE of the four classifiers with the five FS methods using the ADS dataset. Concerning prediction models, Hybrid C5.0 and Hybrid RF are comparatively better than the other two models. Regarding the FS methods, the performance of IG and SU methods are not statistically different. The performance of CHI and MI algorithms standout the performance of IG and SU. The figure reports that the proposed CHIMI successfully improves the performance of the four prediction models and standout the performance of the four FS methods.

Fig. 20 graphically demonstrate the performance of the four classifiers with the five FS methods using the GDS4 dataset. In the context of a larger dataset, the performance of IDBN is significantly improved. The performance of IDBN and Hybrid are not statistically different and the two models standout the performance of Hybrid C5.0 and KNN. The proposed CHIMI roles as the best FS method in selecting the dominant factors for improving the performance of the prediction models.
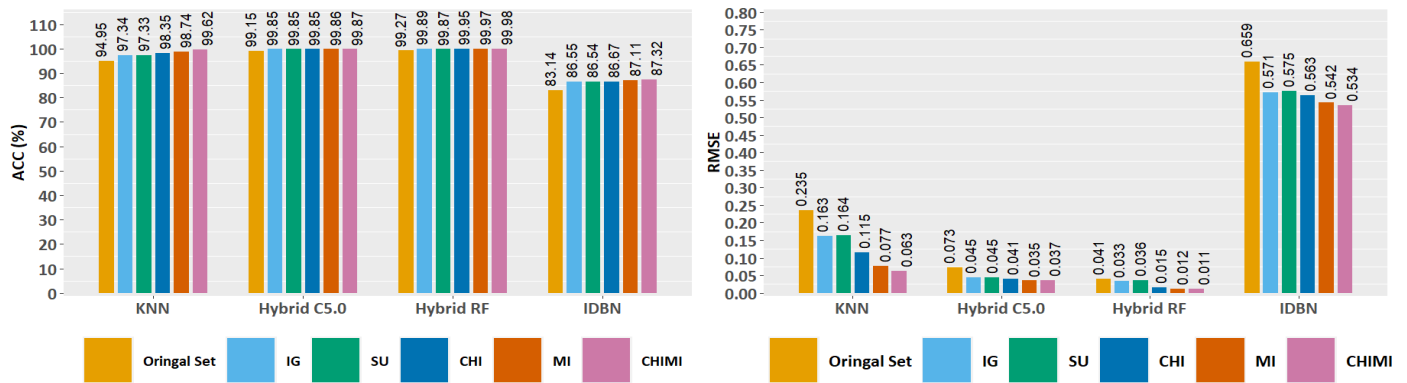
Fig. 19. ACC and RMSE Comparison using ADS Dataset.
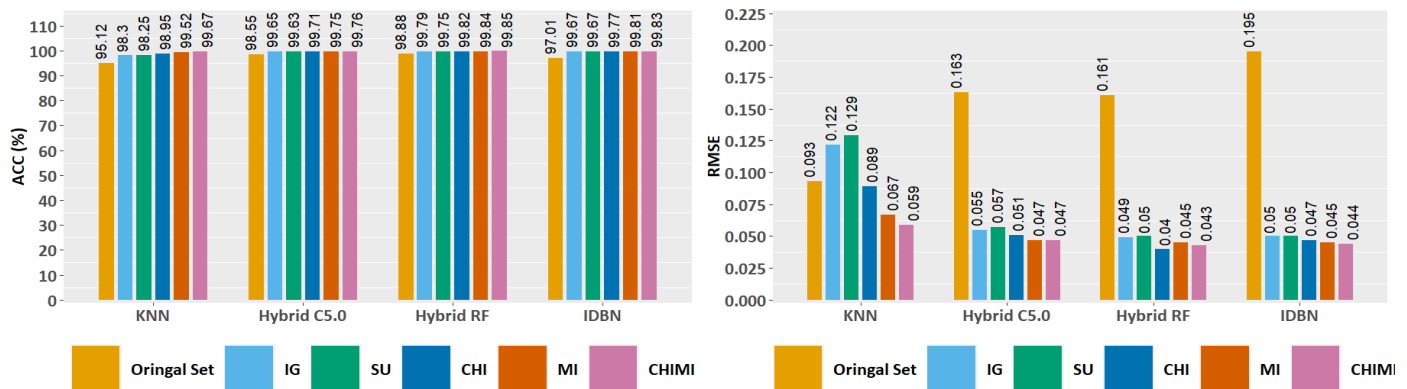
Fig. 20. ACC and RMSE Comparison using GDS4 Dataset.

V. Conclusion

Education is a crucial factor in the development of any country. Predicting student performance and mining their learning behaviors are challenging tasks in an educational environment. This paper presents a study of the analysis of dominant factors using feature selection methods and propose a novel feature selection algorithm for improving prediction performance to get the most successful classification results. The proposed CHIMI of the FS method significantly improves the performance of the proposed prediction models. The dominant set of the CHIMI method enhances the accuracy of prediction models from 1 to 5% increase and the RMSE from 0.06 to 0.2 decrease. The performance of the proposed prediction models reaches the superior classification results that can be effectively used to predict student performance. Once performance levels can be effectively predicted, hence, the at-risk group of students with poor-performing is identified, then they can be timely given intervention and additional assistance.

References

[1]   A. A. Saa, "Educational data mining & student's performance prediction," International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, 2016.

[2]   S. Slater, S. Joksimovic, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for educational data mining," Journal of Educational and Behavioral Statistics, vol. 10, Issue 3, pp. 85-106, 2017.

[3]   C. Romero and S. Ventura, "Data mining in education," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, issue 1, pp. 12-27, 2013.

[4]   P. Thakar, A. Mehta, and Manisha, "Performance analysis and prediction in educational data mining," International Journal of Computer Application, vol. 110, no. 15, pp. 60-68, 2015.

[5]   A. Pena-Ayala, "Educational data mining: Survey and a data mining-based analysis of recent works," Expert Systems with Application, vol. 41, pp. 1432-1462, 2014.

[6]   E. Sosmangegovic, M. Suljic, and H. Gic, "Determining Dominant Factor for Student Performance Prediction by Using Data Mining Classification", Tranzicija, Vol. 6, 2014, pp.147-158.

[7]   H. M. Hard and M.A. Moustafa, "Selecting Optimal Subset of Features for Predicting Student Performance Model", International Journal of Computer Science, Vol. 9, Issue 5, No. 1, pp. 253-262, 2012.

[8]   Jindal P., and Kumar D.., "A Review on Dimensionality Reduction Techniques", International Journal of Computer Application, Vol. 173, No. 2, pp.42-46, 2017.

[9]   L. Ma et al., "Evaluation of Feature Selection Methods for Object-Based Land Cover Mapping of Unmanned Aerial Vehicle Imagery Using Random Forest and Support Vector Machine Classifiers", International Journal of Geo-Information, Vol. 6, No. 51, 21 pages, 2017.

[10] P. Yildirim, "Filter Based Feature Selection Methods for Prediction of Risk in Hepatitis Disease", International Journal of Machine Learning and Computing, Vol. 5, No. 4, pp. 258-263, 2015.

[11] P. J. M. Estrera, et al, "Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School", International Journal of Engineering & Technologies, Vol. 3, No. 5, pp. 147-153, 2017.

[12] R amaswami M., Bhaskaran R., "A Study of Feature Selection Techniques in Educational Data Mining", Journal of Computing, Vol. 1, Issue 1, pp. 7-11, 2009.

[13] J. D. Febro, "Untilizing feature selection in identifying predicting factors of student retention," International Journal of Advanced Computer Science and Applications, vol. 10, no. 9, 2019.

[14] M. Zaffar and K.S. Savita, "A study of feature selection algorithms for predicting students' academic performance," International Journal of Advanced Computer Science and Applications, vol. 9, no. 5, 2018.

[15] A. Alhassan, B. Zaffar, and A. Mueen, "Predict students' academic performance based on their assessment grades and online activity data," International Journal of Advanced and Computer Science and Applications, vol. 11, no. 4, 2020.

[16] Mythili M.S., Mohamed S.A.R., An Analysis of Student Performance Using Classification Algorithms, IOSR Journal of Computer Engineering, Vol. 6, Issue 1, pp. 63-69, January 2014.

[17] Phyu T.Z. and Oo N.N, "Performance Comparison of Feature Selection Methods", MATEC Web of Conference 42, 2016.

[18] A. Bommert et al., "Benchmark for filter methods for feature selection in high-dimensional classification data", Journal of Computational Statistics and Data Analysis, vol. 143, 2020.

[19] Mazumder D.H, and Veilumuthu R., "An enhanced feature selection filter for classification of microarray cancer data", Wiley ETRI Journal, pp.358-370, 2018.

[20] C. F. Tsai and Y. C. Hsiao, "Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches," Decis Support System, vol. 50, no. 1, 258-269, 2010.

[21] A. Thubaity, N. Abanumay, S. Al-Jerayyed, A. Alrukban, Z. Mannaa, "The effect of combining different feature selection methods on Arabic text classification," IEEE: The 14th ACIS International Conference Software Engineering, Artificial Intelligent, Networking and Parallel/distributed Computing (SNPD), 211-216.

[22] S. Oulianroglou and G. Evangelidis, "Dealing with noisy data in the context of k-NN Classification,". Proceeding of the 7th Balkan Conference on Information Conference, Article ID. 28, pp. 1-4, 2015.

[23] P. Sokkhey and T. Okazaki, "Comparative study of prediction models for high school student performance in mathematics," Journal of IEIE Transactions on Smart Processing and Computing, vol. 8, no. 5, pp. 394-404, 2019.

[24] P. Sokkhey and T. Okazaki, "Multi-models of educational data mining for predicting student performance: A case study of high schools in Cambodia," vol. 9, no. 3, pp. 217-229, 2020.

[25] P. Sokkhey and T. Okazaki, "Hybrid machine learning algorithms for prediction academic performance," International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, pp. 32–41, 2020.

[26] P. Sokkhey and T. Okazaki, "Development and optimization of deep belief networks for academic prediction with larger datasets," Journal of IEIE Transactions on Smart Processing and Computing, (Accepted 20-April-2020).

[27] P. Sokkhey and T. Okazaki, "Developing Web-based Support System for Predicitng Poor-performing students Using Educational Data Mining Techniques," International Journal of Advanced Computer Science and Applications, vol. 11, no. 7, pp. 23–32, 2020.

A Novel Framework for Mobile Telecom Network Analysis using Big Data Platform

M. M. Abo Khedra[1]

Dept. of Computer Science
Faculty of Graduate Studies for Statistical Research
Cairo, Egypt

A. A. Abd EL-Aziz[2]

College of Computer and Information Sciences
Jouf University, Al jouf, KSA
Faculty of Graduate Studies for Statistical Research
Cairo University, Cairo, Egypt

Hedi HAMDI[3]

College of Computer and Information Sciences
Jouf University, Al jouf, KSA
University of Manouba, Tunisia

Hesham A. Hefny[4]

Vice-Dean of Faculty of Graduate Studies for Statistical
Research
Cairo, Egypt

Abstract—Social Network Analysis measures the interconnection between humans, entities or communities and the streaming of messages between them. This kind of Analysis studies the relationship between different people in a very deep way; it shows how one node (subscriber) in the network can affect the others. This research studies the connections between the customers in many different ways to help any telecom operator increase the cross and up-selling of its products and services as follows: detect communities of subscribers which are a group of nodes collected together to form a community, identify the connection types and label the links between the customers as (business, friends , family and others), as well as identifying the top influencers in the network who can spread positive or negative messages about products and services provided by the company through communities in the network and determine off-net customers who can be acquired to be targeted by specific marketing campaigns. A real cell phone dataset of 116 Million call detailed records of SMS and Voice Calls of an Egyptian Communication Service Provider (CSP) is used.

Keywords—*SNA; influencer; acquisition; community detection; link prediction; call detailed record; on-net node; off-net node*

I. INTRODUCTION

A Social network is established from the connection of nodes with each other. Much more nodes and links mean much more types of links as friends, colleges, dislike, likes, same perspective or financial exchange. Social Network Analysis (SNA) is a way of discovering social relation by utilizing the network, the structure of the network consists of two main features: vertices (accounts, subscribers inside the network) and linkage (relationships or interconnection) that link them. It analyses the graph of the social network to detect relationship types between subscribers or customers. Social Network Analysis has some metrics to define the importance and figure out the behavior of each node in the graph and structure of the whole graph. The SNA metrics include density, centrality, degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and clique. Data mining based techniques can be utilized to apply the concept of SNA as it is proved a significant impact in this area. SNA can be applied in any industry that can avail some sort of data includes a connection between entities (subscribers, accounts, etc.).

- **Health**

The authors in [1] and [2] tracked and prevented diseases spread, target individuals that are at high-risk positions in the network, track changes in the health status of the population over time, all of this accomplished by using the SNA.

- **Payment fraud detection**

For the financial industry, SNA has a key part to play in detecting fraud as to a large extent it is being committed by organized networks. What the analysis does is putting some context around the content of the data. In payments and banking, the social networks consist of accounts, card numbers and so on. Connections (ties) between these various accounts is the transactions of transfer to figure out the relations between customers in the social graph and to be analyzed to determine patterns that could indicate fraudulent behaviour.

- **Security**

The author in [4] utilized the SNA metrics to detect terrorist networks by tracing communication they made on the events of September 11, 2001, through slicing connections among the hijackers from news statements.

- **Telecommunications**

SNA can be applied easily in the telecom industry due to the availability of data any telecom operator had. A social graph established from using Call Detailed Records (CDRs) and subscribers demographic extracted from the data warehouse or any big data tool like Apache Hadoop. SNA can use this data to discover the hidden relationships between subscribers through representing subscribers as vertices and interactions (calls) as edges between those vertices (nodes). In [3], the author presented the business pains telecom operators had without using SNA. In addition, described how to utilize

the availability of huge good data to increase the Average Revenue Per User (ARPU). Moreover, used to detect influencers, enhancing operations of cross and up-sell campaigns. Telecom operators are facing some main problems when deciding to spend on offers and campaigns that much. Traditional targeting techniques have the drawback of treating every customer as an individual apart from his social interactions and relationships. These social relationships have a great effect on individuals' opinions and usage of products and services. They do not know the relationship between customers (caller and callee) so they cannot be able to make specific offers to customers who are related with each other as a family, friends, co-workers or others. Telecom operators cannot detect the top influencer customers who can attract others to be in the same community. Moreover, the most important off-net customers to be acquired. Who is going to churn from the company, what else they can sell them? Thus, affecting their marketing costs and campaigns ROI. Our first and foremost task is implementing new proposed scientific techniques of Social Network Analysis on large telecom social network to create end-to-end generic SNA solution. To be able to use in the telecom industry to detect communities of customers, Identify relationships between the subscribers of the network, Identify influencer and key players in the network, identify the Ideal members that have the highest probability for acquisition.

The main contribution is proposing a generic, fully integrated, scalable, end-to-end SNA solution built for organizations to gain more insights on their customers from a social perspective. The introduced solution architecture starts with Community detection to identify the one big community for each consumer, which includes all his/her, interactions with other consumers. Then proposed a new Business based Rules for the Community Labelling, Through the extraction of new derived features from the data set to determine the type of relationship between the subscribers in the network such as Family, business, friends or others according to set of business behaviors of customers. Furthermore, identification of influencers in each community by implementing two algorithms used for various types of influence. Finally The creation of a new mathematical equation to acquire off-net customers. The business value behind that is as the following: Offer the customers suitable packages according to the communities they belong to whether it is family, friends or workmates and increasing the company market share by bringing in new subscribers.

The rest of the paper is organized as follows: Section II listed the related work for SNA, Section III explained the full solution architecture and studied the different techniques for implementing SNA, Section IV discussed the results of the techniques used in SNA and finally, Section V presented the conclusion.

II. RELATED WORK

In this section, a literature survey is done for social network analysis, community detection, community labelling and influencer detection.

A. Social Network Analysis

Due to the big size of social network data, there is a need for a big data architecture to host the data and make the analysis on top of it. The author in [5] discussed this issue as they proposed a new prototype platform of SNA upon the architecture of big data using Spark GraphX framework in combination with JavaScript. The author in [7] proposed a study for handling the centrality analysis by using uncertainty methods, they integrated multiple centrality measurements by using fuzzy set and MYCIN theory. The author in [8] used mixed integer linear programming models to a dataset extracted from the social network. This research used two MILP clustering models M1 and M2. M1 model minimizes the maximum diameter between the cluster diameters; on the other hand, M2 reduces the total distance among objects related to the same cluster. M1 has computation time lower than M2.

B. Community Detection

Community detection is one of the key concepts in SNA. Nodes that tied together are grouped to build a cluster or a community. The author in [19] proposed a study for handling the centrality analysis by using uncertainty methods, they integrated multiple centrality measurements by using fuzzy set and MYCIN theory. The author in [6] utilized the clustering technique and centrality measure to detect link intensity among Brazilian Scientific researchers, Researches act as edges and researchers act as nodes. The author in [12] used the Girvan Newman algorithm to detect communities by progressively removing edges from the original network. The author in [13] used K-cliques algorithm for finding overlapping communities that is the combination of two cliques that share a common area that contains (k-1) node. The author in [14] used Fast Unfolding algorithm that unveils hierarchies of communities and allows zooming within communities to discover sub-communities, sub-sub-communities.

C. Community Labelling

The author in [18] predicted links in social networks using deep learning. Proposing computationally efficient and network-size-independent feature vector with deep learning that is fit for the real-time application. The author in [16] used a supervised algorithm (SVM) for relationship identification which gives a low accuracy (55%) and they admit that a specifically designed model for relationship identification will work better. The author in [11] listed some rules to characterize a group of subscribers into one of three different categories (Youth, Corporate, and Homebound). For example, if groups of subscribers have a high usage of messaging service with long duration calls at night and less frequent calls are labelled as Youth. Nevertheless, he did not mention anything related to the below point, which leads to an increase in the accuracy for the community labelling:

- Age difference (between the caller and the receiver, divide it by ranges, this would help to identify friends from families – ex-parent and child have a difference greater than 15 years).

- Home and work location (while taking into consideration that the other user maybe off-net and we would not know either age or home location).

D. Influencer Detection

The detection of influencers is a recent line of research in SNA. An Influencer is the most connected node (customer) in the network. The author in [9] used the HITS (Hyper-link Induced Topic Search) to detect the most important pages by analyzing the links and rates the Web pages. The author in [10] used the IP algorithm on twitter dataset to detect the influencers. The author in [15] used the Limited Recursive Algorithm to calculate social network influence for users on the social network Twitter and Facebook. The author in [17] focused on visualizing and measuring the different metrics of social network and based on those values, it can detect the most influential node.

III. METHODOLOGY

A. Proposed Framework

The framework based on Big Data tools, which contains four layers as shown in Fig. 1:

- Storage layer: a real data came from one of the communication service providers (CSPs) in Egypt to be loaded by the extract transform and load (ETL) process into a relational database management system (RDBMS), and then transferred to the Hadoop HDFS.

- Processing (analysis) layer: Apache Spark is a well-designed engine used to enhance graphs execution. A group of tools built on top of apache spark that contains a tool to handle structured data processing and SQL queries called Spark SQL, a tool for the analysis and mining models called PySpark that used with Python programming language, a tool for graph processing called GraphX.

- Post-Processing Storage layer: The output from the processing layer stored in the PostgreSQL and Neo4J for further analysis and visualization.

- Visualization layer: This layer uses Business Intelligence (BI) software to provide insights on the results of the analysis using reports and dashboards.

B. Data Collection

A cell phone dataset is used containing aggregated Call Data Record (CDR) for three months per the hours of the call for weekdays from (Sunday to Thursday) and the weekend includes (Friday and Saturday) per direction (incoming or outgoing) for 116 Million records CDRS of SMS and Voice Calls. The dataset also contains 12 Million records for customers' profiles that map to the aggregated CDRS with 26 columns like (age, location of the home, work, etc.). The caller and callee represent a network of nodes linked by Calls or SMS. The network based on the centrality measures and adjacency matrix.
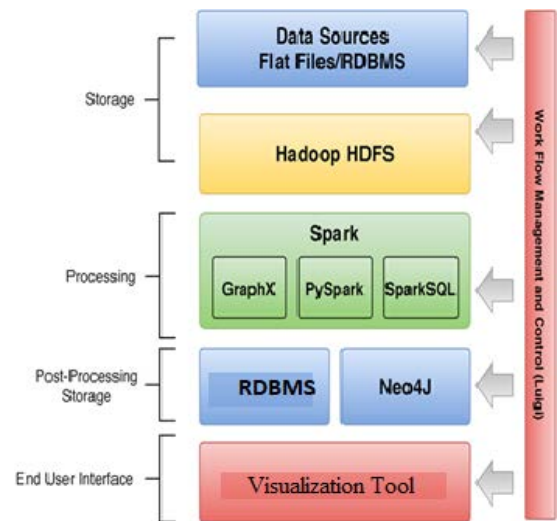
Fig. 1. Solution Architecture for SNA.

C. Social Network Analysis Techniques

Many techniques settled under the umbrella of SNA, the following sub-sections illustrate the flow of SNA methodologies and the approaches used to detect communities, community labelling, define key influencers and acquiring a new customer as shown in Fig. 2.

1) Community detection: Communities or clusters are a number of nodes collectively having a great chance of being linked to each other than to nodes of other groups.

There is no unique definition of a community so far which is widely accepted in social networks. A variety of definitions for a community has been proposed according to different aspects, which can be mainly classified into three main categories: intuitive definition, functional definition and definition from the process of an algorithm.

Community detection can support in predicting churners in a community, i.e. if the rate of churners increases in this community so most probably much more members will churn in a short time, so the operators have to stop that.

The following two algorithms are the most common methods to find out the communities from a given huge telecom network:

- Fast unfolding (Louvain Method) algorithm

- Label propagation (LPA) algorithm.

Both of them represents the heuristic method for greedy modularity optimization. In order to assess the performance and outcome of each technique, both algorithms ran on a sample of nodes and compared the output (nodes divided into clusters) as shown in Table I.

According to the NMI value that measures the accuracy as shown below in Fig. 3 and results in Table I, The Label Propagation algorithm used as a factor of initial splitting to discover communities and built on top of the community labelling.
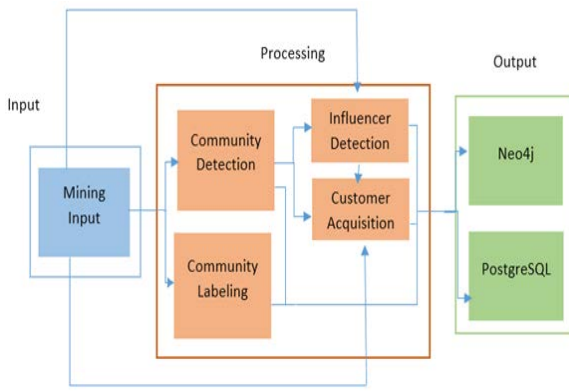
Fig. 2.   The Flow of SNA Techniques.

2) Community labelling: The interactions between every two nodes "edge" indicate some distinct behaviour between them. Based on the interaction weight and time, and by using some defined rules, the edges between nodes could be classified into multiple categories or labels e.g. Friends, family, business or others. These labels could be used further to guide business strategies and plans to maximize the profit e.g. Promos and offers. By applying this technique in the telecom industry, it found that the structure of customer communication networks provides a natural way to understand customers" relationships and to a certain extent the behaviour of highly connected customers. The intention is to identify the type of call whether it is a family, friend, or business. The business value behind community labelling could be by offering the customers suitable packages according to the communities they belong to whether it is family, friends or workmates.

TABLE I.   COMPARISON OF LABEL PROPAGATION AND FAST UNFOLDING
ALGORITHM

Measure	Label propagation	Fast unfolding
Average degree per clusters	1.5363	1.5485
Average density per cluster	0.2286	0.1827
Average in degree per cluster	0.7693	0.7746
Average out degree per cluster	0.7692	0.7746
Average closeness centrality	0.2550	0.2331
Average betweenness centrality	0.0598	0.0080
NMI Value	0.912	0.435
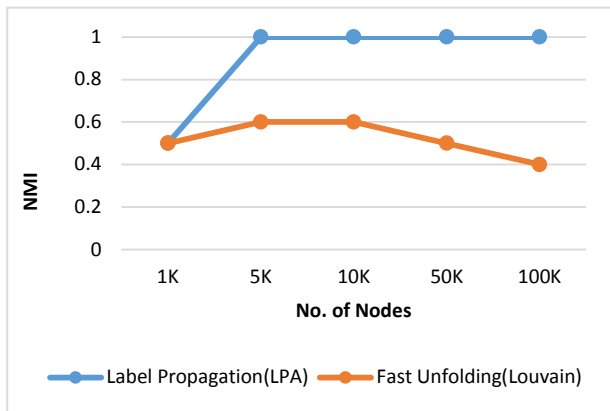
Fig. 3.   Comparison of LPA and Louvain Algorithm.

A new business based rules proposed with taking into consideration age difference, location and some customers might be both a work colleague and a friend.

The day is categorized into four-time slots with different usage pattern:

- 6 am – 8 am (Early morning).
- 8 am – 5 pm (Working hours).
- 5 pm – 12 am (Early evening).
- 12 am - 6 am (late evening).

The Proposed Business based Rules for Community Labelling are as follows:

Family

- v_low_EM_WD =1.
- v_low_EE_WD =1 or v_medium_EE_WD =1.
- v_low_WE =1 or v_medium_WE =1.
- SMS_low_WD=1.
- SMS_medium_WE=1.
- low_age_difference =1 or high_age_difference=1.
- same home location = 1.

Friends

- v_low_EM_WD =1 or v_medium_EM_WD=1.
- v_medium_EE_WD=1 or v_high_EE_WD=1.
- v_medium_LE_WD=1 or v_high_LE_WD=1.
- v_medium_WE=1 or v_high_WE=1.
- SMS_medium_WD =1.
- SMS_high_WE=1.
- low_age_difference=1.

Business

- v_low_EM_WD=1.
- v_medium_WH_WD=1 or v_high_WH_WD=1.
- low_age_difference =1 or medium_age_difference=1.
- SMS_low_WE=1.
- SMS_high_WD=1.
- same work location =1.

3) Influencer detection: Influencer detection module has the ability to identify the most Influential persons in a telecom's operator network. Telecom operators pay attention to identify the influencers in each community to offer them a suitable promotion. Influencers can use viral marketing to talk about the advantages gained from telecom operators they belong. Several companies have produced tools, which supports marketing departments demand to detect leading

users who make some influence on the followers and who will spread more effectively information about the products. This is necessary for the recent trend to invest in the so-called Word of Mouth Marketing, where companies need to discover the influencers who will spread more effectively information about the new products. Which lead to increase the revenue In case of detecting and ranking influential Customers, the following algorithm is the most common method to detect the influencers when coming to the telecom industry and have a distributed implementation on Hadoop unlike the other algorithms, we can implement it on map reduce:

- Page rank: mathematical method

The PageRank algorithm is based on using the hyperlinks as an indicator of a page's importance. Every unique page is assigned a page rank. If a lot of pages' vote for a page by linking to it then the page that is being pointed to will be considered important.

4) Customer acquisition: The telecommunication operators networks could have some off-net customers who can be acquired to further increase the profit and the market share. Using the existing customer base to spread products and ideas beyond the current network borders. The foremost task is to identify top potential off-net consumers to join the network if they were targeted with the right campaign. The acquiring criteria depending on the acquisition score of the off-net customers, more score, more probability to be acquired.

SNA based score formula used to determine the network value for the off-nets. Each node (off-net consumer) assigned a score indicating its network value. The proposed equation to decide which off-net customers to acquire using Eq. (1).

Net Value = $\sum$ (weight of direct neighbors in the community) $\times$ community density (Number of Links divided by Number of Vertices)   (1)

IV. RESULTS AND DISCUSSION

The results presented in this section are obtained from applying the algorithms mentioned in Section III.

A. Community Detection

The output extracted from Label Propagation algorithm used to decide which communities would be targeted by the new campaign. The related customers grouped together to represent one community as shown in Fig. 4.

B. Community Labelling

A social graph created G (V, E) with mobile phone subscribers as the vertices and with edges representing the social links, where (u, v) ∈ E represents the directed link from vertex (U, V) ∈ E to vertex v ∈ V and exists if u initiated a call to v. This graph is built from our set of call detail records as follows. For each region, we collected subscribers after eliminating outliers like automated calling machines and subscribers with low call counts, we pooled subscribers from all the regions and ranked them by their total number of calls

and removed the top and bottom 10%. We then uniformly selected a sample set of 100,000 subscribers to our vertex set V. For each sampled subscriber we added their immediate neighbors to the vertex set and created the directed edges using our set of call detail records. Table II shows the final distributions of the relationship types in our directed graph.

C. Influencer Detection

A sample of nodes examined using the page rank algorithm and the output listed the node id and the authority score per node as shown in Table III.

To interpret the result, the output sorted and extracted the top 10 nodes (can be substituted later on by percentile); the selected nodes would be used as our influencers. Page rank algorithm denotes some node ids that are the new influencers where the telecom companies need to find them out to pay attention.

D. Customer Acquisition

Equation (1) that is mentioned in the customer acquisition helps in identifying which off-net customers would be most profitable for us to acquire. It requires a filtered input based on node type (on-net or off-net) as well as output from the influencer detection algorithms used as well as calculating the community density. The output is a list of recommended off-net nodes to acquire along with their company belong to as shown in Table IV.
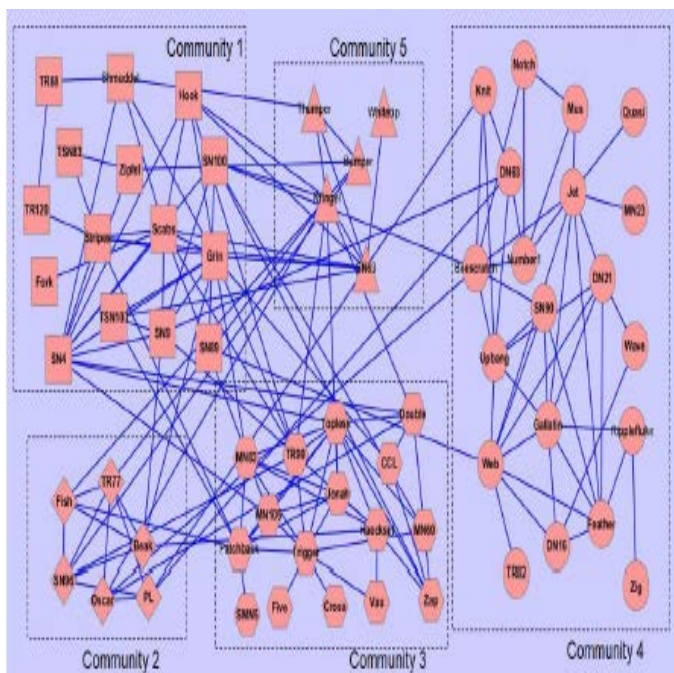
Fig. 4. Output of Community Detection.

TABLE II.   OUTPUT OF COMMUNITY LABELLING

Family Edges	63,350
Business Edges	191,581
Friends Edges	185,214
Others Edges	5267

TABLE III.   OUTPUT OF PAGE RANK ALGORITHM

Page rank Top 10 Scores	
Customer Number	**Authority Score**
599000001	0.31637741
599012461	0.066292251
599013241	0.043618953
599012575	0.039831044
599012839	0.025255006
599011440	0.024302746
599011441	0.024302746
599011817	0.022412784
599010494	0.017147495
599010613	0.017147495

TABLE IV.   OUTPUT OF CUSTOMER ACQUISITION

Off-Net Customer Number	Company Belong To
140254695	Etisalat
157372237	Orange
157372361	Orange
140254731	Etisalat
140254788	Etisalat
50651513	Orange
140254714	Etisalat
157372449	Orange
157373063	Orange

V. CONCLUSION

In this paper, a number of significant challenges pertaining to social network analysis and its implementation in the telecom operators. By focusing on building end-to-end generic SNA solution using big data tools that can be used in the telecom industry. SNA uses the call detail records (CDRs) of the Customer Service Provider (CSP) in conjunction with other data sources like customers' profiles to analyze the social relations of their customers. The paper has a full implementation of the following SNA key points: Community detection to find a community for each subscriber with (90%) accuracy, secondly, to detect the relationship between customers as (friends, family, co-workers, others) to make specific offers, thirdly, to identify the main influencers in the network who can convince other off-net customers to join the company with 99% accuracy and finally created a formula to acquire new customers.

The main goal is to support the decision-makers in the telecom industry to make the right decisions regarding the campaigns and offers to the right customers by providing them with the gaps in their relationships with customers and giving a three-sixty view about the customers through Business object tool for reports and dashboards.

We like to extend this research work by predicting rotational churn clients who left the company and returns back to the company but with a better offer than he took. In addition to, the churn prediction as one of the SNA key points to detect who is the client that has the intent to leave the company.

REFERENCES

[1] I. J. Saldanha, T. Li, C. Yang, C. Ugarte-Gil, G. W. Rutherford, and K. Dickersin, "Social network analysis identified central outcomes for core outcome sets using systematic reviews of hiv/aids," Journal of clinical epidemiology, vol. 70, pp. 164–175, 2016.

[2] K. E. Poundstone, S. A. Strathdee, and D. D. Celentano, "The social epidemiology of human immunodeficiency virus/acquired immunodeficiency syndrome," Epidemiologic reviews, vol. 26, no. 1, pp. 22–35,2004.

[3] S. Doyle, "Social network analysis in the telco sector marketing applications"Journal of Database Marketing & Customer Strategy Management, vol. 15, no. 2, pp. 130–134, 2008.

[4] V. Krebs, "Uncloaking terrorist networks," First Monday, vol. 7, no.4,2002.

[5] I. Sori´c, D. Dinjar, M. ˇ Stajcer, and D. Oreˇsˇcanin, "Efficient social network analysis in big data architectures," in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017 40th International Convention on. IEEE, 2017, pp. 1397–1400.

[6] V. Str¨oele, F. Campos, J. M. N. David, R. Braga, A. Abdalla, P. I. Lancellotta, G. Zimbr˜ao, and J. Souza, "Data abstraction and centrality measures to scientific social network analysis," in Computer Supported Cooperative Work in Design (CSCWD), 2017 IEEE 21st International Conference on. IEEE, 2017, pp. 281–286.

[7] X. Zuo, B. Yang, and W. Zuo, "Exploring uncertainty methods for centrality analysis in social networks," in Data Mining Workshops (ICDMW), 2017 IEEE International Conference on. IEEE, 2017, pp. 163–169.

[8] H. Pirim, "Mathematical programming for social network analysis," in Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017, pp. 2085–2088.

[9] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze (2008). "Introduction to Information Retrieval". Cambridge University Press. Retrieved 2008-11-09.

[10] Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011, September). Influence and passivity in social media. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 18-33). Springer, Berlin, Heidelberg.

[11] SARAVANAN, M., et al. Labeling communities using structural properties. In: 2010 International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2010. p. 217-224.

[12] Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002).

[13] Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3-5), pp. 75-174.

[14] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), pp.100-108.

[15] Hajian, B., & White, T. (2011, October). Modelling influence in a social network: Metrics and evaluation. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing (pp. 497-500). IEEE.

[16] Yu, M., Si, W., Song, G., Li, Z., & Yen, J. (2014, August). Who were you talking to-Mining interpersonal relationships from cellphone networkdata. In 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014) (pp. 485-490). IEEE

[17] Farooq, Aftab, Gulraiz Javaid Joyia, Muhammad Uzair, and Usman Akram. "Detection of influential nodes using social networks analysis based on network metrics." In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1-6. IEEE, 2018

[18] Chiu and J. Zhan, "Deep learning for link prediction in dynamic networks using weak estimators," IEEE Access, vol. 6, pp. 35937-35945, 2018.

[19] Zuo, B. Yang, and W. Zuo, "Exploring uncertainty methods for centrality analysis in social networks," in Data Mining Workshops (ICDMW), IEEE International Conference on. IEEE, pp. 163–169, 2017.

Application of Kinect Technology and Artificial Neural Networks in the Control of Rehabilitation Therapies in People with Knee Injuries

Bisset Gonzales Loayza[1], Alberto Calla Bendita[2], Mario Huaypuna Cjuno[3], Jose Sulla-Torres[4]

Escuela de Ingeniería de Sistemas, Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

Abstract—**In the field of physiotherapy, the recognition of the poses of the human body is obtaining more research so that the patient has an accelerated recovery rate in his rehabilitation. Nowadays, it is not so challenging to have devices like Microsoft Kinect that allow us to interact with the user for the recognition of poses and body gestures. The objective of this work to capture the data of the joints of a person's body through a set of angles using the Kinect device, then artificial neural networks with the Back-Propagation algorithm were used for machine learning, and their precision was determined. The results found on the performance of the neural network show that 99.70% accuracy was achieved in the classification of the patients' postures, which can be used as an alternative in the rehabilitation therapies of patients with knee injuries.**

Keywords—Machine learning; artificial neural network; kinect; physiotherapy; rehabilitation

I. INTRODUCTION

The most common injuries in a person's early stage are bruises and fractures, not to mention that they can present pain from different anatomical structures, such as osteochondritis of the knee (Osgood Schlatter disease) [1]. Knee injuries are common in children, and in people who engage in some form of sports activity or overuse their knee, most studies report traumatic knee injuries [2].

According to the National Institute of Statistics and Informatics (INEI), in Peru, 5.2% of the population (1 million 619 thousand people) have some type of disability, "88.6% of the population with a disability did not receive treatment or therapy for rehabilitation and only 11.4% if they received any treatment or therapy. Among those who received treatment or rehabilitation therapy we can mention physical rehabilitation therapies (46.1%), psychological treatment (18.9%), psychiatric treatment (11.3%), speech therapy (11.0 %), emotional support (3.8%), occupational therapy (3.6%), another type (5.4%)".

The rehabilitation process is commonly carried out in a rehabilitation center, in which two main problems arise: a) in the rehabilitation environment, the patient is motivated by the therapist's guidance, while at home he is completely unmotivated, and b) some patients may perform incorrect exercises by compensating for limited movement with other stereotypes. Knee injuries, which are the most common disabling injuries in athletic and physically active individuals,

can be treated with the use of knee braces and bandaging techniques that are widely used to reduce and / or prevent the severity and incidence of injuries of knee as well as the use of kinesiotaping to improve muscle strength and jump performance in the knee [3]. This is one of the main problems that patients can get bored from repeated rehabilitation activities or cannot properly perform the exercises. This can negatively influence patients to enable them to undergo rehabilitation.

On the other hand, there are several studies where the Kinect tool is used to apply its benefit to different forms of rehabilitation [4] such as the proposal to improve specific rehabilitation such as flexion, abduction and extension of the limbs through digital applications [5], in conjunction with artificial neural networks (ANN) as new approaches for the medical evaluation of injured patients [6].

In this context, the proposal is to carry out a prototype system for monitoring/training physical physiotherapy of patients with knee injuries. This system is based on artificial neural networks, where the input data corresponds to the detection of the joints that belong to the knee; for this, the Microsoft Kinect v1.0 tool is used, which provides exact measurements of this data.

The rest of this document is organized as follows: Section II, review of related works on movement recognition with Kinect; Section III, a summary of methods, data set, neural network model; Section IV presents experimental results on the recognition rate of human movements and the effectiveness in monitoring the patient, and the discussion regarding other works. Finally, we conclude the results in Section V.

II. RELATED WORK

The recognition of human movements has been studied for several years [7], and the techniques for machines to learn to recognize them have been evolving. It is where the concept of Machine Learning appears, which, through its techniques, have helped improve the performance and precision of machines to recognize specific human activities [8] [9].

In the study by Da Gama et al. [10], a review of interactive systems is made to help patients in various therapies. Among these technologically advanced systems is Microsoft's Kinect, which has helped lead the way in how user interaction

technology facilitates and complements many clinical applications.

Another work presented by Morando et al. [11] makes use of various technologies such as Kinect, Leap Motion, Band 2. The work presents a whole system that helps the patient to perform his rehabilitation exercises based on serious games. Some indicators are defined that provide information about the patient's performance and an upcoming evaluation to determine if the patient performs an adequate or wrong movement. In [12] it is corroborated that the Kinect allows the measurement of patient movement during training and exercise, providing good quality medical images with sufficient precision for clinical practice, in addition to being less expensive than most medical detection devices, which made it viable and a good fit for this job.

Other studies such as in [13] that propose techniques to improve the precision of gesture recognition and movement analysis, for this, based on the Kinect data, they extract three essential characteristics for physiotherapy exercises: body posture, movement trajectory, and range of movement. With these data, using techniques such as Hidden Markov Model (HMM) and Dynamic Time Warping (DTW), improved accuracies were obtained with up to 56% for HMM and 32% for DTW.

There are several works in this regard, such as in [14], which is one of the first to use two techniques, such as Support Vector Machines (SVM) [15] and Random Forests (RF) to classify the cinematic activities read by Kinect accurately. This study establishes a comparison in terms of classification performance between these two techniques, wherein the end, it is concluded that RF exceeds SVM in precision, but SVM requires less time to train compared to RF.

Neural networks have been used in [16] where a method for estimating human pose in real-time for multiple people from video using convolutional neural networks has been presented. In other work of Morando et al. [17], which is the continuation of [11] where Machine Learning techniques are used to recognize human movements and evaluate patient performance in an automated way, the same idea has come in several recent research but using newer techniques.

The reliability and validity of a new software program based on Kinect is seen in the work of Ressman, Rasmussen-Barr an Grooten [18], the objective of this study was to establish the test-retest reliability and the validity of the Qinematic (TM) construct to evaluate the activity of the leg. The results of the construct validity study indicate that Qinematic (TM) at 6 degrees of medial displacement can identify subjects with a knee-standing position.

As described, most jobs vary based on knee injury problems. From the reviewed studies, we saw that an alternative way to support knee injury rehabilitation is by using a Kinect device. Also, the data that is captured with the Kinect device can be adequately managed using machine learning techniques such as artificial neural networks.

III. MATERIALS AND METHODS

The present investigation is exploratory and descriptive. The perception of therapists in hospitals for the recovery of knee injuries in patients is analyzed, as well as the characteristics of their body postures. In this way, the facilities for rehabilitation in people's health are identified.

A. Materials

The following tools were used for software development with Kinect and with the c # programming language:

- Kinect v1.0 sensor.
- Visual Studio 2017.
- Kinect for Windows SDK v1.8.

B. Description of the Ontology of the Proposal

Our proposal consists of a physiotherapeutic rehabilitation monitoring system; for this, we use the Kinect sensor to capture the user's joints (corresponding to the lower limbs) in real-time. On the other hand, we have artificial intelligence in charge of recognizing and classifying the movements that said user is making during his physical rehabilitation session. To understand the problem domain, an ontology was designed. Fig. 1 shows all the elements that intervene in the functionality of the system and how they are related, all based on classes, class properties, associations or dependencies, and inheritance.

As can see in the hierarchy, several elements intervene, but of all of them, the main classes are the ones that list:

1) User: General information about who the user is, plus the anatomical parts of the body and joints observed: Body classes, part of the body, part, limb, arm, leg, bone and joint: Classes Body, BodyPart, Part, Limb, Arm, Leg, Bone, and Joint.

2) Sensor: Raw data provided by an acquisition device, such as a sensor like Kinect or other.

3) Gestures and poses: Gesture, GestureSegment, and Pose classes.

4) NeuralNetwork: It provides the machine learning model, as well as the algorithm to be trained and recognize or predict a posture.
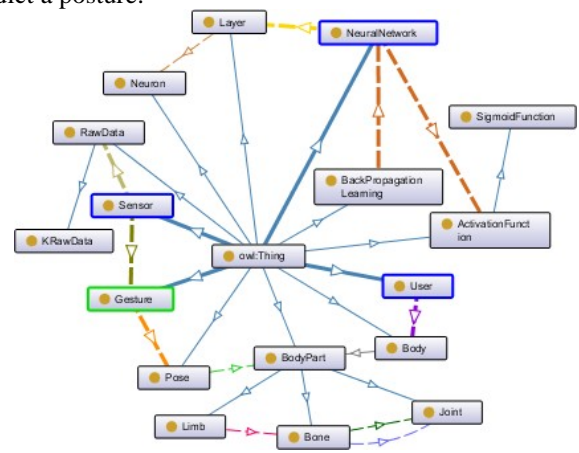

Fig. 1. Ontology of the system.

C. Description of the Physiotherapeutic Monitoring System

Our proposal follows a three-layer architecture. As seen in Fig. 2, the left part of the system interacts with the physiotherapist through a web application that can be accessed from a PC or smartphone; this provides maintainability and scalability; while on the other hand, the patient interacts with an application based on Windows Presentation Foundation (WPF) of Microsoft.NET to create Windows applications of secure development. This application is connected to a Microsoft Kinect v1.0 sensor. Communication between the application server and web client is done through HTTPS, which is recommended for security, since the application treats sensitive private data of people, such as the health of a patient.

Table I shows the data dictionary used in the system.

Both the physiotherapist and the patient must validate their user to enter the system; in the case of the patient, this is very important since their account also maintains information (stored in a database) about their progress with rehabilitation exercises. The physical therapist reviews such progress. The WPF application is the one that implements the Back-Propagation algorithm in order to recognize the patient's movements and detect if the patient performs the proposed exercise correctly or incorrectly.
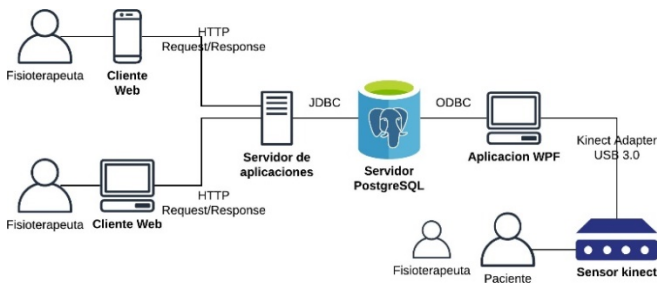
Fig. 2. System's Architecture Proposed.

TABLE I. DESCRIPTION OF DATA USED IN ARCHITECTURE

Entity	Attribute	Type	Length
Patient	Name	varchar	30
Patient	Mother's last name	varchar	30
Patient	Last name	varchar	30
Patient	DNI	Integer	8
Patient	Age	Integer	2
Physiotherapist	Name	varchar	30
Physiotherapist	Mother's last name	varchar	30
Physiotherapist	Last name	varchar	30
Physiotherapist	DNI	Integer	8
Therapy	Code	Integer	4
Therapy	Names	varchar	30
Therapy	Date	Date	-
Therapy	Hour	Datetime	-
Therapy	Punctuation	Float	5, 3
Position	Code	Integer	4
Position	Name	varchar	30
Position	Value	Float	

D. Artificial Neural Networks

The method used to classify human postures involves using an artificial neural network with dense layers. Entries for ANN are expressed by feature vectors associated with human poses captured by Kinect.

Likewise, the output of ANN data is expressed by values 0 and 1, which corresponds to an incorrect and correct posture, respectively. The output is calculated by an activation function f (1). ANN training was performed using the Back-Propagation algorithm, as shown in Eq. 1.

$$sigmoid\ function: f(x) = \frac{1}{1+e^{-ax}} \tag{1}$$

Where: α is a parameter of the model.

On the other hand, for the development part of the neural network, it has been decided to use AForge.NET [19]. It has a simple API for the construction of neural networks with different algorithms to train, such as BackPropagation, DeltaRule, ElasticNetwork, among others. AForge.NET is used as a tool to develop the motion detection module, and it is comprised of a set of libraries that are useful for neural networks. Therefore, C # was used for development, and it is easy to integrate into Aforge.NET, the tool selected for the motion detection module [20].

Fig. 3 shows the main components of the recognition system.
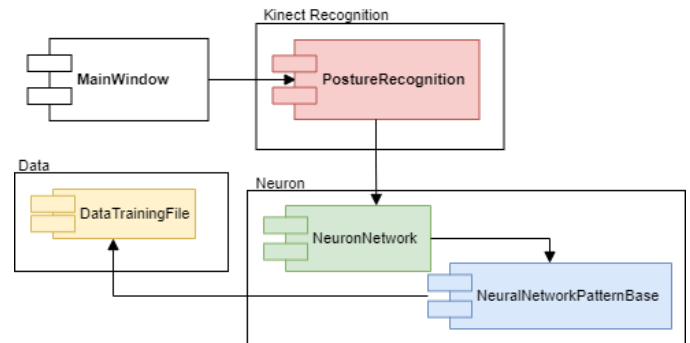
Fig. 3. Main Components of the Recognition System for this Work.

- MainWindows: Represents the WPF application that instantiates components of the Kinect SDK. Displays a frame that draws the skeleton in interconnected green segments. Also, here the posture recording processes are executed, managed by concurrent flags and threads. Kinect v1 provides 30 FPS, and since each frame represents a SkeletonFrame (containing the required Skeleton object), it can achieve 30 Skeleton objects per second.

- PostureRecognition: Instances a NeuronNetwork object, in addition to having the main training and predict methods, to be used directly by a user.

- NeuronNetwork: Uses the ActivationNetwork class, which is the neural network itself, and creates an instance of BackPropagationLearning (both AForge.NET classes) that executes the neural network training.

- NeuralNetworkPatternBase: Provides how data is processed before entering the neural network. Our approach is based on the use of angles as in [21], that is, we take the Skeleton objects stored somewhere, then we access the Joins properties of each object, and with these data, we calculate these angles, which serve as inputs to the neural network.

E. Data Collection and Exercises

The data set that we use for the training and evaluation of the model was obtained from the postures performed by the patients with the supervision of a specialist. These data were captured through a Microsoft Kinect device, where the time of each session lasts between 7 and 15 seconds, where the correct posture is first recorded, and then the patient performs the postures to determine if they are doing it correctly.

In Fig. 4, the points recognized by the Kinect device are shown, which correspond to the joints of the human body that consider 20 points of the skeleton. The skeletal tracking system of the Kinect provides for each posture a set of twenty joints expressed by coordinates (x, y, z), where these coordinates are evaluated as vectors for every three joints that give the angles that serve for the predictions of the postures through the supervised neural network. We focus only on the knee joint; it also takes some arm movements for the balance of the patient, to describe the specific physiotherapy postures [22] [23] [24].
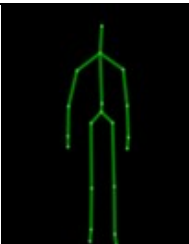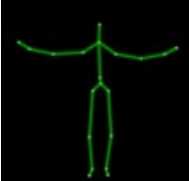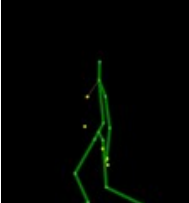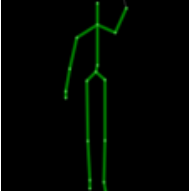
Fig. 4. Capture by Kinect of the Joint of the Second Posture of the Patient.

The coordinates x_i, y_i, z_i ($0 < i < 19$) depend on the size and position of the patient in the scene. The approach used in this work is that with the angles formed between three points through geometry, the angles of the joints are obtained that will give us a better perception of the results, Those are the entries for the network after doing the training is saved in a file in .td format, inside a folder called Training-3, then the trained network is saved in a file for later use.

Table II shows the four postures that are considered for knee therapy. The postures that the child must simulate for his recovery and that the program will tell him if I do the correct posture. The first posture is Standing since the knees must be very straight, posture 2 is Frog here the knee bends an angle of

more than 90 degrees, posture 3 is Lateral Frog here you can see from another view the angles that the knees form if they are uniform and posture 4 in Right knee raised were to raise the bent right knee with the support of the arm lift for the balance of the child. Then the patient will stand in front of the Kinect, and do they pose so that the system displays the name of the pose only if it was done correctly, as shown in Fig. 5.

TABLE II.   BODY POSTURE

Body posture	Name
	Standing
	Frog
	Lateral Frog
	Right knee raised
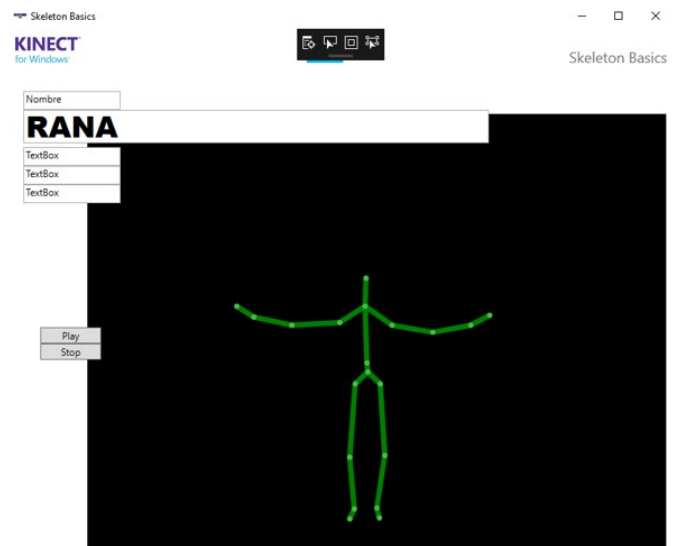
Fig. 5. Recognition of Correct Patient Posture using Kinect.

F. Pre-Processing of Skeleton Data

To improve and optimize the training process is necessary to have the normalized data, we focus on having a representation of the human body independent of the position of the sensor, the distance of the body and the sensor and the structure of the patient's body.

Since the skeleton is represented in three-dimensional space, the first feature extracted is the position. As each joint has three values of three coordinates (x, y, z), and a skeleton consists of 20 joints. However, we only use 14 joints; therefore, our data will have a dimension of 32. This can be seen in Fig. 5, where each join (A, B, C ..., T) that is relevant, is labeled for the capture of the different postures except for the joins L, P, CT, and R.

On the other hand, in the pre-processing, it is necessary that, from the 14 joints, the ten angles are obtained for the characteristic vector. These angles are generated from the union of 2 adjacent vectors, for example, the vectors $(\overrightarrow{HS}, \overrightarrow{Htt})$, $(\overrightarrow{FQ}, \overrightarrow{FW})$ they represent angles that are obtained from the GHS and EFQ joins respectively. Therefore, the feature vector consists of 10 angles, as shown in Eq. 2:

$$\dot{f} = (\theta 1, \theta 2, ... \theta 10) \in R10 \qquad (2)$$

Fig. 6 shows the Joint captured by the Kinect.

G. Model Training

For training, a person executes the postures defined for rehabilitation [9]. This allows 90 samples to be recorded for each pose. In the end, it has been possible to collect (21 poses x 90 samples) 1890 samples distributed among 11 different positions, of which four were used, which are shown in Table II, for more details of how the data of a sample are, before processing it is shown in Fig. 7.

These samples are divided into two random sets: m_{train} for training and m_{test} to calculate the precision of the model.

The method used to classify human postures involves using a dense, layered artificial neural network. The choice of the ANN was made after making a comparison of the precision between different techniques in recognition of postures. Table III shows the comparison of the precision of techniques in recognition of postures.
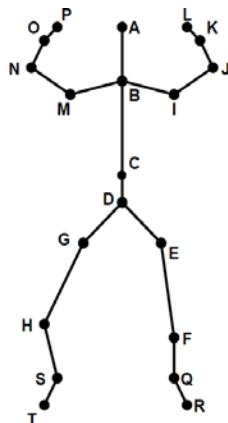
Fig. 6. Joint Captured by the Kinect.
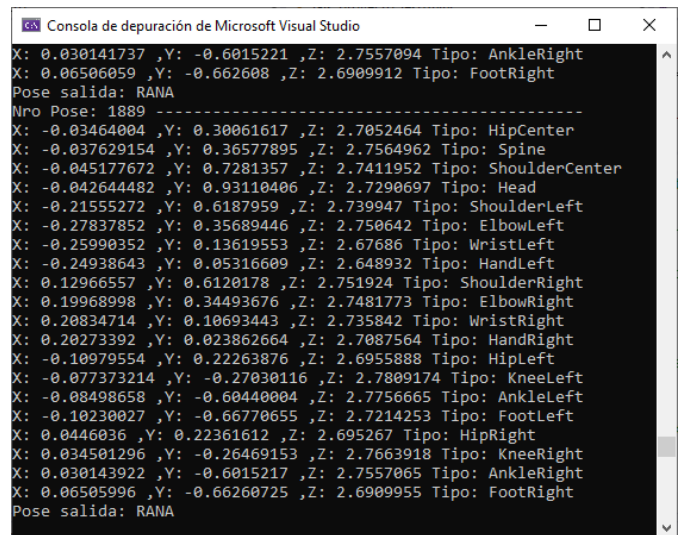
Fig. 7. Sample of the Coordinates of a Frog Posture.

TABLE III. COMPARISON OF THE PRECISION OF TECHNIQUES IN RECOGNITION OF POSTURES

Technique	Accuracy
BLSTM	70.72%
Random Forest (RF)	85.17%
SVM	83.05%
ANN	92.00%

Table III shows that for the BLSTM (bidirectional short-term memory neural network) [25] where the 3D skeleton is detected for each user in the study group, the precision achieved is 70.72%. On the other hand, RF obtained an average general precision of 85.17%, surpassing SVM, which obtained 83.05%, also mentioning that SVM has performance advantages in training time, and RF is more suitable for classifying activities in real-time. However, the use of ANN in the creation of a platform for the recognition of human poses and gestures in order to help patients suffering from stroke disease, an accuracy of 92.00% was obtained [8]. Therefore, we see that ANN's offer excellent precision when dealing with body postures.

Entries for ANN are expressed by feature vectors associated with human poses captured by Kinect. The ANN also has a configuration adapted to the number of input values and the number of output classes. Here are more details:

- the neural network is trained together with the Back-Propagation algorithm.

- the α parameter of the neural network is configured with an initial value of 1.0.

- the network receives inputs consisting of 10 real numbers (angles from 0 to 180 degrees),

- the network has a hidden layer of 20 neurons.

- the network output is made up of 11 neurons that correspond to the total number of positions defined to train the model, and

- the number of iterations for the training, when dealing with angles between 0 to 180 degrees, up to 100,000 iterations had to be used to achieve reasonable precision rates.

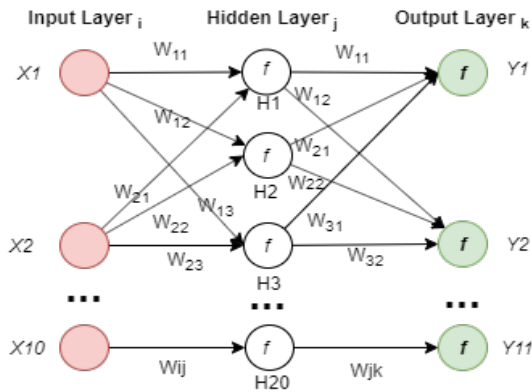Fig. 8 shows the Neural network diagram.

Fig. 8.   Neural Network Diagram.

In the testing stage, some postures made by a patient are recorded, each sample is processed, and ten angles that form a vector of characteristics are extracted, and then this vector is entered into the neural network model and the predicted class and its associated precision.

IV. RESULTS AND DISCUSSION

The tests were available, with hundreds of samples distributed among the four positions considered. The results are displayed on the console, following the form: precision and classified posture. Fig. 8 shows the results that were obtained from the tests carried out with the stored objects.

The first seven samples were extracted from each test posture. The prediction results are shown in Table III. For reasons of space, each posture has been identified by a letter:

- A - Right knee raised

- B – Stopped

- C - Frog (sideways)

- D - Frog (from the front)

Fig. 9 shows the sample of tests performed on saved objects.

Table IV shows that the samples reach a significantly high precision (the highest of 99.70% and the lowest of 98.22%, both averages), but a detail to consider is that the most challenging positions to execute present a slight variation between the training data and are therefore less likely to achieve the best accuracy.

On the other hand, in the neural network, we used a Sigmoid activation function, which gave good results for a 2-layer model; also the initial parameter α of 1.0 was not so decisive, this is derived from the results seen by Choubik and Mahmoudi [9] when using a network with a number of different layers (from 1 to 4), where the parameter α did not make the difference but, using another activation function.
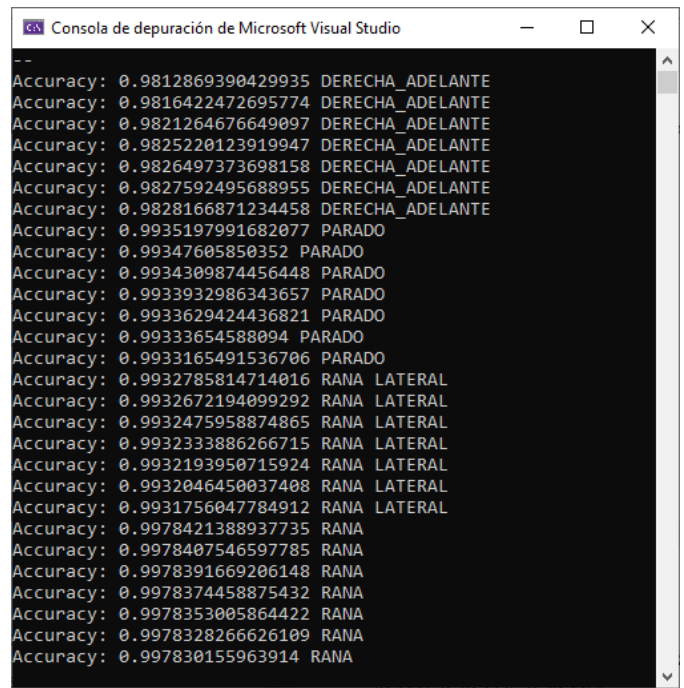
Fig. 9.   Sample of Tests Performed on Saved Objects.

TABLE IV.   RESULTS OF THE RECOGNITION OF 4 POSITIONS FOR 28 SAMPLES (7 FOR EACH POSTURE)

Sample / Posture	A	B	C	D
Sample 1	0.9813	0.9935	0.9933	0.9970
Sample 2	0.9816	0.9935	0.9933	0.9970
Sample 3	0.9821	0.9934	0.9932	0.9970
Sample 4	0.9825	0.9934	0.9932	0.9970
Sample 5	0.9826	0.9934	0.9932	0.9970
Sample 6	0.9828	0.9933	0.9932	0.9970
Sample 7	0.9828	0.9933	0.9932	0.9970
Average accuracy %	0.9822	0.9934	0.9932	0.9970

However, Choubik and Mahmoudi also show other clustering-based learning techniques such as Support Vector Machines (SVM) or k-nearest neighbors (KNN), in the case of SVM, accuracies of up to 100% were obtained using a linear kernel function, while with KNN accuracies were also obtained of 100%, although SVM showed higher precision rates as the amount of data increased more. Le and Nguyen [21] used SVM in the same way, obtaining a precision of 100%, testing with different combinations of angles formed by the joints of the body.

V. CONCLUSIONS

In this work, we proposed one more way to take the data provided by Kinect to convert it into an entry that represents as much information as possible about the body joints, all in order to recognize postures. The entrances are represented by ten angles made up of different groups of three-body joints.

The developed ontology allowed to broadly identify the most outstanding classes or entities of the application, as well as their relationships with other parts of the system.

The neural network trained with the Back-propagation algorithm allowed to give good predictions of the postures based on the angles of the joints, and the more data have, the more significant the recognition precision.

Despite having obtained high classification precision such as 99.70%, there is still room for improvement, considering that Clustering techniques are usually better with data of many dimensions, we may consider extending this work in the future using SVM as learning.

REFERENCES

[1] R. S. Valtanen, A. Arshi, B. V. Kelley, P. D. Fabricant, and K. J. Jones, "Articular Cartilage Repair of the Pediatric and Adolescent Knee with Regard to Minimal Clinically Important Difference: A Systematic Review," Cartilage. 2020.

[2] T. Junge, L. Runge, B. Juul-Kristensen, and N. Wedderkopp, "Risk factors for knee injuries in children 8 to 15 years: The CHAMPS study DK," Med. Sci. Sports Exerc., 2016.

[3] G. Aktas and G. Baltaci, "Does kinesiotaping increase knee muscles strength and functional performance?," Isokinet. Exerc. Sci., 2011.

[4] C. Sinpithakkul, W. Kusakunniran, S. Bovonsunthonchai, and P. Wattananon, "Game-based Enhancement for Rehabilitation Based on Action Recognition Using Kinect," in IEEE Region 10 Annual International Conference, Proceedings/TENCON, 2019.

[5] G. F. Muñoz, J. G. Casero, and R. A. M. Cárdenas, "Usability study of a kinect-based rehabilitation tool for the upper limbs," Adv. Intell. Syst. Comput., vol. 931, pp. 755–763, 2019.

[6] A. Vitali, D. Regazzoni, C. Rizzi, and F. Maffioletti, "A new approach for medical assessment of patient's injured shoulder," in Proceedings of the ASME Design Engineering Technical Conference, 2019.

[7] R. J. Nogales and M. Benalcázar, "Reconocimiento de Gestos de la Mano en Tiempo Real Usando Leap Motion Controller y Machine Learning," in Conference Proceedings, 2019, vol. 3, no. 1, pp. 823–835.

[8] F. Cary, O. Postolache, and P. S. Girão, "Kinect based system and Artificial Neural Networks classifiers for physiotherapy assessment," in IEEE MeMeA 2014 - IEEE International Symposium on Medical Measurements and Applications, Proceedings, 2014.

[9] C. Youness and M. Abdelhak, "Machine Learning for Real Time Poses Classification Using Kinect Skeleton Data," in Proceedings - Computer Graphics, Imaging and Visualization: New Techniques and Trends, CGiV 2016, 2016.

[10] A. Da Gama, P. Fallavollita, V. Teichrieb, and N. Navab, "Motor Rehabilitation Using Kinect: A Systematic Review," Games for Health Journal, vol. 4, no. 2. Mary Ann Liebert Inc., pp. 123–135, 01-Apr-2015.

[11] M. Morando, S. Ponte, E. Ferrara, and S. Dellepiane, "Definition of motion and biophysical indicators for home-based rehabilitation through serious games," Inf., 2018.

[12] S. T. L. Pöhlmann, E. F. Harkness, C. J. Taylor, and S. M. Astley, "Evaluation of Kinect 3D Sensor for Healthcare Imaging," J. Med. Biol. Eng., 2016.

[13] A. D. Calin, H. F. Pop, and R. F. Boian, "Improving movement analysis in physical therapy systems based on kinect interaction," in HCI 2017: Digital Make Believe - Proceedings of the 31st International BCS Human Computer Interaction Conference, HCI 2017, 2017.

[14] D. Leightley, J. Darby, B. Li, J. S. Mcphee, and M. H. Yap, "Human activity recognition for physical rehabilitation," in Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013, 2013.

[15] G. Paraskevopoulos, E. Spyrou, and D. Sgouropoulos, "A real-time approach for gesture recognition using the Kinect sensor," in Proceedings of the 9th Hellenic Conference on Artificial Intelligence, 2016, pp. 1–4.

[16] M. Linna, J. Kannala, and E. Rahtu, "Real-time human pose estimation with convolutional neural networks," in VISIGRAPP 2018 - Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2018.

[17] M. Morando, M. Trombini, and S. Dellepiane, "Application of SVM for evaluation of training performance in exergames for motion rehabilitation," in ACM International Conference Proceeding Series, 2019.

[18] J. Ressman, E. Rasmussen-Barr, and W. J. A. Grooten, "Reliability and validity of a novel Kinect-based software program for measuring a single leg squat," BMC Sports Sci. Med. Rehabil., 2020.

[19] Aforge.net, "AForge.NET :: Framework," AForge.NET Framework, 2020. [Online]. Available: http://www.aforgenet.com/framework/. [Accessed: 25-Jul-2020].

[20] N. B. Pham, V. T. Phan, V. B. Nguyen, and V. D. Nguyen, "Game-based virtual rehabilitation system for upper extremity using low-cost camera," in BMEiCON 2015 - 8th Biomedical Engineering International Conference, 2016.

[21] T. L. Le, M. Q. Nguyen, and T. T. M. Nguyen, "Human posture recognition using human skeleton provided by Kinect," in 2013 International Conference on Computing, Management and Telecommunications, ComManTel 2013, 2013.

[22] Y. Li, "Hand gesture recognition using Kinect," in ICSESS 2012 - Proceedings of 2012 IEEE 3rd International Conference on Software Engineering and Service Science, 2012.

[23] M. Peris and K. Fukui, "Both-hand gesture recognition based on KOMSM with volume subspaces for robot teleoperation," in Proceedings - 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, CYBER 2012, 2012.

[24] J. M. K. Westlund and C. Breazeal, "Transparency, teleoperation, and children's understanding of social robots," in ACM/IEEE International Conference on Human-Robot Interaction, 2016.

[25] R. Saini, P. Kumar, B. Kaur, P. P. Roy, D. P. Dogra, and K. C. Santosh, "Kinect sensor-based interaction monitoring system using the BLSTM neural network in healthcare," Int. J. Mach. Learn. Cybern., 2019.

Detecting Violent Radical Accounts on Twitter

Ahmed I. A. Abd-Elaal[1]

Computer and Systems Engineering
Department Faculty of Engineering
Ain Shams University

Ahmed Z. Badr[2], Hani M. K. Mahdi[3]

Professor of Computer Systems
Faculty of Engineering
Ain Shams University

Abstract—**In the past few years and as a result of the enormous spreading of social media platforms worldwide, many radical groups tried to invade social media cyber space in order to disseminate their ideologies and destructive plans. This brutal invasion to society daily life style must be resisted as social media networks are interacted with on daily basis. As some violent radical groups such as ISIS has developed well designed propaganda strategies that enables them to recruit more members and supporters all over the world using social media facilities. So it is crucial to find an efficient way to detect the violent-radical accounts in social media networks. In this paper, an intelligent system that autonomously detects ISIS online community in Twitter social media platform is proposed. The proposed system analyzes both linguistic features and behavioral features such as hashtags, mentions and who they follow. The system consists of two main sub-systems, namely the crawling and the inquiring subsystems. The crawling subsystem uses the initially known ISIS-related accounts to establish an ISIS-account detector. The inquiring subsystem aims to detect Pro ISIS-accounts.**

Keywords—*Machine learning; ISIS; Daesh; extremism; data mining; social media; Twitter*

I. INTRODUCTION

Social media become an essential part of everyone's life style nowadays. Everyone now can easily express his thoughts, feelings and even his emotions through internet and these ideas will spread within seconds among the world. These posts may be viewed by millions who will interact with them. Online Social Network (OSN) has grown significantly over the past decade. Within a population of 7.676 billion human being which is the entire population on earth, there are 4.388 billion Internet users, 3.256 billion smart phones and 3.484 billion active social media users in 2019 [1]. Its existence gave humanity the gift of spreading civilization, literature, science, arts, and others which are means of fulfilling prosperity and welfare all over the world. On the other hand, if this power was misused, many unpleasant sequences may occur. Hatred may be spread instead of constructive ideas, violence instead of literature and science, war instead of welfare.

Many extremist radical hate groups and violent associations are consistently trying to spread their ideologies and hate speech through various social media platforms. In other words, these groups are using the social media facilities to recruit new members and to distribute destructive ideas and plans. It is very important to identify the members of these groups to prevent them from spreading their harmful ideologies, disseminating violence and hatred on social media platforms that may result war and conflicts in peaceful societies. A living example on

these radical groups is Islamic State of Iraq and al-Sham "ISIS", also is known as "Daesh" in Middle East. It is recognized by its adherence to the fundamentalist Salafi faith of Sunni Islam [2]. It gained international fame in early 2014 when it expelled Iraqi government forces from the main cities of Western Iraq then seizing Mosul and committing the massacre of Sinjar. Since its appearance, "ISIS" is continuously trying to leverage its ideology through social media platforms.

Twitter social media is one of the most popular online social media networks in the world with 330 million monthly active users and 500 million tweets per day. A tweet is a message composed of 140 characters that any user can easily share among millions of accounts. Tweets may contain hashtags that highlight the tweet's main topic. Although there is no accurate statistics shows the existence of ISIS members but in 2014, it was an expected that from 46,000 to 90,000 Twitter accounts that upheld for ISIS or they were controlled by its supporters. Over 1.2 million accounts suspended for terrorist content since August 2015 [3]. Suspension criteria are mainly based on other accounts reporting an account that generates violent radical tweets. However, inspecting the posts of a single user in a social media application such as Twitter is a tedious work. The need for an artificial intelligent approach that mimics the human inspection to solve this problem is a must.

In this paper, a new architecture that can autonomously detect ISIS members' community in the Twitter social media network is proposed. The proposed system analyzes both linguistic features and behavioral features such as hashtags, mentions and the following lists. These will enable us to continuously have a live image of how ISIS members interact with online social media. The rest of the paper is organized as follows: Section 2 reviews the previous studies and related work. Section 3 clarifies dataset gathering and analysis. Section 4 introduces the proposed architecture. Section 5 discusses the experimental results. Finally, Section 6 includes the conclusions and future work.

II. RELATED WORK

ISIS existence in twitter has been expanded enormously that catch the concerns of the international society. That resulted the appearing of some online groups volunteering their time and resources to report those violent accounts. One of these groups was Ctrl-sec [4] that was responsible of deactivation of about 25,000 twitter accounts in three years by identifying radical accounts manually [5]. Others could suspend about 25,000 Pro-ISIS accounts through crowd-

sourced reporting [3]. Many researchers were attracted to this topic. They are trying to find new means to discover and limit the vast spread of these accounts. Ashcroft et al. [6] adopted machine learning techniques based on a list of English predefined hashtags related to ISIS to detect ISIS related tweets using three feature classes, namely sentiment, temporal and stylometric features. He found that his approach is highly dependent on data and the approach of detecting radical content should always be a helpful tool to support humans to asset accounts not to replace them. Choudhary et al. [7] tried to detect behavior patterns and key player features to identify terrorist community.

In a trial to understand what happens through the phase of an account being converted from an ordinary account to a Pro-ISIS one sharing Pro-ISIS content, Rowe and Saif [8] defining the Pro-ISIS account by the account that shares more radical Pro-ISIS content than the Anti-ISIS one. Although this approach seems effective but couldn't deal with the lexical diversity.

Klausen et al. [9] studied the communication flow among ISIS members on twitter using 59 manually assessed Pro-ISIS accounts and found that female members has a key role in ISIS propaganda. Carter et al. [10] found that newly bounded ISIS members seek guidance through online spiritual ISIS figures on twitter. Chatfield et al. [11] investigated how ISIS recruits new members through disseminating their ideology on social media platforms. Also Vergani et al. [12] studied how ISIS uses emotional speech and religious quotes to recruit online supporters. Berger et al. [13] found that twitter users who follow ISIS members are highly affected by their ideology. Saif et al. [14] found that semantic features based models out performs other lexical, topic and sentiment based models in detecting Pro-ISIS accounts. Berger et al. [15] found that Pro-ISIS accounts can be identified through their profile description. Agarwal et al. [16] expressed the presence of offensive, war and hate speech terms in ISIS propaganda.

III. DATA COLLECTION AND ANALYSIS

A. Data Gathering

As the research objective is to automatically detect Arabic speaking violent radical accounts on twitter, dataset should be found and properly cleaned to be analyzed, studied and to apply various machine learning algorithms on it. Although native ISIS members are Arabic speakers as they mainly located in Syria and Iraq, no Arabic ISIS related dataset was found due to the lack of proper Arabic resources in science society especially in machine learning field. In order to prepare a suitable Arabic dataset that can be used in this research, two approaches for collecting data were adopted. The first approach is to collect proper dataset from Arabic speaking twitter accounts that represent Pro-ISIS, Anti-ISIS and non-ISIS. This dataset is referred to hereafter as the collected dataset. The second approach is to collect the published available non-Arabic ISIS related datasets that can be found in online data-science communities and translate them. This dataset is referred to hereafter as the translated dataset.

1) The collected dataset

By studying and examining extremist accounts in Twitter, the most frequent hashtags in ISIS propaganda were manually identified such as تتمدد ,باقية ,الإسلامية_الدولة ,وأعدوا. Using these hashtags, 42 accounts were collected and annotated as the most violent and ISIS influential accounts. Using twitter API [17] the tweets feed of the annotated accounts were collected which resulted to downloading of 21,000 tweets. These accounts will be referenced as "collected Pro-ISIS accounts". Similarly and in order to collect balanced dataset 21,000 Anti-ISIS tweets were gathered in the same way using the following hashtags جرائم داعش, داعش تجار الدم, مسلمون ضد داعش, فضائح داعش, بلغ such as عن داعش as they were the most used hashtags in the most active Anti-ISIS accounts namely "collected Anti-ISIS accounts". 21,000 random non-ISIS related tweets were collected as-well from different domains: "News – Religion – Sports – Art" to represent "collected non-ISIS related accounts". Data pre-processing such as URL and mentions removal, discarding non alpha characters such as (@, #, $, %, _), characters normalization such as (ئ ,ؤ ,آ, إ ,أ,) and stop words removal such as (في, على, أي, من, إلى) , tashkeel removal such as (َ◌, ُ◌, ِ◌, ْ◌, ّ◌) and prefix/suffix removal such as (الـ, نى, نا, و) was applied to prepare the collected dataset for training stage.

2) The translated dataset

By searching online data-science communities such as "Kaggle", three non-Arabic ISIS related datasets were found:

a) Fifth-tribe "How ISIS Uses Twitter" dataset [18]. It is consisted of 17,000 tweets was collected from 112 Pro-ISIS twitter accounts from all over the world that supported 2015 terrorist Paris attacks [19]. These tweets are mostly written in English.

b) "Religious Texts Used By ISIS" [20]. 2,685 religious texts dataset which was collected by scrapping 24 issues of Dabiq and Rumiyah English-based ISIS magazines that ISIS uses to spread their ideology in Europe and western world.

c) "Tweets Targeting ISIS" dataset which contained 122,000 ISIS related tweets was collected all over the world in many languages, mostly in English [18]. These tweets were collected by following ISIS related terms such as (ISIS, Daesh, Islamic State, Raqqa, Mosul) 13,000 tweets that were against ISIS ideology and terrorism were translated into Arabic.

Translating into Arabic language was carried out by custom python scripts using Google translating service [21]. Translated dataset was manually reviewed to correct mistranslated words/expressions. Finally, data pre-processing steps were carried out including data cleaning, normalization, stop words removal and stemming as mentioned in the collected dataset subsection.

B. Text Features Vectoring

Dataset was collected from 2-main different sources collected dataset and translated one. These two datasets may suffer from the domain divergence problem [22]. To make sure of suitability of applying any recognition and detection technique for both of these datasets, visual testing was adopted. In the first step to carry out this test, the datasets are first represented as vectors. Then visualization techniques are used to represent the converted words. "Mazajak" word embedding

was used [23] to convert dataset corpus into vector domain. Mazajak is considered to be the largest Arabic word embedding models based on a corpus of 250 million tweets converted using skip-gram architecture [24]. In Mazajak, each word token is converted to a 300-D vector. Tweets can be represented by the mean of its contained word embedding's. Similarly the whole user's tweets thread can be represented by the mean of its vector tweets embedding.

The second step in the proposed visual testing is dimensionality reduction of the resulting vectors because it is not proper to visualize 300-D vectors. So embedding vectors dimensions have to be reduced in order to visualize and study the dataset. TSNE [25] is a machine learning technique for dimensionality reduction. This technique is applied here to reduce the vector dimensions from 300-D to just 2-D.

C. Text Features Analytics

After converting the dataset's corpus into a vector form using "Mazajak" word embedding model [23] where the distance between any two vectors is proportional to the difference in the meaning of the words they represent. The vectors should be plotted graphically so they can be studied and analyzed to get better understanding for dataset. That is how the consistency of the dataset can be checked as it was collected from multiple sources and to make sure that vectors that represent the same class can be clustered in spite of the lexical diversity between the collected and translated data.

Fig. 1 illustrates both the collected and the translated tweets. The "collected Pro-ISIS accounts" tweets and both the translated "How ISIS Uses Twitter" tweets and the translated "Religious Texts Used By ISIS." These tweets were labeled as Pro-ISIS in "Red", Anti-ISIS in "Green" and Random in "Blue" colors. From Fig. 1, it can be noticed that some tweets from both Anti-ISIS and Pro-ISIS classes are overlapped with random tweets. This can be easily interpreted because ISIS related tweeters may have other interests in different topics such as sports and news. Also it is noticed that non-ISIS related tweets are distributed over a wide range of fields that belong to different domains such as "news – religion – sports – art".
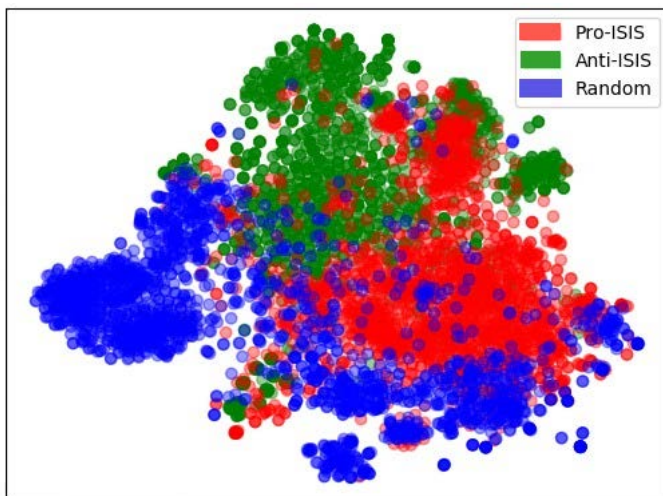
Fig 1.   Datasets in 2-D Representation.
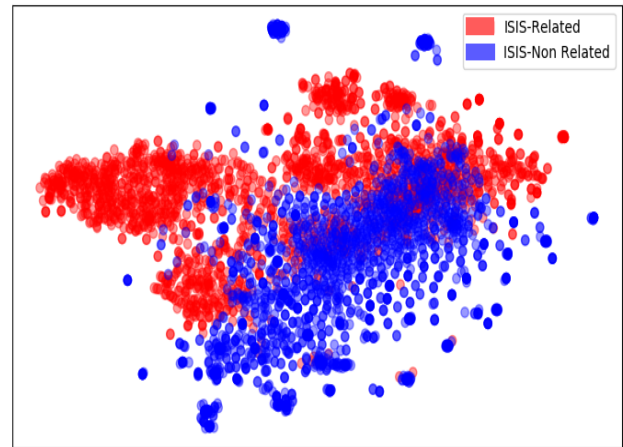
Fig 2.   ISIS and Non-ISIS in 2-D Representation.

Non-ISIS related dataset should be collected in order to avoid detection accuracy degradation if the system is not trained to detect non-ISIS related tweets. Fig. 2 represents in "Red" ISIS related tweets from both classes Anti-ISIS and Pro-ISIS tweets, where non-ISIS related tweets are represented in "Blue". It is obvious that ISIS related class is clearly separated from non-ISIS related.

Finally the relation between Pro-ISIS and Anti-ISIS tweets has to be expressed. They may share the same vocabularies, as twitters from both sides (Pro and Anti ISIS) regularly discuss similar subjects in their tweets what will make it harder for us to separate between them. To gather appropriate tweets that represent these two classes, the translated 13,000 tweets from "Tweets Targeting ISIS" dataset and 18,000 manually labeled tweets from collected Anti-ISIS tweets were used to represent the Anti-ISIS data class. On the other side, the translated 12,000 tweets from "How ISIS Uses Twitter" plus 16,200 tweets from collected Pro-ISIS and 2,600 translated texts from "Religious Texts Used By ISIS" were used to represent Pro-ISIS data class. Fig. 3 shows that the two classes can't be separated linearly.

D. Behavioral Features Organization

In addition to the usage of the lexical features of the dataset, it is important to make use of other behavioral features got from the collected data such as mentions, hashtags, likes, retweets and follow lists. These features express the behavior of the user, as the user's used hashtags, retweets and likes defines the topics he is interested in. On the other hand, his mentions and follow list express his spiritual leaders. With the aid of unstructured Mongo database management system [26], the collected mentions, likes, retweets, hashtags and follow lists can be aggregated. Each of these features will have a score that will be identified with its commitment in ISIS community. If a certain hashtag or mention is found more often in ISIS propaganda, it will have higher score which will reflect how much it is related to ISIS. This facilitates capturing the topics they are keen on, how they influence their beliefs and who they follow. Finally a list of the most followed accounts, the most liked and retweeted tweets in addition to the most tweeted hashtags by ISIS supporters will be obtained.
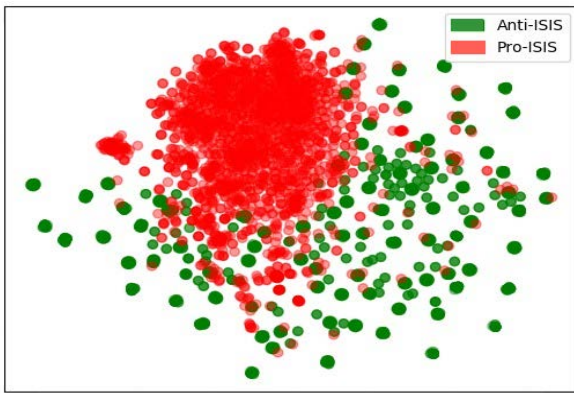
Fig 3.   Anti-ISIS, Pro-ISIS in 2-D Representation.

IV. PROPOSED ARCHITECTURE

Although Pro-ISIS propaganda is distinguished by the presence of offensive, war and hate speech terms [16], the lexicon itself varies according to the undergoing events [8]. They always use the world trending topics to ensure the wide spread of their propaganda. Other problem is that ISIS propaganda evolves during the process of recruitment itself within many stages that includes religious, emotional and hate speech [7]. These problems cause degradation in the detection accuracy on the long term for non-maintainable detection systems, as they rely on old detection models or outdated lexical dictionaries in the detection process.

The main challenge of the proposed system is to maintain it up to date autonomously. This includes tracking the changes in the online-behavior of ISIS members as well as tracking the evolution of their ideologies and propaganda plans. Moreover, it is essential to identify the fundamental key-players or profound pioneers that originate hate speech content. In a trial to overcome these challenges, the proposed architecture is designed to continuously crawl on ISIS online community updating its corpora and other semantic features such as used hashtags, mentions, who they follow and what they share. That will give us a live image of how ISIS behaves on Twitter social media platform. In order to invade ISIS online community on Twitter, an initial seed of ISIS related accounts will be needed which can be found in the collected dataset. The proposed architecture consists of two main subsystems, namely the crawling subsystem and the inquiring subsystem. The details of these subsystems are explained in the next two subsections. They are depicted separately in Fig. 4 and Fig. 5.

A. The Crawling Subsystem

The Crawling subsystem enables us to invade ISIS online Twitter community. Fig. 4 depicts the subsystem components. The crawling subsystem is started by targeting the most followed accounts from the follow lists in the collected dataset. Using Twitter's REST API [17], the account info, followers list and the last posted 3,000 tweets for each targeted account can be downloaded. The downloaded tweets will undergo data pre-processing steps. The steps include data cleaning, normalization, stop words removal and stemming prior inputting to ISIS-Content detector. ISIS-Content detector will detect ISIS related tweets in order to define for how far the user is involved into ISIS community.
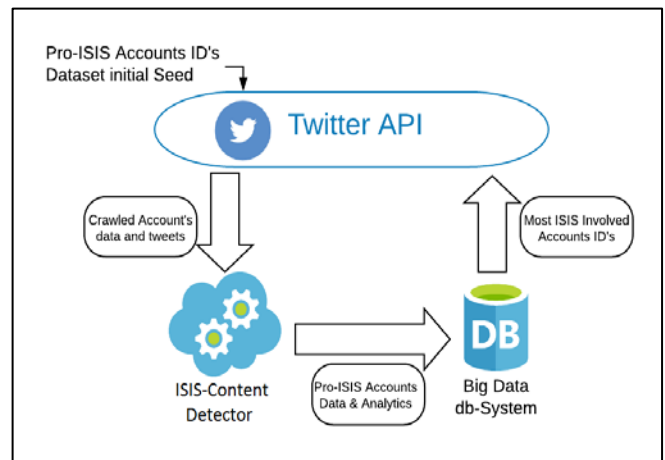
Fig 4.   The Crawling Subsystem.

As the ISIS-Content detector will assess all of the downloaded tweets for each user, it will distinguish whether each tweet is Pro, Anti or non-related to ISIS in addition to updating hashtags and mention lists in the system's database if the tweet is predicted as Pro-ISIS. If the anticipated Pro-ISIS tweets ratio for the tested account surpassed certain ratio (Pro-ISIS threshold) the account is considered to be Pro-ISIS. Finally all the collected and analyzed data for each account will be stored in the unstructured MongoDB [26]. The collected database can be effortlessly used to investigate and aggregate most followed/active accounts, hashtags and mentions. Updated most followed ISIS accounts will be sent back to be cushioned as a contribution to the following cycle to keep updating the database. As a result of this subsystem, ISIS members, supporters and leaders can be easily tracked in addition to addressing hot topics undergoing in ISIS. Thus the gain of this subsystem is the expansion of the user's knowledge. Likewise an up to date dataset of ISIS-related Arabic labeled tweets will be continuously available that can be used in further studies.

B. The Inquiring Subsystem

Inquiring subsystem enables the user to inspect specific twitter account by twitter ID as an input. Using Twitter API [17], the system will download account data along with the latest posted 3,000 tweets.
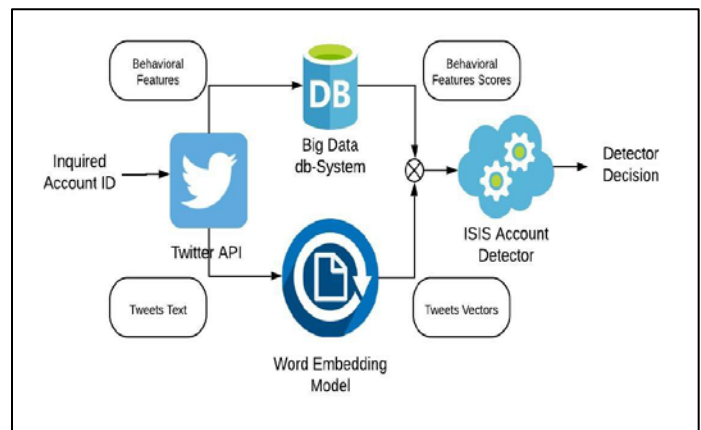

Fig 5.   The Inquiring Subsystem.

The downloaded tweets will undergo the same cleaning and data pre-processing steps before usage as in the crawling subsystem. Downloaded hashtags, mentions and following list will be calculated and correlated with the data stored in the system DB which was collected from crawling subsystem to calculate its behavioral features. Behavioral scores will be determined by contrasting the calculated behavioral features by the blacklisted data from crawling subsystem. Behavioral scores along with the pre-processed tweets will be the input to ISIS Accounts Detector which will detect whether the account is Pro-ISIS or not.

C. System Maintenance

In order to maintain the system up to date with high-performance and acceptable detection rates, it must be periodically updated and retrained. It should be introduced to undergoing ISIS related events and its members reactions on them, track their spiritual leaders and key players, learn about their ideologies and propaganda methods. As a result of continues pursue of ISIS community, a live image of how ISIS members spread their news, ideology and even guidance on social media will be obtained. Helpful reports of tracking ISIS members and leaders can be easily developed. Also a proper up to date dataset will be developed which can be used to periodically retrain our detectors on up to date data to avoid deprecation that causes degradation in detection accuracy.

V. EXPERIMENTAL RESULTS

The proposed system includes two pre-trained supervised classification detectors which are key nodes in the system. They detect whether the tweet is Pro-ISIS or not and whether the account is Pro-ISIS or not. Their accuracies define the overall performance of the system as they categorize and define the quality of the crawled data which is important in system evolution. So their accuracy should be boosted and lower down the possible False Positive or False Negative rates.

Experiments were carried out using six different machine learning algorithms namely, Bernoulli Naive Bayes, Decision Tree Classifier, K Neighbors Classifier, Linear Support Vector Classifier, Logistic Regression and Random Forest Classifier [27]. Pre-processed tweets were converted to vectors by multiple word embedding algorithms: "Mazajak" word embedding [23], skip-gram scheme [24] and TF-IDF embedding [28] in [unigram "UG" – bigram "BG"- teragram "TG"] forms. The dataset was then shuffled and spitted into ratios 80% as training dataset and 20% as testing dataset.

A. ISIS-Content Detector

ISIS-Content Detector should process on all of the downloaded tweets downloaded from crawling subsystem that targets ISIS community on Twitter. Although tracking ISIS members in the proposed system, also their tweets should be checked before labeling them where some of ISIS members may have other interests such as news, sports or religion so Pro-ISIS tweets obtained from stalked ISIS members should be inspected in order to increase the quality of the collected dataset.

Other task is to determine how far the stalked account owner is involved into ISIS community, as the ratio of his Pro-

ISIS tweets to all of its tweets is calculated. If it exceeds a certain value, the account will be labeled as Pro-ISIS account in the collected database. So a supervised pre-trained detector on labeled data classes that represent Pro-ISIS, Anti-ISIS and non-ISIS related tweets should be prepared. In order to collect these three classes, the translated 13,000 tweets from "Tweets Targeting ISIS" dataset and 18,000 manually labeled tweets from collected Anti-ISIS tweets were used for representing the Anti-ISIS class. For Pro-ISIS data class 12,000 translated tweets from "How ISIS Uses Twitter" plus 16,200 tweets from collected Pro-ISIS and 2,600 translated texts from "Religious Texts Used By ISIS" were used. Finally, the collected 21,000 random tweets were used to represent the non-ISIS related class. Table I shows the results of the testing process.

TABLE I. ISIS-CONTENT DETECTOR'S TESTING RESULTS

Algorithm	TF-IDF			Skip-gram "Mazajak"
	UG	BG	TG	
Bernoulli NB	0.88	0.87	0.86	**0.86**
Decision Tree Classifier	0.80	0.80	0.80	0.79
K Neighbors Classifier	0.48	0.51	0.53	0.77
Linear SVC	0.87	0.87	0.89	**0.84**
Logistic Regression	0.85	0.87	0.88	**0.84**
Random Forest Classifier	0.18	0.62	0.64	0.80

Table I shows that TF-IDF embedding outperforms skip-gram embedding in all experiments. Linear SVC achieved accuracy 89% followed by Bernulli NB in "UG" TF-IDF [27] and Logistic Regression [27] in "TG" TF-IDF achieved accuracy 88% regarding F1-score [27]. Detector's accuracy can be boosted by using all the three detectors as a voting detector as an ensemble learning system [29]. So each tweet will be labeled as the majority of the three detectors decide which will avoid any weaknesses found in any individual one of them in addition to avoiding over fitting.

B. ISIS-Account Detector

ISIS-Account Detector is a key node in the proposed system. It should check if the account inquired by the system user is Pro-ISIS or not. Unlike ISIS-Content detector that operates on discrete tweets, Pro-ISIS accounts detector run on all account data including tweets as textual features, in addition to hashtags, mentions, likes, retweets and following list as behavioral features. Thus beside vectoring dataset corpus by word embedding techniques, other behavioral features such as used hashtags, mentions and following list were vectored by comparing them by the collected data from the crawling subsystem and assigning scores to them.
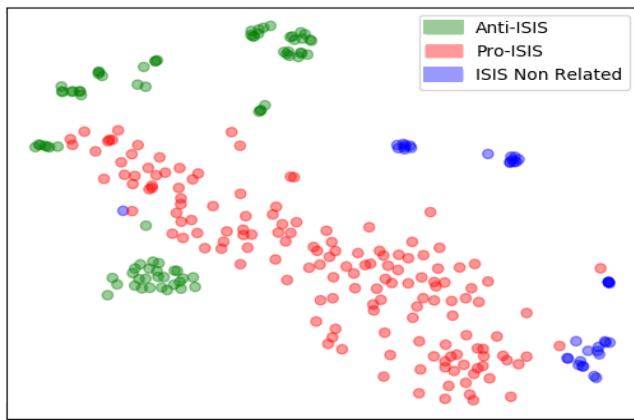
Fig 6.   ISIS, Non-ISIS in 2-D Representation.

TABLE II.   Pro-ISIS Account Detector's Testing Results

Algorithm	TF-IDF			Skip-gram "Mazajak"
	UG	BG	TG	
Bernoulli NB	0.71	0.58	0.55	0.86
Decision Tree Classifier	0.81	0.81	0.83	**0.89**
K Neighbors Classifier	0.68	0.67	0.70	0.87
Linear SVC	0.62	0.62	0.62	**0.94**
Logistic Regression	0.56	0.59	0.62	**0.94**
Random Forest Classifier	0.57	0.45	0.48	0.80

In order to train this detector, two data classes that properly represent Pro-ISIS accounts and non-Pro-ISIS were created. The translated "How ISIS Uses Twitter" dataset in addition to the collected Pro-ISIS accounts were used to represent Pro-ISIS data class. On the other side the collected Anti-ISIS accounts plus collected non-ISIS related accounts were used to represent non-ISIS data class. Non-ISIS related accounts were introduced to the detector to avoid detection accuracy degradation if a non-ISIS related account was inquired in the system. Fig. 6 expresses the scattering of the accounts in the dataset. However Pro-ISIS and Anti-ISIS accounts may share the same lexicons as they are both involved into the same topics, Fig. 6 shows that Pro-ISIS accounts can easily be separated from both Anti-ISIS and non-ISIS-related accounts.

Table II shows that this time skip-gram embedding [24] outperforms TF-IDF embedding. Both Linear SVC and Logistic Regression [27] achieved F1-score 94% followed by Decision Tree [27] by score 89%. Such as "ISIS-Content Detector", accuracy can be boosted and avoid over fitting by using all three in a voting ensemble learning system [29].

VI.   Conclusions and Future Work

A new intelligent system architecture that autonomously detects Pro ISIS-accounts in twitter social media platform is introduced. The system consists of two main sub-systems, namely the crawling and the inquiring subsystems. The kernels

of the two subsystems are intelligent detectors. The attributes of these detectors are both linguistic features and behavioral features.

For proper testing the proposed system, a new collected dataset of Pro-ISIS, Anti-ISIS and non-ISIS-related accounts were gathered. Three English datasets about ISIS were translated to Arabic. All datasets were represented as vectors using "Mazajak" word embedding "skip-gram" scheme to 300-D vectors. Vectors dimensions were reduced into 2-D and plotted to check their consistency.

The intelligent detectors kernels for the crawling and the inquiring subsystems were developed using supervised machine learning techniques. The results show that ISIS-Content Detector gave best accuracy 89% according to f1-score by linear SVM algorithm using TF-IDF embedding. They show also that ISIS-Account Detector gave pest accuracy 94% according to f1-score by linear SVM algorithm Skip-gram embedding.

As next steps, extension of the proposed architecture to other social media applications such as Facebook and Instagram is planned. Enlarge the used dataset to include different Arabic delegates.

References

[1]   S. Akar, Ezgi and Mardikyan, "Analyzing factors affecting users' behavior intention to use social media: Twitter case," Int. J. Bus. Soc. Sci., vol. 5, no. 11, 2014.

[2]   A. Chaliand, Gerard and Blin, The history of terrorism: From antiquity to ISIS. Univ of California Press, 2016.

[3]   K. M. Benigni, Matthew C and Joseph, Kenneth and Carley, "Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter," PLoS One, vol. 12, no. 12, p. e0181405, 2017.

[4]   H. Fernandez, Miriam and Alani, Contextual semantics for radicalisation detection on Twitter. CEUR, 2018.

[5]   E. Ferrara, "Contagion dynamics of extremist propaganda in social networks," Inf. Sci. (Ny)., vol. 418, pp. 1–12, 2017.

[6]   N. Ashcroft, Michael and Fisher, Ali and Kaati, Lisa and Omer, Enghin and Prucha, "Detecting jihadist messages on twitter," in 2015 European Intelligence and Security Informatics Conference, 2015, pp. 161–164.

[7]   U. Choudhary, Pankaj and Singh, "A survey on social network analysis for counter-terrorism," Int. J. Comput. Appl., vol. 112, no. 9, pp. 24–29, 2015.

[8]   H. Rowe, Matthew and Saif, "Mining pro-ISIS radicalisation signals from social media users," in tenth international AAAI conference on web and social media, 2016.

[9]   J. Klausen, "Tweeting the Jihad: Social media networks of Western foreign fighters in Syria and Iraq," Stud. Confl. Terror., vol. 38, no. 1, pp. 1–22, 2015.

[10] P. R. Carter, Joseph A and Maher, Shiraz and Neumann, Greenbirds: Measuring importance and influence in Syrian foreign fighter networks. Citeseer, 2014.

[11] U. Chatfield, Akemi Takeoka and Reddick, Christopher G and Brajawidagda, "Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks," in Proceedings of the 16th Annual International Conference on Digital Government Research, 2015, pp. 239–249.

[12] A.-M. Vergani, Matteo and Bliuc, "The evolution of the ISIS'language: a quantitative analysis of the language of the first year of Dabiq magazine," Sicurezza, Terror. e Soc., vol. 2, pp. 7–20, 2015.

[13] B. Berger, John M and Strathearn, Who Matters Online: Measuring Influence, Evaluating Content and Countering Violent Exremism in Online Social Networks. International Centre for the Study of Radicalisation and Political Violence, 2013.

[14] H. Saif, Hassan and Dickinson, Thomas and Kastler, Leon and Fernandez, Miriam and Alani, "A semantic graph-based approach for radicalisation detection on social media," in European semantic web conference, 2017, pp. 571–587.

[15] J. Berger, Jonathon M and Morgan, "The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter," Brookings Proj. US relations with Islam. world, vol. 3, no. 20, pp. 1–4, 2015.

[16] A. Agarwal, Swati and Sureka, "Using knn and svm based one-class classifier for detecting online radicalization on twitter," in International Conference on Distributed Computing and Internet Technology, 2015, pp. 431–442.

[17] H. Kumar, Shamanth and Morstatter, Fred and Liu, Twitter data analytics. Springer, 2014.

[18] Fifthtribe, "How ISIS Uses Twitter," kaggle, 2019. [Online]. Available: https://www.kaggle.com/ activegalaxy/isis-related-tweets. [Accessed: 20-June-2020].

[19] R. K. Cragin, "The November 2015 Paris attacks: the impact of foreign fighter returnees," Orbis, vol. 61, no. 2, pp. 217–226, 2017.

[20] S. Fuhriman, Christopher and Medina, Richard M and Brewer, "Introducing a Dataset of Multi-Scale Geographies of ISIS Ideology from ISIS Sources," Terror. Polit. Violence, pp. 1–18, 2020.

[21] H. Somers, "Example-based machine translation," Mach. Transl., vol. 14, no. 2, pp. 113–157, 1999.

[22] W. Wang, Mei and Deng, "Deep visual domain adaptation: A survey," Neurocomputing, vol. 312, no. 135–153, 2018.

[23] W. Farha, Ibrahim Abu and Magdy, "Mazajak: An online Arabic sentiment analyser," in Proceedings of the Fourth Arabic Natural Language Processing Workshop, 2019, pp. 192–198.

[24] A. Neelakantan, Arvind and Shankar, Jeevan and Passos, Alexandre and McCallum, "Efficient non-parametric estimation of multiple embeddings per word in vector space," arXiv Prepr. arXiv1504.06654, 2015.

[25] G. Maaten, Laurens van der and Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, no. Nov, pp. 2579–2605, 2008.

[26] K. Chodorow, MongoDB: the definitive guide: powerful and scalable data storage. "O'Reilly Media, Inc.," 2013.

[27] K. Khan, Aurangzeb and Baharudin, Baharum and Lee, Lam Hong and Khan, "A review of machine learning algorithms for text-documents classification," J. Adv. Inf. Technol., vol. 1, no. 1, pp. 4–20, 2010.

[28] R. Qaiser, Shahzad and Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," Int. J. Comput. Appl., vol. 181, no. 1, pp. 25–29, 2018.

[29] T. G. Dietterich, "Ensemble methods in machine learning," in International workshop on multiple classifier systems, 2000, pp. 1–15.

Intelligent Tutoring Supported Collaborative Learning (ITSCL): A Hybrid Framework

Ijaz Ul Haq[1], Aamir Anwar[2], Iqra Basharat[3], Kashif Sultan[4,*]

Department of Software Engineering, Bahria University Islamabad Campus, Pakistan[1, 3, 4]

Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science & Technology

Islamabad Campus, Pakistan[2]

Abstract—Recently Intelligent Tutoring Systems (ITS) and Computer-Supported Collaborative Learning (CSCL) have got much attention in the field of computer science, artificial intelligence, cognitive psychology, and educational technologies. An ITS is a technologically intelligent system that provides an adaptive learning paradigm for an individual learner only, while CSCL is also a technology-driven learning paradigm that supports groups of learners in pertaining knowledge by collaboration. In a multidisciplinary research field—the Learning Sciences, both individual and collaborative learning have their own significance. This research aims to extend ITS for collaborative constructivist view of learning using CSCL. Integrating both design architecture of CSCL and ITS, this research model propose a new conceptual framework underpinning "Intelligent Tutoring Supported Collaborative Learning (ITSCL)". ITSCL extend ITS by supporting multiple learners interacting system. ITSCL support three different types of interaction levels. The first level of interaction supports individual learning by learner-tutor interaction. The second and third level of interaction support collaborative learning, by learner-learner interaction and tutor-group of collaborative learners' interactions, respectively. To evaluate ITSCL, a prototype model was implemented to conduct few experiments. The statistical results extrapolate the learning gains, measured from Paired T-Test and frequency analysis, contend a significant learning gain and improvement in the learning process with enhanced learning performance.

Keywords—*Intelligent Tutoring System (ITS); Computer-Supported Collaborative Learning (CSCL); Artificial Intelligence (AI); individual learning; collaborative learning*

I. INTRODUCTION

Intelligent Tutoring System (ITS) uses Artificial Intelligence (AI) techniques to provide intelligent tutors in some domains without human intervention. The intelligent tutoring system discovers the diverse status of student experiences and gives adaptive feedback to enhance the learning process [1]. The research community in Artificial Intelligence Education (AIED) has been mostly focusing on the development of one-on-one ITS. While the field of learning sciences is now focused on the integration of individual and collaborative learning. In collaborative learning, peer learners collaborate to solve a problem, encourages peers to explore ideas, present and defend arguments, exchange of ideas, conceptual mapping, reflect and elaborate upon their knowledge. Traditional non-collaborative ITS, geared towards a single learner, despite the different positive impacts of individual and collaborative learning. The one-on-one intelligent tutoring system is not in accordance with a collaborative constructivist view of learning. In single learner (learner-tutor) ITS, there are boundaries among the learners with no real-time interaction or knowledge sharing on problem-solving. The adaptive and flexible intelligent tutoring model allows us to extend this to collaborative intelligent tutoring. To extend the scope and application of ITS for collaborative learning, Computer-Supported collaborative learning (CSCL) holds the potential in a broader range for collaborative learning.

Computer-Supported Collaborative Learning (CSCL) uses computer technology for promoting students' collaboration [2]. CSCL aims to support groups of students in getting knowledge by collaborating in a specific domain using a computer as mediation [3]. Collaboration in a coordinated system is a more challenging task [4]. An important study concluded that developing an intelligent system for CSCL is more challenging to improve collaborative learning and development of collaboration skills [5].

ITS and CSCL both are the multidisciplinary areas of cognitive psychiatry, computer science, and educational technologies, etc. CSCL and ITS provide pedagogically, cognitive and scaffolding of learning. In computer-based learning environments, ITS is assisting students in acquiring knowledge [6]. ITS and CSCL are computer-mediated platforms that monitor interaction patterns and provide feedback to learners. Communication in CSCL is interactive, dynamic, varied and unpredictable while ITS intelligently communicates with the learner. The important feature in CSCL learning is the collaborative peer learner having a shared understanding within the learning environment, on the other side in ITS environment learning is only from an intelligent system that supports in improving personal learning skills. CSCL researchers are focusing on exploring issues of adaptivity, interaction analysis, and feedback. These research aspects of CSCL are getting closer to the techniques of ITS research. Leveraging CSCL approaches could be promising towards combining individual and collaborative learning within ITS. CSCL in ITS integration means, multiple learners interact with the system and with each other for collaborative learning. This endeavor holds the potential for extending the scope and application of ITS in a broader range.

This research shifts the paradigm from traditional ITS that is limited to the single learner to ITSCL, as ITSCL encapsulates both individual and collaborative learning. To conceptualize ITSCL, the architecture of ITS with individual

*Corresponding Author

learning capability and collaborative learning concepts from CSCL is considered. In ITSCL, individual and collaborative learning could be achieved through different learners' interactions. ITSCL provides three different types of interactions. The first level of interaction is learner tutor interaction which is the focus of ITS development. ITSCL second level interaction will be between peer learners. This depicts that learners may reap benefits from the peer learners. The third level of interaction will enable ITSCL interaction between a tutor and a small group of collaborative learners. ITSCL is promising towards combining individual and collaborative learning of a small group of learners within ITS environment.

Following introduction, this paper is organized as: Section II deliberates the review of literature. The architectures of ITS and CSCL is presented in Section III and IV. Based on these architectures ITSCL framework is proposed in Section V. The prototype implementation of the proposed model is presented in Section VI. Evaluation and findings are mentioned in Section VII, following by experimental results and discussion in Section VIII. Lastly, conclusion and future is drawn in Section IX.

II. Literature Work

From prior research, it is inferred that CSCL environments lacks cognitive and adoptive support [7]. An ITS is a prime AI application that provides cognitive and adoptive support, that can be beneficial for collaboration to be successful. Very few studies incorporate the optimal mode of combining individual and collaborative learning in integrated CSCL and ITS environments, mostly collaboration achieved by group formation outside ITS environment through asynchronous communication. Effective collaboration script does not follow simply by group formation. Several exemplary works using CSCL with ITS are described here to provide our research background.

Pierre Tchounikine et al. explored interaction analysis i.e. individual's and collaborative group's interactions are used. In the CSCL environment to provide students with adoptive support, ITS has been used. They outlined the requirements for technological platforms, supporting learners with adaptive guidance and feedback [8]. Maria Virvou proposed a system that incorporates an adaptive system, learner module, error diagnosis, advice generative module, and collaborative module. Collaboration is carried out by asynchronous text messaging in two different languages (French and English) [9]. Another investigation by Maria Virvou et al, the study considered user modelling and machine learning. User modelling is used to collect and analyze user characteristics, also considering these characterizes incorporated to create student groups by using machine learning. These resulting groups promote win-win collaboration [10]. This study only focuses on collaborative learning but they did not explore in ITS environment. Another important study was done out by Daniel Epstein et al, a text-mining tool elaborates on the interaction of the learners within ITS environment. Students are requested to make questions. The learners are required to post questions, using a machine learning approach system to extract meaningful and relevant terms from their text and the

triggering text to extract keywords. These keywords are sent to a web search that retrieves its information [11]. This study does not encourage learner's collaboration instead they seek help from the web. Ronald Cole et al, studied learners' interactions with Marni (virtual science subject tutor). This research investigated that the impact of interactions on learners demonstrated similar significance gains in learning achievement [12]. The collaboration of the students was recorded without involving a tutor; the students discuss and provide a group answer to the tutor. Pravin Chopade et al. enable the ITSs that facilitate collaborative problem solving (CPS). Team interaction data was collected from log data, eye-tracking, and video/audio [13]. The students discuss and collaborate though audio/video platform. Since students achieved collaboration outside ITS system A Generalized Intelligent Framework for Tutoring (GIFT) also supports effective adaptive feedback to an individual learner or team of learners [14]. GIFT assist teams of soldiers, but its implication is only for soldier's teamwork. Another study applied common concepts in the human-agent team, such as task allocation, adaptive triggers, and behavior modelling [15]. Most of the studies achieve collaborative learning outside the ITS environment. Some worked on learning group formation, sharing problem view, learner response sharing, and learner communication audibly and through text messaging or chat communication, etc.

Jennifer K. Olsen et al, Continuously worked for many years to integrate CSCL and ITS and to achieve collaborative learning. ITS authoring tool is extended to Cognitive Tutor Authoring Tools (CTAT), to support collaboration scripts using example-tracing collaborative tutors. This collaboration was obtained by applying multiple parallel example-tracing tutor engines, one engine for each student. These engines collect all the inputs from every student and send outputs to all students [16]. These engines collaborate to share information on peer learners. Another study tests the hypothesis that collaboration may be more beneficial for conceptual knowledge, and less optimal for procedural skills. The dual eye-tracking technique was used to evaluate this hypothesis about collaborative learning [17]. The collaborative version of Cognitive Tutor Authoring Tools (CTAT), where students have shared problem view that can be differentiated through adoptive guidance/feedback, and problem information [18] Jennifer et al, explore; collaboration support for elementary level students using ITS system, raising the strength of collaborative and individual learning in conceptual and procedural knowledge and benefits of two learning methods better than either one alone [19]. The research focused on the analysis of students' interactive dialogue and their behavior in an Intelligent Tutoring System (ITS). The research found that the frequency of interactive talk and errors overtime decrease in the group of two learners working together on conceptual problems [20]. The study investigated the benefits of combining individual and collaborative learning. The results showed that the combined condition had higher learning gains than either alone in individual or collaborative [21]. In cognitive group awareness, every learner answers the question individually, the peers shared their answers, and then they are acquired to provide a collaborative answer [22]. Jennifer K. Olsen et al further investigated how the system constructs

knowledge collaborative using log data, student interactive dialogue, and eye-gaze analysis [23]. All these studies explore collaborative learning in ITS with different angles. The author successfully extends ITS for collaborative learning as well. Although this research work nicely contributes to integrating ITS and CSCL, this research has opened an inquiry into how collaborative and individual learning can be effectively combined. Also, there are some limitations to their work. First, CTAT is only of mathematical fraction problem. The authors did not give any explanation that this domain allows a certain degree of discussion among the students. Selecting the nominator or denominator by each learner or selecting an answer from radio buttons does not allow a certain degree of communication and collaboration. Second, this collaboration is limited to only two learners. As in CTAT equivalent fraction is a two-student activity. An increase in several learners could complex the interaction analysis; most importantly that collaborative learning is not only between learners. The optimal collaborative learners' group size is four. Third, this study more focuses to investigate the outcome of collaborative learning in procedural and conceptual problems rather than general. Fourth, with no prior group formation of the learners, teachers randomly assign learners to a group based on their abilities. Fifth, the learner communicates with another learner through Skype which is recorded but this is outside ITS environment. CTAT itself not capture learner-learner interactions. Sixth, Poor communication among students, as students communicate outside ITS environment, so ITS could not analyze their interaction, engagements, reasoning and sharing knowledge and ideas.

Another important contribution to the collaborative learning in ITS is carried out by Richel Harsley, at the University of Illinois at Chicago. This is also the most relevant study that integrates CSCL and ITS. Harsley extends ChiQat tutor for collaborative learning as well. ChiQat works for programming problems. ChiQat is extended to Collab-ChiQat to support collaborative learning between two pair programmers. Collab-ChiQat provides two types of collaboration. First Non-structured collaboration, that does not provide any feedback on collaborative pair. Second is structured collaboration, that visual feedback on the group and individual performance. Both students logged to the system in a collaborative mode. One student involves in coding and the system monitors the performance while the other helps the peer coder and then turn change accordingly. The students used the headset to record their Skype dialogue [24] [25] [26] [27]. Richel Harsley successfully extends ChiQat to Collab-ChiQat. Collaborative-ChiQat achieves the meaningful collaboration of two learners for a pair programming task. But there is some limitation of these research findings. First, Collab-ChiQat provides collaboration between two learners, but collaboration may involve more than two users. Second, Collab-Chiqat is for pair programming domains that involve only two learners in collaboration. It does not provide a higher degree of collaborative learning. The pair programming domain does not allow a certain degree of discussion among the students. Collaborative groups of more learners could achieve higher learning gains rather than two learner's collaboration. Third, Learners are stationed on a single stationed computer, and their collaboration is not recorded

inside ITS. Fourth, these studies mostly focus on unstructured and structured collaboration rather than a high collaboration script. Fifth, the ITS only measures the performance of the coder, with no collaboration script inside ITS system. Sixth, Students wear headsets to record spoken dialogue. This collaboration could be analyzed by ITS. The students seek help from a peer by spoken dialogue that is outside ITS domain.

The novelty from recent work with real-time support of peer learner's collaboration is facilitated by ITSCL. ITSCL drove the design of ITS for collaborative learning. This enhancement to the framework of traditional one-on-one ITS will provide a collaboration platform where the learner can seek help from peer learners. In the above literature work, ITS is extended for the only group of two students. Also, ITS domain selected was pair programming and fraction problems, which does not involve logical reasoning between two students. ITSCL support a small group of students and provide a different level of interaction among learners. ITSCL provides individual learning by Learner-ITS interaction. It also supports collaborative learning by learner-learner interaction and a tutor-small group of learners. This means of collaboration by ITSCL will increase its learning effectiveness. CSCL community necessitates exploring how students acquire knowledge with real-time support of peer learner's collaboration inside an ITS. The ability to integrate these two ideas (CSCL and ITS) to effectively ameliorate learning remnant is a challenge. This endeavor holds the potential for extending the scope and application of ITS in a broader range for collaborative learning.

III. ITS ARCHITECTURE

Historically research has been focusing on one-on-one ITS, its different tutoring techniques, different scaffolding techniques, student and domain modules. The architecture of traditional one-on-one ITS consists of four modules namely, tutor module, domain module, learner module, and user interface [28] shown in Fig. 1.

A. Tutor Module

It contains all the tutoring strategies for the student. It accepts information from the domain module and the student module. Making use of student module it decides the tutoring strategies, techniques and tactics as well.

B. Domain Module

This module contains the domain knowledge of a specific subject. It includes information about a specific domain/topic. Domain model serves as an expert knowledge which tackles different issues, diagnosis of error provides a standard for learner performance evaluation or to response questions postured by students.

C. Student Module

The student module stores all the information about the student. The student model is used to interpret leaner behavior using ITS. So, the student model emphasizes on cognitive (learner knowledge) and effective (behavioral) states.

D. Interface Module

Learners interact with ITS through the interface. The learner acquires knowledge from ITS interacting by the user interface. The learner acquires learning material from ITS and gives reactions to the learning material. Learning material may include text, graphics, audio, and multimedia, etc. Traditional one-on-one ITS provides a single learner interface as shown in Fig. 1.
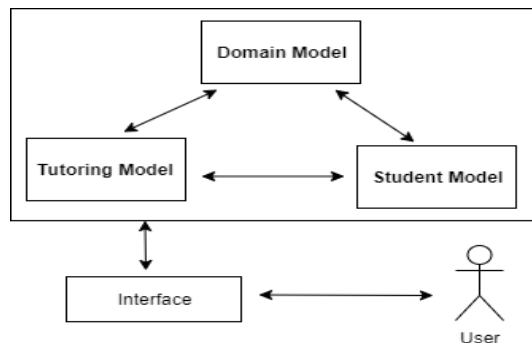

Fig. 1. ITS Architecture [28].

IV. CSCL ARCHITECTURE

CSCL script has different architectures. The most relevant architecture of CSCL that relates to our study is the distributed architecture of CSCL.

A distributed architecture is characterized by the distribution of the Model View Controller (MVC) architecture. MVC is the sharing of components across multiple hosts. In MVC architecture model lives on a shared server and each user has its view and controller as shown in Fig. 2. Each user sits on his computer having his interface view and control. The distributed architecture of CSCL shares features on a live shared and centralized server where specification of view and controller are maintained on the server and sent to user. This supports multiple users connecting across a shared server. Daniel D. Suthers present distributed architecture for CSCL [29].

As distributed architecture supports multiple users in the CSCL environment, so using this architecture collaboration can be fostered in ITS as well. To achieve collaboration in ITS, the distributed architecture of CSCL will support multiple learners to collaborate in the learning process having their own view and control of the system. Multiple learners could interact with ITS and ITS could respond to many learners. This study uses the distributed architecture of CSCL with ITS to achieve both individual and collaborative learning in the ITS environment. Collaboration is one of the emerging learning paradigms in intelligent tutoring systems and education.
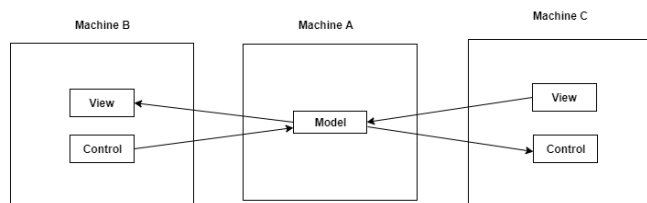
Fig. 2. CSCL Distribute Architecture [29].

V. PROPOSED FRAMEWORK

This proposed framework of an intelligent tutoring supporting collaborative learning (ITSCL) supports both individual and collaborative learning. This system supports three types of interactions. First, Learner-Tutor interaction, second is learner-learner and third is a group of learners-ITS interaction.

A. Individual Learning

State of the art ITS provide an intelligent individual learning facility. This traditional ITS is also referred to as one-on-one ITS (learner-ITS). In one-on-one, ITS single learner interacts with ITS and ITS provides an intelligent support for learning. ITSCL facilitates individual learning by learner-ITS interaction (as shown in Fig. 3 single learner interact using individual learning interface). ITS ask questions from the learner and learner respond to ITS questions. ITS provides adaptive, intelligent feedback and multi scaffolding teaching without human teacher support.

B. Collaborative Learning

One-on-one non-collaborative learning is extended to collaborative learning in the proposed ITSCL framework. Thus, ITSCL supports the collaborative view of learning as well. This collaborative learning is supported by learner-learner interaction and a small group of learners-ITS interaction.

1) Group of learners-ITS interaction: Collaborative learning is accomplished by ITSCL through the group of learners-ITS interaction. All the learners have the same problem view of the ITS. Every learner sits on their computer having a control view that also has a shared problem view with peer learners. System asks the question from a group of learners. This question or problem is shared with all groups of collaborative learners on their interface. Every learner has its control to respond to the system question. So, each learner provides an individual answer at the first step. The answer provided by each learner is shared among all the peer learners. Answer sharing helps learners to get a concept or idea of different responses of the peer learners. After sharing of answers, the learner can collaborate with peer learners by commenting on the peer learners. If the peer learner misses any point or concept in the answer, the peer learner can help and guide on commenting on the shared answers. Every learner is capable of sharing knowledge with a peer by commenting on peer answers. This activity will help learners to share knowledge, ask a question, clear their concept, conceptual understanding transformation, articulate their misconception, collaborate and reflect and apply their knowledge. This also helps students to be more interactive and collaborative. Its helps to increase their interest in response to peer's answer and to express how much he agrees or satisfy with peer answer. After getting a response from peer comments and rating answers, the learner can edit their answers according to peer help to respond to his answer. The system allows the student to modify up to two iterations. If the students edit their answers, the system will show the edited

iteration. Then students finalize their answers. After finalizing, the system will use the Natural Language Process (NLP) to compare or match the finalized answers with the answers in the database. The system chooses that answer as a collaborative answer that highly and best match the system answer. The system accepts the most identical answer as a collaborative answer and provides an adoptive response to all the learners. So, using this activity the learners can best collaborate with the peer learners.

This activity is very useful in many ways. Foremost, every leaner has a shared problem view and control. This breaks the boundaries among learners and makes learners able to work on a single problem together. It also supports every learner to interact with the system and respond. Moreover, this system captures the individual response of each learner. This helps that every learner must have to participate in the activity. Also, we can measure his learning gains from peer collaboration that how much he modifies his answer. Thirdly, the system is highly helpful in knowledge sharing among collaborative learners. Every learner could respond to the peer's answers by commenting on answers. This system ensures highly appreciation and facilitation learners to share knowledge, argue, guide, modify concept, enhance learning and reflect upon their knowledge utilizing commenting on peer answers. Likewise, Students could reflect upon their knowledge by getting responses, guidance, and knowledge sharing in comments. Learners could modify their answers from getting feedback or knowledge from peers. These answer modifications help the learners to enhance their knowledge level and to be more productive.

2) Learners-learner interaction: Learner-learner interaction frequently used technique connecting a group of learners via computer networks. Tighter integration of the learners through chat provides a richer collaborative learning environment [30]. ITSCL provide the third level of interaction by mean of learner-learner interaction. The learner could interact and could seek help from peer learners in a private chat tool. This chat between learner and learner is more focused and productive.

The integrated model of ITS and CSCL into ITSCL, that support individual and collaborative learning by mean of three levels of interaction (learner-ITS interaction, Group of Learner-ITS interaction and learner-learner interaction) is shown in Fig. 3.
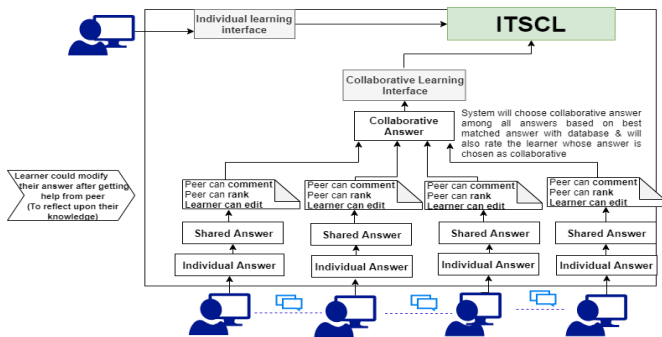
Fig. 3. Proposed ITSCL Framework.

VI. IMPLEMENTATION

The proposed framework of ITSCL is practically implemented into a prototype. This prototype of ITS extends traditional ITS for collaborative learning as well. While many ITS have been used for one-on-one learning, but we used NDLtutor (Negotiation-driven learning tutor). NDLtutor was introduced as an ITS that uses negotiations as a platform for providing instructions to the learners [31]. ITSCL prototype contains all the basic functionalities of NDLtutor along with the deployment of collaborative learning techniques. ITSCL prototype was developed HTML5 (Hypertext Markup Language) and PHP (Hypertext Pre-processor) along with MySQL database. ITSCL performs two main functionalities, a) One-on-one ITS, b) Collaborative ITS.

A. Individual Learning

This is the same functionality that is already provided by existing one-to-one ITS. Inside this environment, the learner works individually without any communication or collaboration with other learners as shown in Fig. 4. ITSCL provides a natural language interactive interface. ITSCL provides tutoring to single learning by the Negotiation Driven Learning (NDL) paradigm of NDLTutor [31]. NDL provides learners to interact with the ITS in an intuitive natural language paradigm. ITSCL asks a question from the learner that guides and facilitate to construct their knowledge. The learner provides the response and Tutor interprets learner response and provides adaptive feedback.

B. ITSCL Collaborative Environment

Collaborative learning involves a group of learners in a learning activity by communicating and collaborating. ITSCL provides collaborative support for learners to work on the same problems. A small group of leaners collaborate with peer learners in the learning process and headed towards the solution of the problem. Learners interact with the tutor from their personal computer having the same problem view. ITSCL first asks the students to record individual responses to the question posed. ITSCL provides an opportunity for every learner to respond a step individually to the ITS question before working on the step as a group. This step is very helpful to actively involve every learner in the collaborative process.
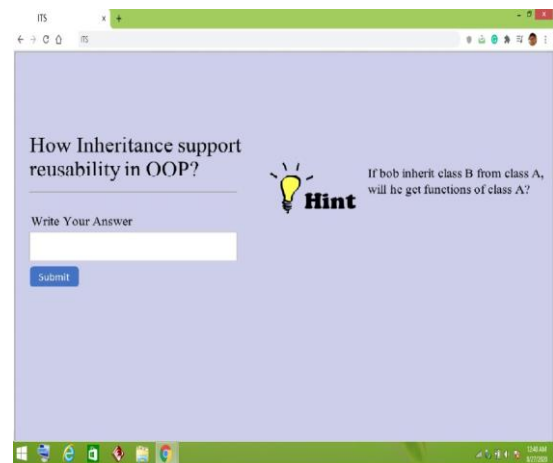

Fig. 4. Individual Learning Interface.

The learner responses are then shared with the peer learners. Every learner can see the peer answers. Sharing of knowledge with collaborative learners is an essential approach in the collaborative learning process. This helps the learner to access peer knowledge in a specific domain. If learners with their different knowledge levels working on the same problem domain, then knowledge flow from high performing students to weak students. This also helps to construct knowledge from peer knowledge if learner missed in individual response.

After individual answering and sharing of answers with colleagues, ITSCL allows the student to access peers' answers and give them rating as stars to the answers and the updated answers. Rating peer response in a small collaborative group of learners could be a method to foster collaboration and provide an encouraging result [32]. This rating procedure allows learners to read and rate the contribution of their peer collaborative members, which can led to a common understanding and constructing knowledge. This helps the learners to look into the peer's responses and how much they agree to the proposed solution. The responses having a high rating might be the most relevant or correct answer. This rating expresses the cognitive contribution of the participants.

As the responses of the learners are shared and can be viewed to other learners. Every learner is also able to guide and help the peer learner by commenting on peer answers. The learner could point out the missing concepts, provide useful guidance and positive feedback on peer answers. This is a valuable activity that helps the learners to reflect upon their knowledge. This allows learners to construct reasoning on their knowledge level, address their knowledge gap and tend towards the problem solution. When learners actively involve themselves, by sharing knowledge and performing collaborative activities with peer learners, it could influence each other's thought processes, articulate reasoning, memory retrieval, rational behavior, and advance their knowledge. In this process, information or knowledge flows from high performing to low performing students among the collaborative group. After getting clues, help, guidance, concept, and knowledge from peer participants, learners can reflect upon their knowledge and can review their answers twice for the same problem. After finalizing the answers, ITSCL analyzes each learner's responses. ITSCL select authentic answer as collaborative and also identify student whose answer is chosen as collaborative. The system also measures every individual response and shows its result to the student while the collaborative chosen answer is shared with all of the members from the collaborative group. This helps learner's individual as well as collaborative accountability in a collaborative group. ITSCL collaborative mode of learning is presented in Fig. 5.

VII. EVALUATION AND FINDINGS

A. Method

For ITCSL evaluation an experiment was performed to test system performance and efficiency and later on comparison was made with one-to-one ITS. The experiment consisted of two different ways of evaluation.
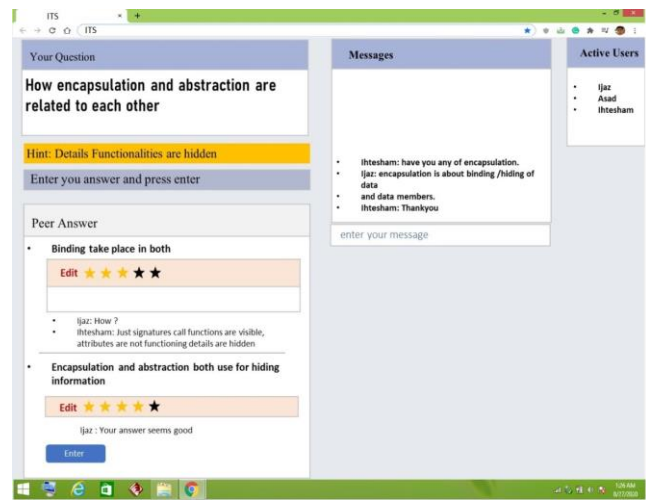
Fig. 5. Collaborative Learning Interface.

Firstly, students learning gain was measured in individual and collaborative conditions. These learning gains were calculated from Pre-Test scores and Post-Test scores of the individual learner. The evaluation further shows the efficiency and performance of the ITSCL concerning the learning gain of concerned learners. Learning gains were analyzed using statistical methods of Paired T-Tests. Students learning gains were analyzed in both individual and collaborative conditions. Secondly, a post-experiment survey was conducted to record user response regarding the usability, effectiveness, and application of both systems and the final verdict of preferring rather a single tutoring system or ITCSL.

B. Participants

For experimentation purpose, we have selected 28 undergraduate students from Bahria University Islamabad Pakistan. This study took place in one session in a computer science programming lab. Those participants were students from the under-graduate program and were selected because they were taught programming concepts for four semesters. They had the experience of using online learning environments, but none of the participants had previous experience or any idea about intelligent tutoring systems. The experiment was conducted on ITSCL that uses NDLTutor technique [31]. Before the session, participants were oriented on how to use the ITSCL and the different available options about the usage.

C. Student's Group Selection

In CSCL generally, small groups of learners are organized using random, self-selection, quiz/assignment or grading-wise selection. In this experimental procedure, the teacher divided students into different groups. The size of the group also matters in CSCL. Smaller groups (less than three) contain less diversity and lack social constructivism, whereas participants in the larger group are difficult to ensure full participation and can lead to complex communication structures. Furthermore, we used the group size of four students. As the students were participating, thus the availability of seven groups of four students was confirmed.

D. Procedures

The experiment was conducted during the students' regular class timings. It was a single interactive session divided into the following stepwise procedure also shown in Fig. 6.

Firstly computer-based Pre-Test was conducted to determine the learner's knowledge level in the object-oriented programming domain. Each test had a total of 8 questions. The questions target the students' concepts of object-oriented programming. We designed Pre-Test and Post-Test and then got it approved by an object-oriented programming instructor. The questions were graded with digit 1 point for each correct answer and digit 0 for the wrong answer.

Secondly, Students were tutored regarding the use of the system and functionality. The learners were explained regarding the user interface, functionalities and especially collaborative perspective of the system that how they can best collaborate with their peers. Students were encouraged to follow the collaborative structure of collaborative learning with their peers. Participants were encouraged to discuss/negotiate with peers to construct a new level of knowledge. Also learning participants were briefed on how they can best collaborate and every member of the group must be given equal opportunity to contribute his/her concepts. After knowledge sharing and helping peers, the learners were directed to an understanding level and a group solution.

Thirdly, intervention session, where students first experienced ITSCL individually and then collaboratively. During Individual learning, single student intervenes ITSCL. This intervention was provided for learner-ITS interaction. ITSCL taught eight related concepts of pre and Post-Tests. ITSCL posted questions to each learner, and the learner answered the questions individually. ITSCL provided adaptive feedback and hints to improve the knowledge level of the students. During Collaborative learning procedures, students log in to the collaborative view of learning and join the peer learners for collaborative learning. The total number of students was divided into seven groups; each collaborative group was consisting of four students. Students were observed to be active, engaged and interested in the collaborative learning environment of ITSCL.

Fourthly, Students attempted Post-Test individually for both individual and collaboratively told conditions for statistical analysis of Paired T-Tests.

Finally, learners fill the post-experiment survey to share their experience about ITSCL usability, performance, collaborative nature and future prospective of the system.

Overall, experiment procedures are presented in Fig. 6.
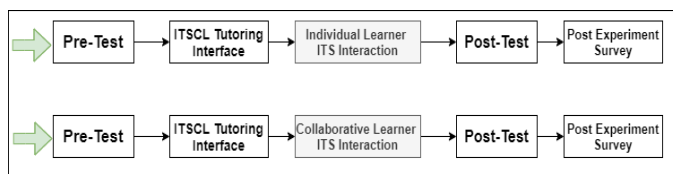
Fig. 6. Study Procedure.

E. Measures

Basically, two measures were taken for the evaluation of the ITSCL. Firstly, students learning gains were taken from pre and Post-Test in individual and collaborative conditions. Secondly, students' perceptions about the ITSCL were recorded through the post-experiment survey.

1) Measurement of learning gain: Pre-Test and Post-Test were taken individually to access learner's conceptual knowledge of object-oriented programming domain in individual and collaborative conditions. These questions were interrelated and counter checked for balance. The test was administered on the computer. There were total 8 questions in each test. For each correct answer, learners received 1 mark. For one incorrect answer learners were marked as 0. Points obtained from all questions are added. Learning gain is determined with the following expression.

Learning gain = Post_Test score – Pre_Test score   (1)

We further performed descriptive statistics of frequency analysis. Frequency analysis is the number of occurrence of scores obtained by the students that is, 5 students get 2 marks, and then its frequency is 5. Frequency analysis showed overall students' performance on Pre-Test and Post-Test in both individual and collaborative conditions.

2) Post experiment survey: After sessions with participants, research conducted a post-experiment survey to record the participant's response and experience about ITSCL. Post experiment survey consisted of five questions regarding ITSCL performance, efficiency, usability, and application. These five questions tend to find the different perceptions of using ITSCL about usability, knowledge gain, student satisfaction, a collaborative learning environment, and its application.

Total of 28 participants answered these questions and their responses are measured on the scale from 1 to 5.

F. Analysis

To analyze the learning gains from Pre-Test and Post-Test, the study performed statistical analysis of Paired T-Tests. To trace student learning gain, a Paired T-Test was applied on the Pre-Test and Post-Test scores in both individual and collaborative conditions. Paired T-Test analysis measure the dependency between dependent variables. The Paired T-Test is a parametric approach that compares the means of the same population applying two different procedures. Paired T-Test gives t-value, which shows the significant difference between Pre-Test and Post-Test. Paired T-Test works on two hypothesizes, null and alternative hypothesis.

Null Hypothesis:

H_0: Pre-Test and Post-Test have no difference

Alternative Hypothesis:

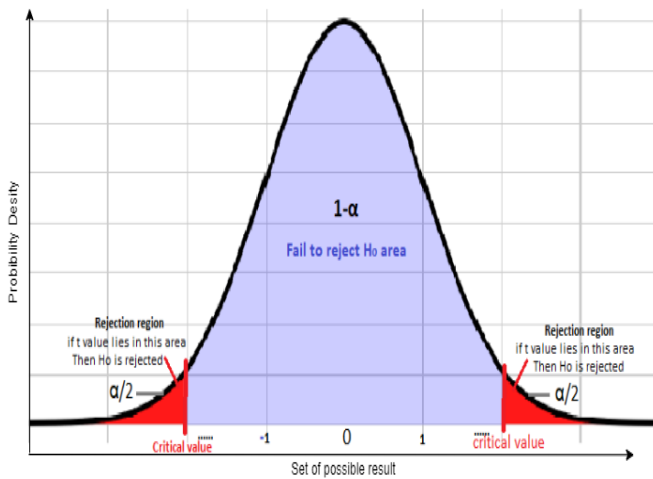H_1: Pre-Test and Post-Test difference matter

Fig. 7.   P Valued Graph.

Performing Paired T-Test on individual and collaborative conditions if T value lies in the rejection region then the null hypothesis is rejected otherwise null hypothesis fails to reject and the alternative hypothesis is accepted. This result is presented on a two-tailed p-value in Fig. 7.

The critical value is the area that is a critical point between rejected areas and failed to reject area and can be found out from t-table using the degree of freedom.

Degree of freedom = df = n – 1   (2)

Keeping 95 % confidence level is inferential statistics that means there is 95% chance that the null hypothesis will be rejected if T values lie in the rejected region. So 95% confidence level means:

α = 0.95   (3)

The above p-value graph shows the significant difference, that how much improvement is observed and is analyzed from the mean value, standard deviation, t-value, and significance. Results of Paired T-Tests are also evaluated on the mean value, standard deviation, t-value, and significance show learning gain improvement.

We also performed descriptive statistics of frequency analysis. Frequency analysis showed overall students' performance on the Pre-Test and Post-Test scores. This analysis shows that some weak students got zero, or could score one or two marks in the Pre-Test, improved their marks in the Post-Test. So we trace the frequency table of marks obtained in Pre-Test and Post-Test scores in individual and collaborative conditions to trace the effectiveness and performance of ITSCL.

Subsequently, we also analyzed student perception of using ITSCL. A Post-experiment survey was performed using a Likert scale. A total of 28 students participated in the survey. We calculated the mean and percentages of students' performance.

VIII. RESULTS AND DISCUSSION

The results of Paired T-Test and frequency analysis show improvement in learning gain and performance. The result of the post-experiment shows the student's perception of using ITSCL. These are given below.

A. Learning Gain

Learning gain was measured from Paired T-Test and frequency analysis. Results from both these evaluations are given below:

1) Paired T-Test Result: A Paired T-Test was applied on Pre-Test and Post-Test in individual and collaborative conditions.

T value of Paired T-Test was -4.121 that lie in the rejection region as shown in Fig. 8. However, the null hypothesis is rejected and it concludes that it is significantly different in Pre-Test and Post-Test in the individual condition. Therefore, it concluded to fail the H_o hypothesis.

Further interpretation of Paired T-Test gives mean, standard deviation, t-value, and significant difference. For individual condition, Paired T-Test was applied on Pre-Test and Post-Test. Keeping 27 degrees of freedom, the mean of Post-Test (4.46) was higher than the pre-test (3.04). Moreover, there was a significant difference of p (0.0001) that is less than 0.05 between Pre-Test and Post-Test. These results show a significant difference between Pre-Test and Post-Test. Learning gain was measured by taking the difference between Post-Test and Pre-Test, then t-test was applied to calculate mean (1.892), standard deviation (1.448) and highly significant p value (p<0.05).
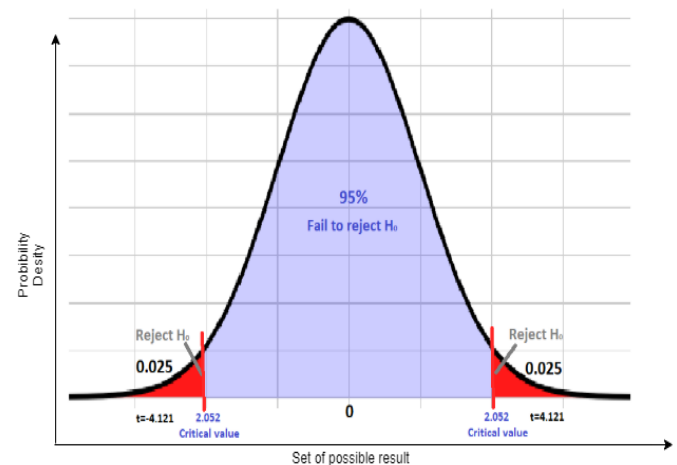
Fig. 8.   P-Value Paired Test Graph for Individual Condition.

In collaborative conditions, T value of Paired T-Test was -5.872 that lie in the rejection region as shown in Fig. 9, thus null hypothesis is rejected and concluded that there is a significant difference in Pre-Test and Post-Test in the individual condition. Therefore, it concluded to fail the H_o hypothesis.
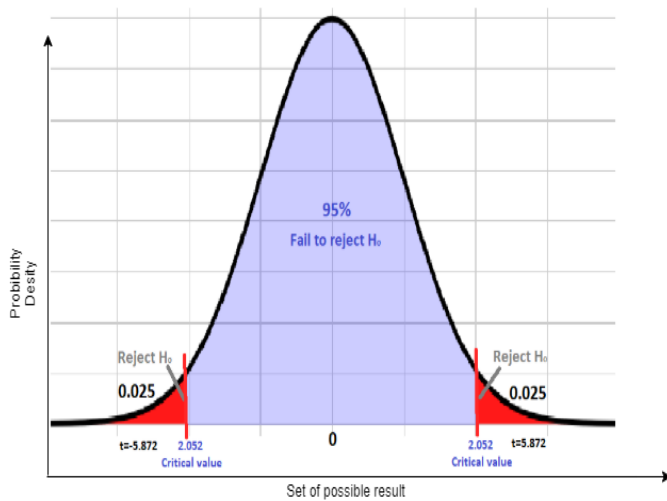
Fig. 9. P-Value Paired Test Graph for Collaborative Condition.

To measure the improvement level is analyzed from the mean value, standard deviation, t-value, and significance value. Keeping 27 degrees of freedom, the mean of Post-Test (6.035) was higher than the Pre-Test (4.035). Also, there was a significant difference of p (0.0001) that is less than 0.05 between Pre-Test and Post-Test. These results show a high significant difference between Pre-Test and post-test. Learning gain was measured by making a difference between Post-Test and Pre-Test, then t-test was applied to calculate mean (2.571), standard deviation (1.77) and highly significant ($p<0.05$). The overall Paired T-Test result is given in Table I.

TABLE I. PAIRED T-TEST RESULTS

Conditions	N	Pre Test		Post Test		Gain	
		Paired T-Test Result					
		Mean (μ)	SD (σ)	Mean (μ)	SD (σ)	Mean (μ)	SD (σ)
Individual	28	3.04	1.972	4.46	1.478	1.892	1.448
Collaborative	28	4.035	1.764	6.035	1.580	2.571	1.77

B. Individual Learning Frequency Measurement

The study calculated the frequency analysis of Pre-Test and Post-Test. Frequency analysis simply counts the number of times that each variable occurs. Here Frequency analysis shows the frequency or number of students in Pre-Test and Post-Test. Applying descriptive statistical analysis showed an interesting measure of student performance. The study observed that in the Pre-Test there are students who performed weak or average, improved their performance in the Post-Test.

Considering Pre-Test scores some students got 1 or 2 marks, while in Post-Test frequency analysis there were no students with results of 1 or 2 marks. This frequency analysis showed improvement students learning gain in individual learning. The frequency analysis table of Pre-Test and Post-Test is given in Fig. 10.
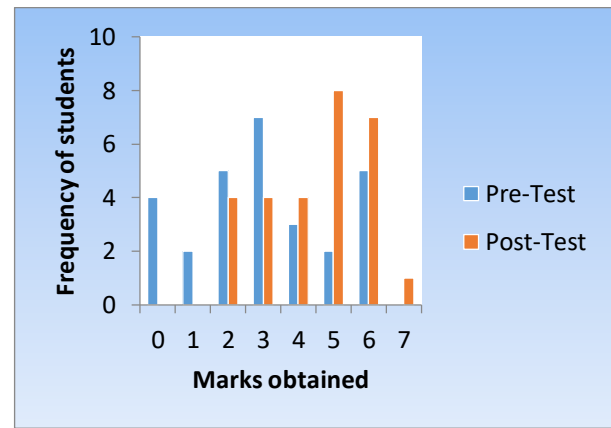
Fig. 10. Pre-Test and Post-Test Frequency Comparison.

C. Collaborative Learning Frequency Measurement

The study evaluates the frequency analysis for collaborative learning as well. Because in collaborative learning there is a small group of learners involved in the learning environment, so it might have a different result. Applying frequency analysis there were interesting results that illustrates improvement in the performance. In Pre-Test frequency analysis some students got 1 or 2 marks, but in the Post-Test, there were no students in 1 or 2 marks. Likewise, only two students in the Pre-Test got only full eight marks, But in the Post-Test, seven students obtained full marks. Overall frequency analysis of Pre-Test and Post-Test is presented in Fig. 11.
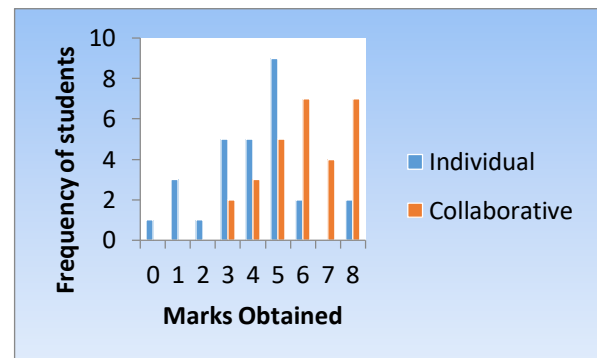
Fig. 11. Frequency Graph in Collaborative Condition.

D. Comparing Individual and Collaborative Learning Gain

Analyzing statistical measurements, we found interesting facts that a collaborative learning environment resulted in high learning gain as compared to individual learning conditions. This conclusion was drawn from calculating the mean value of both learning gain shown in Table II, where collaborative learning condition has a high mean than individual condition.

TABLE II. COMPARING LEARNING GAINS

Conditions	Learning Gain	
	T-Test Result	
	Mean (μ)	SD (σ)
Individual	1.89	1.449
Collaborative	2.571	1.772
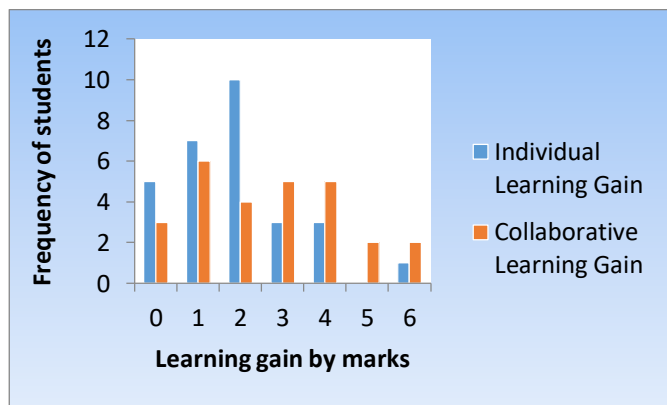
Fig. 12. Individual and Collaborative Learning Gain.

For additional interpretation of comparing individual and collaborative learning gains, bar charts are plotted in Fig. 12. These representations also clearly show that collaborative learning has better results than individual learning.

E. Post Experiment Survey

Table III shows the participants' responses and experience of using ITSCL. Students' responses were recorded for about five different questions. These questions were related to usability, performance, collaborative learning, and future application perspective.

1) ITSCL usability feature: ITSCL provides interactive environment and user interface. User interface in an important aspect of human computer interaction and computer supported collaborative learning. User interface prospective is to facilitate learners to have interactive user interface. The post experiment survey showed significant interest of learners that ITSCL provide interactive environment. Post experiment survey indicated that about 90% as shown in Table IV, learners agreed that user interface was user friendly and facilitated learners to understand system. Graphical representation of experimental results of question 1 is presented in pie char in Fig. 13.

2) ITSCL helps in learning gain: ITSCL provides full control of tutoring and facilitates user in learning process. This is most the significant aspect to evaluate ITSCL. This evaluation helps to find weather ITSCL help learners to tutored and improve their learning gains. The Post experimentation resulted in participants 39% strongly agreed and 51% agreed that ITSCL helped them in tutoring and facilitate in learning process. Overall students' responses to questions 2 are shown in Table V. Fig. 14 below illustrate the results of users' satisfaction using pie chart.

3) User satisfaction: ITSCL is more interactive and enhances user satisfaction due to collaborative nature. ITSCL extend one-to-one ITS for collaboration script. In collaborative view of learning, learners collaborate each other's in learning process. If the system is more interactive due to collaborative nature, this will be helpful in improving learning gain of collaborative students. Post experiment evaluation, showed learners satisfaction that ITSCL facilitate

learners in constructive collaborative knowledge. Post experiment showed 32% strongly agreed and 48% user satisfaction in supportive collaborative nature. Students' responses to questions 3 are presented in Table VI. User satisfaction due to collaborative nature is graphically presented in Fig. 15.

4) ITSCL collaborative methodology: ITSCL methodology is more useful for sharing knowledge with other peers. ITSCL proposed a novel framework model for integrating CSCL and ITS. It is important to evaluate ITSCL for knowledge sharing aspect of collaborative learners. This evaluation shows 57% strongly agree and 32% agree choice of the participants. Table VII displays the usefulness of ITSCL proposed framework for sharing knowledge among the learners and graphically depicts the results in Fig. 16.

TABLE III. POST EXPERIMENT STUDENTS RESPONSES

	<strongly agree…strongly disagree>					Mean
	(5)	(4)	(3)	(2)	(1)	
ITSCL provides an interactive environment and user interface.	08	17	0	2		4.035
ITSCL provides full control of tutoring and facilitates the user in the learning process.	11	15	1	1	0	4.178
ITSCL is more interactive and enhances user's satisfaction due to collaborative nature.	10	15	0	3	0	4.14
ITSCL methodology is more useful for sharing knowledge with other pairs.	16	9	0	2	1	4.321
ITSCL methodology can be used as an effective technique for teaching classes in the future.	10	14	1	2	1	4.071

TABLE IV. QUESTION 1 RESPONSES

Total Participants	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
28	08	17	0	2	1
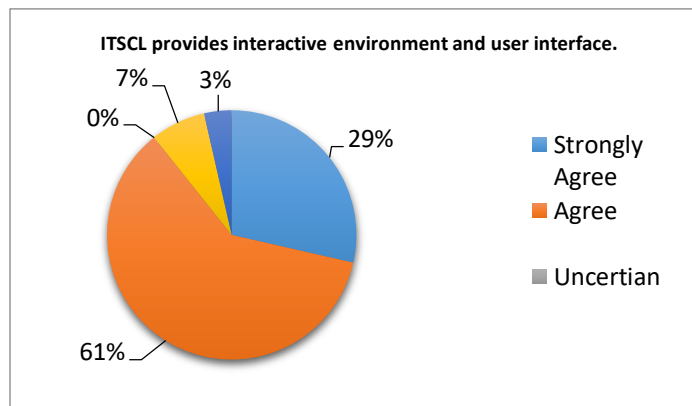
Fig. 13. Overall Results of Question 1.

TABLE V.   QUESTION 2 RESPONSES

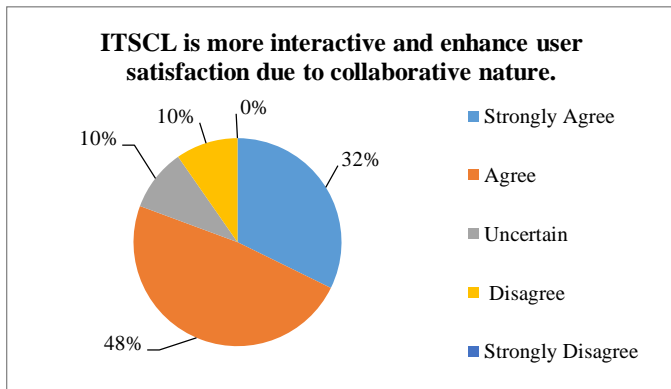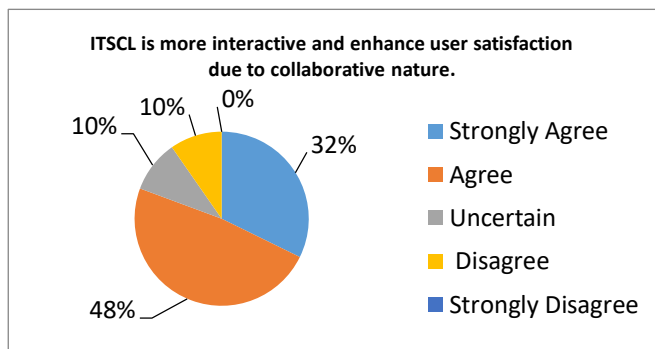| Total Participants	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
28	11	15	1	1	0

ITSCL is more interactive and enhance user satisfaction due to collaborative nature.

Fig. 14. Overall Results of Question 2.

TABLE VI.   QUESTION 3 RESPONSES

Total Participants	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
28	10	15	0	3	0

ITSCL is more interactive and enhance user satisfaction due to collaborative nature.

Fig. 15. Overall Results of Question 3.

TABLE VII.   QUESTION 4 RESPONSES

Total Participants	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
28	16	9	0	2	1

ITSCL methodology is more useful for sharing knowledge with other pairs
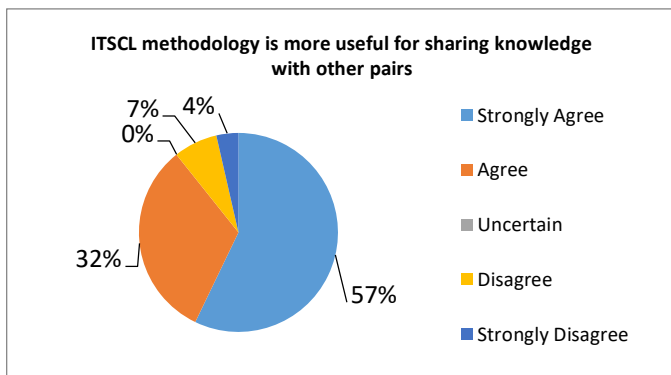
Fig. 16. Overall Results of Question 4.

5) ITSCL Future Implication: ITSCL methodology can be used as an effective technique for teaching classes in future. ITS have been used practically successfully in classroom for teaching. Single learner ITS provide cognitive support for individual learning. However in classroom teaching is almost in collaborative nature. As ITSCL provide support for learners in individual and collaborative learning environment, so it is important to evaluate ITSCL suitability for effectiveness in classroom teaching. Interestingly participants 36% strongly agreed and 50% agreed, ITSCL could be applicable as an effective technique for classroom teaching. Participants' responses to question 5 are shown in Table VIII and illustrated in Fig. 17.

TABLE VIII.   QUESTION 5 RESPONSES

Total Participants	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
28	10	14	1	2	1

ITSCL methodology can be use as an effective technique for teaching classes in future.
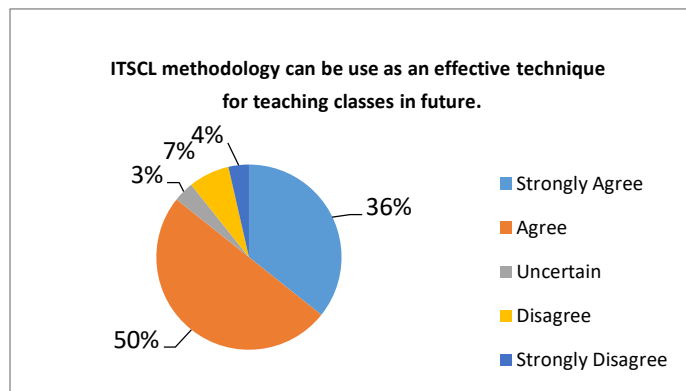
Fig. 17. Overall Results of Question 5.

IX. CONCLUSION AND FUTURE WORK

The research work was related to the redesigning of a traditional one-to-one tutoring system to facilitate collaborative learning. In this research study, we integrated ITS and CSCL to propose a design framework of Intelligent Tutoring Supported Collaborative Learning (ITSCL) that supports both individual and collaborative learning. Individual learning is achieved by leaner-ITSCL interaction. While collaborative learning is achieved by groups of learners-ITSCL interaction and learner-learner interaction. This extension of ITS for collaborative learning could be more effective because it allows students to share knowledge, articulate reasoning and misconceptions and reflect upon their knowledge with peer learners, thus developing a deeper understanding. Learners sharing knowledge, misconception, and reasoning, reflect upon their earlier responses and had an opportunity to build fluency with individual capability. The strengths of individual and collaborative learning are integrated into ITSCL to enhance the students learning. This study also developed a prototype of a proposed model of ITSCL for evaluation. ITSCL provides functionalities to both individual learner and groups of learners. After the implementation, we evaluated ITSCL through experiments. ITSCL evaluated for both individual and collaborative

learning. A group of 28 students participated in experiments and learning gains of students and the post-experiment survey was recorded. Learning gains were measured from Paired T-Test and frequency analysis showed significant learning gains and improvement in the learning process. Another evaluation of the post-experiment survey was collected to evaluate the efficiency and performance of ITSCL. This study achieved its goal to have both collaborative and cognitive support with improved learning performance.

This research indicates a promising direction to explore the support of collaborative learning that affects learning and social participation. In this study, ITSCL is not analyzing intra-group communication and learner interactions, therefore, further research is needed to explore this gap and ITSCL should provide real-time feedback on collaborative activities. Another future research perspective is group cognition by an intelligent tutoring system and predicting student performance in a collaborative learning paradigm. Further research is needed on group cognition of collaborative learners by intelligent tutoring systems. Our study indicates that this would be a promising direction for future research to explore.

REFERENCES

[1] Hartley, JR, and Sleeman, DH, "Towards more intelligent tutoring system", International Journal of machine studies, Volume 5, Issue 2, pp. 215-236, April 1973.

[2] Rachel Carlos, Seiji Isotania, Carla Rodriguezac, Kamila Lyra, Patrícia Jaques, Ig Bittencour, "Affective states in computer-supported collaborative learning: Studying the past to drive the future", Computer and Education 120, Volume 120, pp.29-50, 2018.

[3] Yiping Lou, Philip C. Abrami and Sylvia d'Apollonia, "Small group and individual learning with technology: A meta-analysis", Review of Educational Research, Vol. 71, No. 3, pp.449–521, 2001.

[4] I. Magnisalis, S. Demetriadis and A. Karakostas, "Adaptive and Intelligent Systems for Collaborative Learning Support: A Review of the Field," in IEEE Transactions on Learning Technologies, Vol. 4, no. 1, pp. 5-20, Jan.-March 2011.

[5] Jun Xie, Shirin Mojarad, Keith Shubeck, Alfred Essa, Ryan S. Baker and Xiangen Hu, "Student Learning Strategies and Behaviors to Predict Success in an Online Adaptive Mathematics Tutoring System", 10th International Conference on Educational Data Mining, January 2017.

[6] Salgueiro, Fernando Costa, Guido Cataldi, Zulma Lage, Fernando García Martínez, Ramón, "Redefinition of basic modules of an intelligent tutoring system: the tutor module." VII Workshop de Investigadores en Ciencias de la Computación, pp. 444-448, 2005.

[7] A. C. Graesser, P. Chipman, B. C. Haynes and A. Olney, "AutoTutor: an intelligent tutoring system with mixed-initiative dialogue," in IEEE Transactions on Education, vol. 48, no. 4, pp. 612-618, Nov. 2005.

[8] Tchounikine, Pierre & Rummel, Nikol & Mclaren, Bruce, "Computer Supported Collaborative Learning and Intelligent Tutoring Systems", Advances in Intelligent Tutoring Systems, SCI 308, pp. 447–463, 2010.

[9] M. Virvou, C. Troussas and S. Sidiropoulos, "Collaborative Support in a Multilingual Tutoring System," 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Piraeus, pp. 502-505, 2012.

[10] M. Virvou, E. Alepis and C. Troussas, "User Modeling on Communication Characteristics Using Machine Learning in Computer-Supported Collaborative Multiple Language Learning," 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, Athens, pp. 1088-1093, 2012.

[11] Epstein, Daniel & Pinho, Isis & Acosta, Otavio & Reategui, Eliseo. (2013). "Inquiry-based learning environment using intelligent tutoring system", Proceedings - Frontiers in Education Conference, 1072-1074. 10.1109/FIE.2013.6684991, 2013.

[12] Ronald Cole, Cindy Martin, Timothy Weston, Liam Devine, Jeannine Myatt, Brandon Helding, Sameer Pradhan, Margaret McKeown, Samantha Messier, Jennifer Borum, and Wayne Ward, "One-on-one and Small Group Conversations with an Intelligent Virtual Science Tutor", Computer Speech & Language, Volume 50, pp. 157-174, July 2018.

[13] Chopade, P., Khan, S., Stoeffler, K., Edwards, D.T., Rosen, Y., & Davier, A.V, "Framework for Effective Teamwork Assessment in Collaborative Learning and Problem Solving", International Journal of Artificial Intelligence in Education, pp. 1-40, 2017.

[14] Sottilare, R.A., Baker, R.S., Graesser, A.C. et al. Special Issue on the Generalized Intelligent Framework for Tutoring (GIFT): Creating a Stable and Flexible Platform for Innovations in AIED Research. Int J Artif Intell Educ 28,139–151, 2018.

[15] Kaitlyn Ouverson, Mariangely Iglesias-Pena, Jamiahus Walton, Stephen B Gilbert,and Michael C Dorneich, "What Intelligent Team Tutoring Systems Can Learn from Human-Agent Teams", proceeding of Techmind Society, USAArticle No. 28 , pp. 1, 2018.

[16] Jennifer K. Olsen, Daniel M. Belenky, Vincent Aleven, Nikol Rummel, "Intelligent Tutoring Systems for Collaborative Learning: Enhancements to Authoring Tools", International Conference on Artificial Intelligence in Education, AIED: Artificial Intelligence in Education, pp. 900-903, 2013.

[17] Daniel Belenky1, Michael Ringenberg1, Jennifer Olsen1, et al, "Using Dual Eye-Tracking Measures to Differentiate Between Collaboration on Procedural and Conceptual Learning Activities", 10th International Conference on Computer Supported Collaborative Learning Madison, Wisconsin, pp. 15-19, June 2013.

[18] Jennifer K. Olsen et al, "Using an Intelligent Tutoring System to Support Collaborative as well as Individual Learning", ITS, 2014, LNCS 8474, pp. 134–143, 2014.

[19] Jennifer K. Olsen, Vincent Aleven, and Nikol Rummel, "Toward Combining Individual and Collaborative Learning Within an Intelligent Tutoring System", AIED, Springer International Publishing Switzerland, LNAI 9112, pp. 848–851, June 2015.

[20] Jennifer K. Olsen, Vincent Aleven, and Nikol Rummel, "Adapting Collaboration Dialogue in Response to Intelligent Tutoring System Feedback", Artificial Intelligence In Education AIED, 2015, LNAI 9112, pp. 748–75, 2015.

[21] Jennifer K. Olsen, Vincent Aleven, and Nikol Rummel, "Learning Alone or Together? A Combination Can Be Best!", CSCL Proceedings, 2017.

[22] Olsen, J., Sharma, K., Aleven, V., & Rummel, N., "Combining Gaze, Dialogue, and Action from a Collaborative Intelligent Tutoring System to Inform Student Learning Processes", In Kay, J. and Luckin, R. (Eds.) Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 13th International Conference of the Learning Sciences (ICLS), V.No. 2, 2018.

[23] Jennifer K. Olsen, Nikol Rummel and Vincent Aleven, "It is not either or: An initial investigation into combining collaborative and individual learning using an ITS", International Society of the Learning Sciences, Inc.14, pp. 353–381, 2019.

[24] Rachel Harsley, Nick Green, Barbara Di Di Eugenio, Satabdi Aditya, Davide Fossati and Omar Al Zoubi, "Collab-ChiQat: A Collaborative Remaking of a Computer Science Intelligent Tutoring System", CSCW '16 Companion: Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, pp.281–284, February 2016.

[25] Harsley R., Di Eugenio B., Green N., Fossati D., Acharya S. "Integrating Support for Collaboration in a Computer Science Intelligent Tutoring System". In: Micarelli A., Stamper J., Panourgia K. (eds) Intelligent Tutoring Systems. ITS 2016. Lecture Notes in Computer Science, vol. 9684, pp. 227–233, 2016.

[26] Rachel Harsley et al, "Interactions of Individual and Pair Programmers with an Intelligent Tutoring System for Computer Science", SIGCSE '17 Proceedings of the ACM SIGCSE Technical Symposium on Computer Science Education, Washington USA. PP. 285-290, March 2017.

[27] Rachel Harsley et al "Enhancing an Intelligent Tutoring System to Support Student Collaboration: Effects on Learning and Behavior", AIED, LNAI 10331, pp. 519–522, June 2017.

[28] Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi, "Introduction: What Are Intelligent Tutoring Systems, and Why This Book?", Advances in Intelligent Tutoring Systems, SCI 308, pp. 1–12, 2010.

[29] Daniel D. Suthers, "Architectures for Computer-Supported Collaborative Learning", ICALT '01: Proceedings of the IEEE International Conference on Advanced Learning Technologies , August 2001.

[30] Levette Dames et al, "Active Student Engagement through the Use of WebEx, MindTap, and a Residency Component to Teach a Masters Online Group Counseling Course", Handbook of Research on Transformative Digital Content and Learning Technologies, Chapter: 14, Publisher: IGI Global, Editors: Jared Keengwe, Prince Hycy Bull, pp.245-268, 2016.

[31] Raja M. Suleman et al, "NDLtutor: An Automated Conversational Agent to Facilitate Metacognitive Skills in Fully-Negotiated OLMs", International Conference on Intelligent Tutoring Systems ITS, pp. 354-360, 2016.

[32] Ioannis Magnisalis et al, "Can peers rate reliably as experts in small CSCL groups?", 13th International Conference on Intelligent Tutoring Systems, ITS, 2016.

Secure Access Control Model for Cloud Computing Environment with Fuzzy Max Interval Trust Values

Aakib Jawed Khan[1]
Research scholar
Department of Electrical Engineering
JamiaMilliaIslamia
New Delhi, India

Shabana Mehfuz[2]
Professor
Department of Electrical Engineering
JamiaMilliaIslamia
New Delhi

Abstract—Cloud computing needs service provider with reliable communication for increasing the user trust. As existence of cloud depends on quality of services, evaluation of this trust value needs to be carried out by the cloud. Many of the web services provided by E-commerce, social sites, digital platform maintain this for the faith of user by estimating the reliability of service provider. This paper focuses on a model that can identify real nodes by its behavior in cloud. Here fuzzy max interval values have been evaluated from the transactional behavior of the node in fixed interval. By increase in transaction count, trust value of real node trust increases and trust value of malicious nodes decreases. The work is based on Role based Access Control (RBAC), which has three type of roles (Admin, Data owner, Node). Data owner content security was achieved by AES algorithm and only trusted node can access those resources. Experiment was performed by carrying out simulations on ideal and environment under attack. Analysis of evaluation parameters values shows that proposed model of fuzzy max interval trust is better as compared to other existing Domain Partition Trust Model (DPTM), for identification of malicious nodes.

Keywords—*Cloud computing; encryption; fuzzy logic; trust computing; role based access control; resource management*

I. INTRODUCTION

Cloud computing offers web supported services on an efficacy source to the trade development. Shared resources are allotted to the service provider on rent where working and management of allotted resource is done independently. Therefore safety is a primary importance in the cloud atmosphere. The client loses the control of information in the cloud environment and therefore an appropriate faith system is necessary to guarantee safety and security of records [1]. As the cloud computing is collection of diverse narrow systems and embraces the associates from various surroundings, the safety in cloud is obscure. On one side, the safety machinery should offer warranty protection adequate to the client, on the further side, the safety machinery should not be too compound to put the clients into an inopportune circumstances. The flexibility of the client nodes and commercial operating systems has significant features for sustaining their wide acceptance. Conversely, that extremely similar directness and suppleness have been confirmed to be a dual edged weapon, since it brings difficulty, diminishes conviction degree and danger against safety. So there should be stability among the safety and the expediency [2]. While downloading records

from the web, the clients mistakenly downloads damaging software such as key logger. The user-sensitive information such as login and password gets hacked by the software such as Spyware, Trojans etc. The service provider works with the client interface in order to access the cloud services. The information in the contaminated PC is no longer secure. Thus even after engaging every protection method such as installing antivirus software, there exist the threats of information getting hacked when we utilize the internet services of cloud computing [3].

A trust management system assists cloud service suppliers and clients to gather the advantage brought about by cloud computing equipment. Regardless of the advantages of trust management, more than a few problems connected to universal trust measurement mechanisms, doubted comments, poor verification of comments, solitude of applicants and the need of comments incorporation is still required to be tackled. Customary trust organization approaches such as the exercise of Service Level Agreement (SLA) are insufficient for compound cloud atmospheres. The unclear rules and uncertain scientific specifications of SLAs can guide cloud service clients to be incapable to recognize dependable cloud services [4, 5].

In this paper, a fuzzy max interval based trust model is given which uses centralized data updation. It identifies real node behavior by calculating the trust value and secures the data by using role based model and owner resource encryption.

Rest of the paper is organized as follows. Section II gives a brief literature survey of trust models in cloud computing. In Section III, proposed model has been presented. Section IV describes the results and implementation. Conclusion is given in Section V.

II. LITERATURE SURVEY

In [6], authors has concentrated on investigation of secrecy and information sensitivity & safety issues in cloud structural design and atmosphere covering all the phases of existence cycle of statistics. In this paper, the authors detailed privacy protection, information protection, records separation, cloud safety and cloud computing. They have examined these subjects and as well given a key for determining these problems. These problems are chiefly at SPI (SaaS, PaaS, IaaS) level and the most important dispute is information

distribution. Once work achieves security of data and privacy of data owner, paper has evaluated the trust on the basis of social relationship based on past transaction and energy availability of nodes.

In [7] authors have provided a platform to utilize an extensive variety of services that are based on the web to deal with our business events & a variety of services of data equipment. But in addition to all the benefits, it also boosts the risk for safety when a TTP (Trusted Third Party) is concerned. By connecting a TTP (Trusted Third Party) there still remains a possibility of heterogeneity of clients which affects safety on a cloud. In this investigation, the authors suggest a TTP (Trusted Third Party) autonomous approach for IDM (Identity Management) with the ability of utilizing unique information on unreliable information protection procedures for Building faith in Cloud Computing. By means of predicate information over the set information and utilizing multi association computation and computing and vigorous package method are the approaches utilized at this point. In this system the package has self-reliability checking procedure. It comprises of protection methods, privacy plans and virtual mechanism for strategy enforcement of these plans. The declaration lets the usage of IDM solicitation on untrustworthy clouds. Cloud computing is extremely effectual safety service that is based on conceptual knowledge. Information retrieval and safety of the data is the major area of work in cloud computing.

In [8] authors have suggested an innovative trust model with the help of an algorithm to reduce trust management load on system and increase malicious node detection accuracy. Detachment of nodes into domains is helpful for decreasing the overhead of trust management in terms of trust storage and computation. Domain and cross-domain sliding-windows are planned and operated to accumulate the nearly all fresh conviction values. Then, an algorithm is intended to calculate domain and cross-domain trust values for nodes, and a filter process is implemented to eliminate hateful trust assessments and hateful nodes from a domain.

Azad et al. [9] have proposed a node to node trust evaluation algorithm in IOT. Only standing social belief metric is measured in this learning. The applicants allocate a trust rate to the machine based on their knowledge and communications with the machine. Then, they drive faith values' cryptograms to the bulletin board. Using safe multi-party calculation procedures, the reputation activist analyzes the universal status of machine by using the information cryptograms in the bulletin panel.

Rafey et al. [10] have improved collaboration among trusted nodes and regulated the faith scores dynamically based on the node performance. In this work, node operation characteristics (e.g., node calculation power, assurance, perspective importance, and response), and node community characteristics (e.g., friendship, centrality, and connection) are measured. In the trust calculation stage, each node calculates in general faith values of further nodes supports on its own straight communications and proposals from further nodes. In addition, their representation incorporates the community relationships and background of contacts in the faith calculation. The trust correctness in this representation can be amplified by suggestions from false nodes that allocate superior trust values to their collection of associates.

In [11] author has developed a health environment for an effective network services with resistance against malicious node trust attacks. To calculate the trust paper has utilized node energy as well as social performance of node in the network based on social similarities. This paper has not considered the dynamic nature of nodes in the network.

So a model was required to overcome the problem of increase in the decision time for interdomain communication in de-centralized systems, as untrustworthy values are stored in the domain or cross-domain Algorithm [8]. Malicious nodes can increase the trust of other malicious node, hence detection of these kind of nodes is essential. An algorithm to monitor the behavior should be developed for the classification of nodes, as malicious nodes always produce some unfair pattern which can be identified.

Proposed model presents a fuzzy max interval based trust model in unreliable cloud access environment. The main contributions of the present work are:

1) For dynamic service adoption in cloud environment, a trust evaluation model with fuzzy max interval is provided while using centralized data updation which enhances the efficiency of proposed model.

2) Proposed algorithm identifies the real nodes by node behavior in cloud environment and calculates trust value for each node to assert the authenticity of the node.

3) To secure the data, it used role based model and owner resource encryption. It evaluates the proposed fuzzy max interval trust model by calculating the malicious node convergence and trusted node convergence.

Access Control Systems: There are three type of access control systems present, discretionary access control (DAC), Role based Access control and mandatory access control (MAC). Proposed strategy is works on RBAC. It is a model used in secured network for defining role and privileges. As per role, access of resources is granted. This model can be further enhance by monitoring role to role sequence of actions for finding malicious activity by the user / employee / etc. in a organization or network. This paper has also improved this monitoring system by estimating the user's social trust as per the series of actions performed.

III. Proposed Model

This section gives a brief review of the proposed model with the help of a block diagram along with the explanation of different section of the blocks. The proposed algorithm provides a complete architecture of the work as well. This work assumes an area with N number of nodes where one node communicates with other. Communication approach between two nodes is termed as transaction T. Paper has three types of roles; first is Admin, second is Data Owner, third is Node. Information of each node and transaction count, with successful number of transaction was centralized. Hence trust evaluation of the nodes was also centralized. Table I shows various symbols used in the proposed algorithm.

Symbol	Meaning
N	Node
D_{ij}	Direct Trust between i, j nodes
CB	Centralize Bridge
T_t	Trust of ith node
S_t	Successful Transaction
M	Malicious Node
R	Real Node
K	Total Number of transaction between i, j node
Iij	Fuzzy Interval value of Ith node w.r.t. jth node
T_k	Kth transaction between nodes
M	Number of Window frame

A. Roles Definition

1) Admin (CB): This is centralized body in the model which takes care of all data owners, Node, and maintains different other information which is required for working securely. So admin manage all set of password, number of transaction, and trust value. Here all set of password, number of transaction, trust are also stored and updated by admin.

2) Data Owner (DO): This role acts as service provider where data stored by the owner is accessed by other type of Node. To enhance the security of data owner CB provides security access to DO for creating initial access of data. To further enhance the security, data is encrypted by using AES algorithm. Hence each data owner has its own set of Keys shared by DO to its user (Node). AES algorithm first encrypts the data and stores on cloud while its user can access this by search query.

3) Node: This role is created by DO, so if DO creates a Node then it can access DO data by search option. As DO has created this Node so reverse keys of AES were applied for that data owner Node. Search query of the node is first encrypted, so security of user query maintained by this operation. Entire searching function does not de-crypt resource (DO data). Search query is also encrypted for increasing the communication security, as work is performed on unreliable cloud.

4) Hierarchal cluster: All set of nodes as per types of services were grouped into set of clusters. As cluster size increases to a certain threshold, further clustering of nodes takes place. This is shown in Fig. 1. This reduces communication cost [12]. Information of these nodes in form of number of total transactions, number of successful transactions and trust score was maintained by centralized bridge CB.

5) Window: In this work, m size window moves to monitor the transaction behavior of various nodes in different domain of the network. Here after each m number of time frames, trust value of the nodes were evaluated. This can be understood by diagram present in Fig. 2, where each block is

m^{th} time frame where any number of transactions may occur between nodes. This transaction may be of same or different domain. Fig. 3 shows the flowchart of malicious node detection.
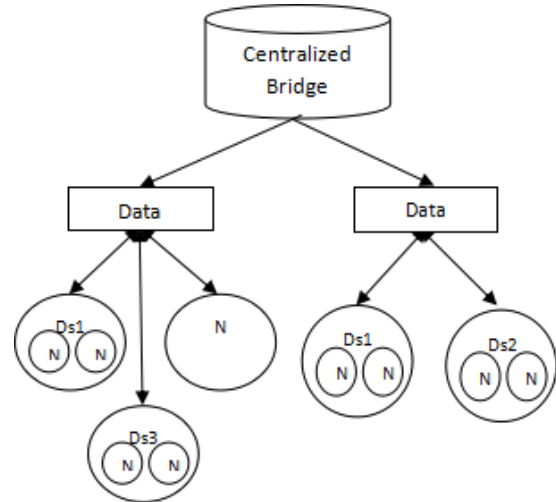
Fig. 1.   Centralized Clustering of Nodes.
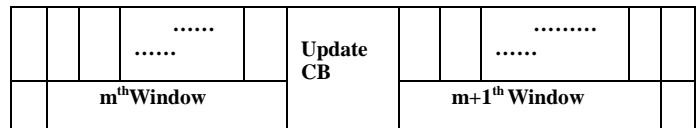
Fig. 2.   Monitoring the Transaction behavior of Nodes.
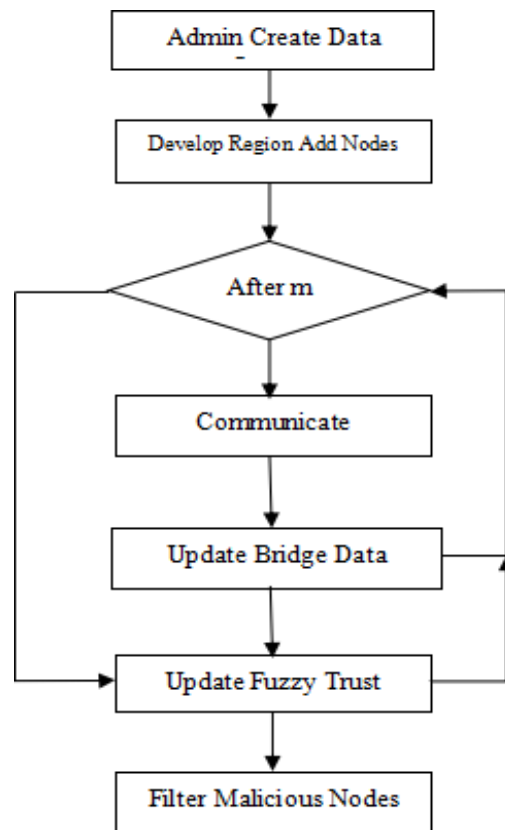
Fig. 3.   Flowchart of Malicious Node Detection.

B. Direct Trust

A node can generate its trust on other node from its previous set of transactions. If Node i request a service of node j then transaction happen between $i \rightarrow j$. Hence direct trust between them is evaluated by Eq. 1 and 2.

$$D_{ij} = \frac{S_t}{k} \tag{1}$$

$$S_t = \sum_i^k t_i \tag{2}$$

In equation 2 t_i is i_{th} transaction out of total k transaction happen between $i \rightarrow j$. t_i is 1 when transaction is successful otherwise its value is 0. D_{ij} is direct transaction value for $i \rightarrow j$. It means direct transaction value for $j \rightarrow I$ is different.

1) Centralized bridge: In this model all set of information related to node activities has been managed by this bridge [13]. Hence number of transactions between nodes and number of successful transactions S_t were managed by centralized bridge. This has trust value for the nodes as well. To evaluate trust of a node, proposed model uses Fuzzy Max Interval approach.

C. Fuzzy Max interval

Here a matrix of N x N was developed for the network. In this matrix each row is representing number of different combination of transactions occur between nodes of respected row, column [14]. Let that matrix is F whose dimension are NxN where each cell of M represents direct trust D_{ij}, value between nodes.

Eq. 3 generates max interval value.

$$I_{ik} = (Max(F_i) - F_{ik}) \quad k = \{1, 2, \ldots \ldots N \tag{3}$$

In equation 4 I_{ik} is the interval value of ith node for other set of k nodes. $Max(F_i)$ is the maximum value of F matrix for i^{th} row.

D. Trust Score

In this step one single value calculated corresponds to all set of fuzzy max interval, so this term is called as Trust score [15]. This is very simple as the step given in Eq. 3 of interval values is sum of trust score.

$$T_i = \frac{1}{\sum_{j=1}^k I_{ij}} \tag{4}$$

Hence by Eq. 4 trust value of nodes were evaluated by the centralized bridge.

E. Proposed Algorithm

Input: CB, m // S: Services

Output: T // Trust

1. DO←Create_DataOwner(CB)
2. N←Create_DataUser(DO)
3. R←Resources(DO) // R : Resources
4. CB←Hierarchal_cluster(N, R, CB)
5. Loop 1:m
6. i ←Random(N)

7. j ←Random(N)
8. if i and j belong to same Domain D_s
9. CB←Increase_Total_Transaction(CB, i, j)

   If t_{ij} is successful

   CB←Increase_Successful_Transaction(CB, i, j)

   EndIf
10. OtherwiseifTj>β
11. CB←Increase_Total_Transaction(CB, i, j)

    If t_{ij} is successful

    CB←Increase_Successful_Transaction(CB, i, j)

    EndIf
12. EndLoop
13. Loop i=1:N
14. Loop j=1:N
15. Dij←CB
16. Fi←Dij
17. EndLoop
18. $I_{ik} = (Max(F_i) - F_{ik})$
19. Ti←Trust_Score(Iik)
20. CB[i]←Ti
21. EndLoop

Attacks: Proposed algorithm was tested for two types of attacks. The first one is black hole attack. In this attack malicious node takes a service request and do not fulfill that request hence successful count of transaction get reduced. So detection of these kinds of malicious nodes was carried out by checking the trust value where trust values decrease with increase in number of unsuccessful transactions. Here direct trust value of nodes reduces to zero. Malicious node detection by only using direct trust value is not sufficient.

The second type of attack for which algorithm was tested was, Group Attack. In this attack if more than one malicious node mutually increases successful transaction than direct trust of this node is higher, but it's relation with other set of nodes will help to identify this malicious action. Let us consider six nodes, one to four out of six are real nodes and last two are malicious in nature. After *m* frame centralized bridge have direct trust values as shown in Table II.

TABLE II. NODE DIRECT TRUST VALUE MATRIX AT CENTRALIZED BRIDGE

Nodes	1	2	3	4	5	6
1	Null	2/2	2/3	¾	0/2	0/4
2	3/3	Null	2/2	2/3	0/4	0/1
3	2/2	¾	Null	2/3	0/3	0/4
4	¾	2/3	¾	Null	0/1	0/2
5	2/3	3/3	2/3	4/5	Null	5/5
6	2/3	1/1	2/3	Nil	4/4	Null

In Table II, each cell is direct trust value between existing nodes. So interval score of 1st node is Max (Column 1) is 1, hence.

	I_{11}	I_{12}	I_{14}	I_{15}	I_{16}
I =	0	0	0.25	0.33	0.33

In similar way Interval value of 5th malicious node where Max(Column 5) is 1.

	I_{51}	I_{52}	I_{53}	I_{54}	I_{56}
I =	1	1	1	1	0

Hence Trust score of 1st node is 1.098 while 5th node trust value is 0.25. Hence in this way proposed model detects malicious node.

IV. EXPERIMENT AND RESULTS

Implementation of service provider trust evaluation in un-reliable cloud environment was done on MATLAB platform. Here different number of virtual nodes with transaction were perform on this platform under windows operating system having 4GB RAM with I3 processor.

A. Evaluation Parameters

As various techniques involve different steps of working for classifying user query into appropriate category, it is highly required that proposed techniques or existing work should be compared on same experimental environment. This paper has used following parameters:

Malicious Node Convergence (MNC): This term is the ratio of number of malicious nodes to number transaction required to detect. Hence higher value of MNC ratio is better as it take low number of transaction for detection of malicious nodes.

$$MNC = \frac{Number of Malicious Node in Network}{Number of Transaction Need to Detect}$$

Trusted Node Convergence: This term is the ratio of number of Trusted nodes to number of transaction required to detect. Hence higher ratio value is better as it take low number of transaction.

$$TNC = \frac{Number of Trusted Node in Network}{Number of Transaction Need to Detect}$$

B. Results

Comparison was done on three environmental situations, first was Ideal (No attack), second was Black Hole Attack and third was Group Attack. Comparison was done with Domain Partition Trust Model (DPTM) method proposed in [8].

1) Ideal Condition: In this environment proposed model FMI-TM (Fuzzy Max Interval-Trust Model) performs better as shown in Fig. 4 that convergence of trusted node detection was done in less number of transaction as compared to previous approach DPTM in [8]. As proposed model takes approx. 950 transaction for trusted node detection and DPTM takes 6300 transaction for same situation.

Fig. 5 shows that proposed model has increased the trusted node trust value in less number of transactions and maintained the value for further transactions effectively. While previous approach DPTM has also increased the trusted node trust value but it take large number of transaction. This high increase in trust value was achieved by using the trust score obtained from the fuzzy max interval in FMI-TM.

Above Fig. 6 shows that proposed model FMI-TM has increase maintain the trust value of trusted score above 0.8 in all sets of transaction. This trust value was retaining by centralized storage as Fuzzy Max Interval value got updated regularly in each domain.

a) Black hole and group attack: In this environment, proposed model FMI-TM (Fuzzy Max Interval-Trust Model) performs better as Fig. 7 shows that convergence of malicious node detection was done in less number of transactions as compared to previous approach DPTM in [8]. As proposed model takes approx. 8260 transaction for trusted node detection and DPTM takes 11780 transaction for same. In case of group attack as malicious node increases the trust value of other malicious node, so DP-TM takes larger transaction, while proposed FMI-TM takes less number of transactions for detection. This reduction of transaction was achieved by fuzzy based trust score.
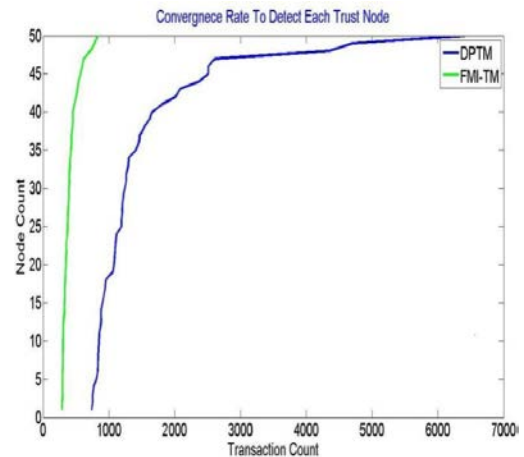
Fig. 4. Trusted Node Convergence based Comparison.
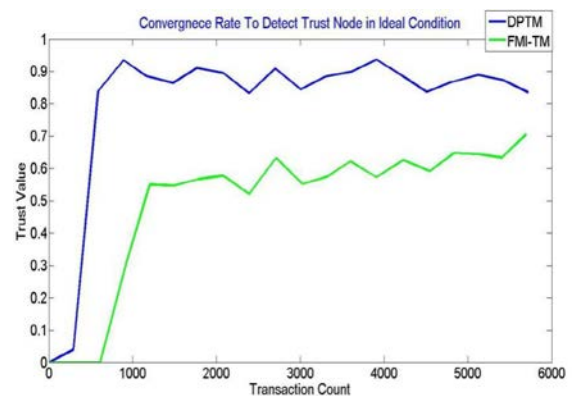
Fig. 5. Trusted Node Convergence based Comparison.
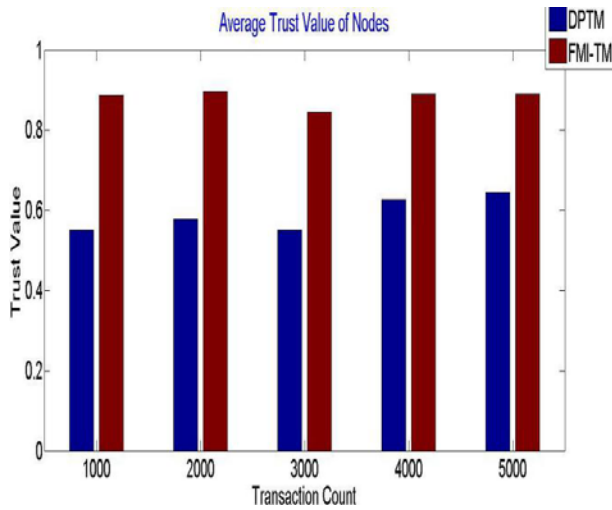
Fig. 6.   Average Trusted node Trust value at different Number of Transactions.
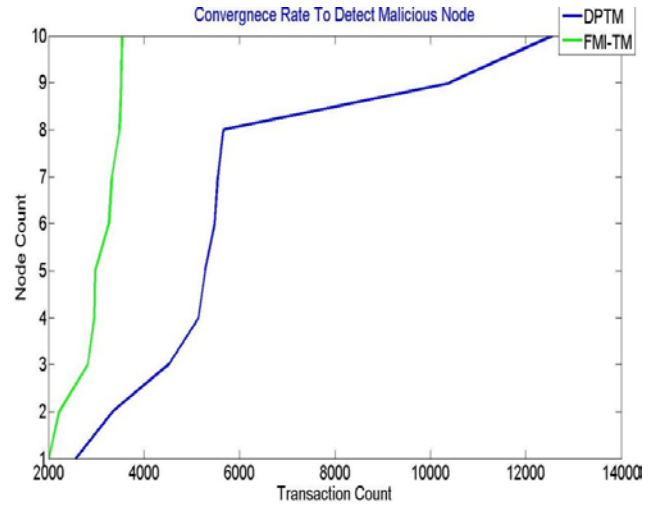
Fig. 8.   Malicious node Convergence based Comparison for Group Attack.
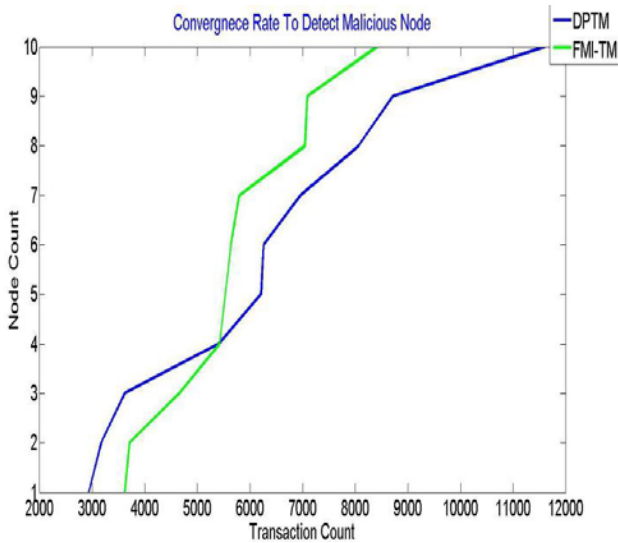
Fig. 7.   Malicious node Convergence based Comparison for Black Hole Attack.
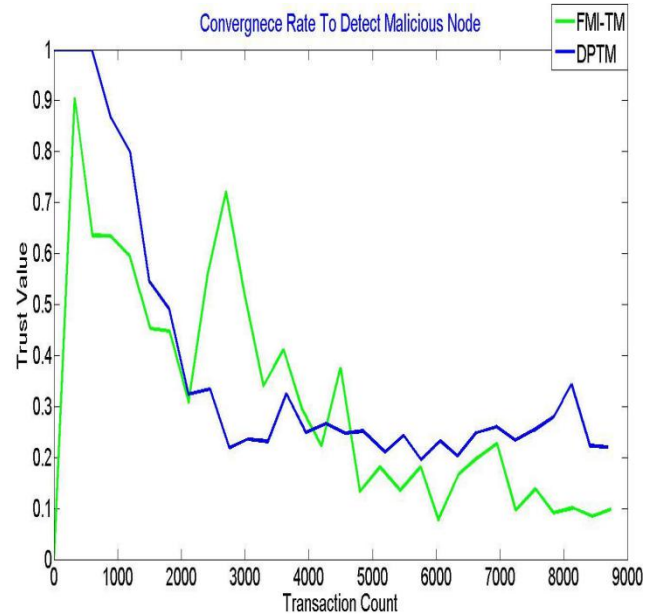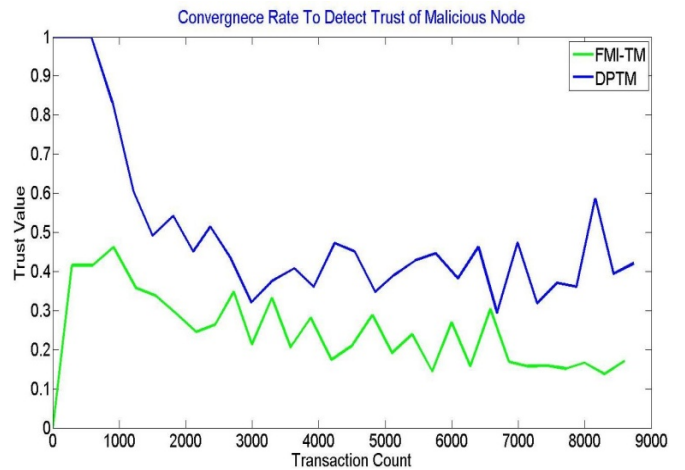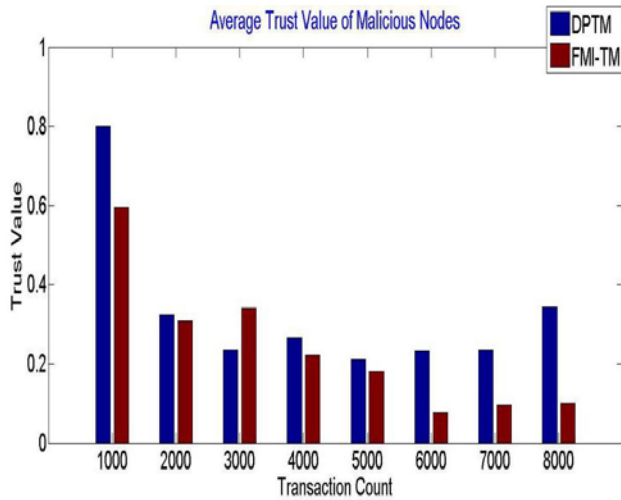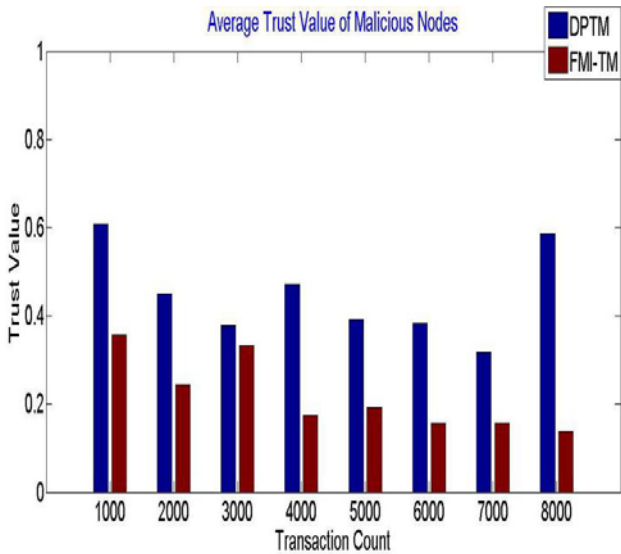

Fig. 9.   Malicious node Convergence based Comparison for Black hole Attack.

Fig. 8, 9 and 10 shows that proposed model has reduced the malicious node trust value in less number of transactions and maintained the value for further transactions effectively. While DPTM technique also decreases the trusted node's trust value but it involves large number of transactions. This high increase in trust value was achieved by using the trust score obtained from the fuzzy max interval in FMI-TM.

Fig. 11 and Fig. 12 shows that proposed model FMI-TM has reduced the average trust value of malicious score with increase in number of transactions. This pattern of decreasing trust value was obtained by the use of centralized storage and Fuzzy Max Interval value concept in proposed model.

Table III shows that proposed model has increased the convergence rate value of malicious and trusted nodes detection. It was obtained that under ideal condition convergence rate value was high while in attack environment this values decreases.

Fig. 10.   Malicious node Convergence based Comparison for Group Attack.

Fig. 11. Malicious node Convergence based Comparison for Black Hole Attack.

Fig. 12. Malicious node Convergence based Comparison for Group Attack.

TABLE III. CONVERGENCE RATE VALUE IN DIFFERENT CONDITIONS

Environment	DPTM	FMI-TM
Ideal	0.0078	0.0597
Black Hole	0.00086	0.0012
Group Attack	0.000796	0.0028

V. CONCLUSION

Cloud computing provide dynamic adaption of service sharing for company, individual, etc. So effective working of proposed model is highly required which directly depends on trust evaluation. This paper proposed a trust evaluation model by using Fuzzy Max Interval with multiple secure access for various role based access control users. Concept of centralized data updation was also enhances in this work as hierarchical cluster of nodes were developed, which was as per resource requirements for different data owners. Access of the resources was secured by involving AES encryption algorithm as well. Here model has increase the trust score of nodes with every successful transaction while its behavior with other

nodes also affect the score value. Proposed model has reduced the malicious node trust score by identifying its working pattern with other nodes. Experiment and analysis was carried out on three environmental conditions ideal, black hole and group attack. Results were compared with existing method and it was obtained that proposed model has improved the malicious node convergence value by 58.6% and Trusted Node convergence value by 86.93%. In future researcher can adopt genetic algorithm for clustering of nodes into real and malicious nodes.

REFERENCES

[1] Talal H Noor, Quan Z Sheng, Abdullah Alfazi, Jeriel Law and Anne HH Ngu, Identifying fake feedback for effective trust management in cloud environments in Service-Oriented Computing, pp.47-58(2013 b).

[2] Talal.H.Noor, Sheng, Q.Yao, L.,Dustdar, S. and Ngu, A.H.H, CloudArmor: Supporting Reputation-based Trust Management for Cloud Services, IEEE Transactions on Parallel and Distributed Systems,99(2014).

[3] WanitaS ,Surya Nepal and Cecile Paris ,A survey of trust in social networks in Journal of ACM Computing Survey ,45(4),pp.1- 33(2013).

[4] Sheikh MahbubHabib, Max Mühlhäuser, Sebastian Ries. "Towards a Trust Management System for Cloud Computing".Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on, At Changsha, China.

[5] M. Alhanahnah, P. Bertok, and Z. Tari, "Trusting cloud service providers: Trust phases and a taxonomy of trust factors," *IEEE Cloud Computing*, vol. 4, no. 1, pp. 44–54, Jan./Feb. 2017.

[6] Satish Kumar and Anita Ganpati, "Multi-Authentication for Cloud Security: A Framework," International Journal of Computer Science & Engineering Technology (IJCSET),Vol. 5, Issue 4, pp. 295-303, Apr. 2014.

[7] V. Sulochana and R. Parimelazhagan, "A Puzzle Based Authentication Scheme for Cloud Computing," International Journal of Computer Trends and Technology (IJCTT), Vol. 6, Issue 4, pp. 210-213, Dec. 2013.

[8] Peiyun Zhang, *Senior Member, IEEE*, Yang Kong, AndMengchu Zhou. "A Domain Partition-Based Trust Model For Unreliable Clouds". IEEE Transactions On Information Forensics And Security, VOL. 13, NO. 9, SEPTEMBER 2018.

[9] Azad M.A., Bag S., Hao F., Salah K. M2m-rep: Reputation system for machines in the internet of things. Comput. Secur. 2018;79:1–16. doi: 10.1016/j.cose.2018.07.014.

[10] Rafey S.E.A., Abdel-Hamid A., El-Nasr M.A. CBSTM-IoT: Context-based social trust model for the Internet of Things; Proceedings of the 2016 International Conference on Selected Topics in Mobile & Wireless Networking (MoWNeT); Cairo, Egypt. 11–13 April 2016; pp. 1–8.

[11] Chen Z., Ling R., Huang C.M., Zhu X. A scheme of access service recommendation for the Social Internet of Things. Int. J. Commun. Syst. 2016;29:694–706. doi: 10.1002/dac.2930.

[12] Yubiao Wang School of Big Data and Software Engineering, Chongqing University, Chongqing, China ; Junhao Wen ; Wei Zhou ; Bamei Tao ; Quanwang Wu ; Zhiyong Tao. "A Cloud Service Selection Method Based on Trust and User Preference Clustering" IEEE Access Volume 7, 12 August 2019.

[13] L. Minh Dang , Md. JalilPiran, Dongil Han, Kyungbok Min and Hyeonjoon Moon. "A Survey on Internet of Things and Cloud Computing for Healthcare". MDPI, journal/electronics 6 July 2019;

[14] Xiuqin Ma, Hongwu Qin, NorrozilaSulaiman, TututHerawan, and Jemal H. Abawajy . "The Parameter Reduction of the Interval-Valued Fuzzy Soft Sets and Its Related Algorithms". IEEE Transactions On Fuzzy Systems, Vol. 22, NO. 1, FEBRUARY 2014.

[15] Mahdi Ghafoorian, DariushAbbasinezhad-Mood, and Hassan Shakeri. "A Thorough Trust and Reputation Based RBAC Model for Secure Data Storage in the Cloud". IEEE Transactions On Parallel And Distributed Systems 2018.

Improving Palmprint based Biometric System Performance using Novel Multispectral Image Fusion Scheme

Essia Thamri[1], Kamel Aloui[2], Mohamed Saber Naceur[3]

LTSIRS Laboratory
National Engineering School Tunis
ENIT, Tunisia

Abstract—Nowadays, there are several identification systems which are based on different biometric modalities. In particular, multispectral images of palmprints captured in different spectral bands have a very distinctive biometric identifier. This paper proposes a novel fusion scheme of a biometric recognition system by multiSpectral palmprint. This system is composed of three blocks: (1) extraction of the region of interest (ROI) from multispectral images, (2) a new image fusion architecture based on the measurement of decorrelation, and (3) a scheme of dimension reduction and classification. The proposed image fusion system combines the information from the same left and right spectral band using the 2D discrete wavelet (DWT) transform technique. In addition, a feature extraction using the Log-Gabor transform is performed, while the feature size has been reduced using the Kernel Principal Component Analysis technique (KPCA). In Our experiments we use CASIA multispectral palmprint database. We obtained an accuracy rate (ACC) of 99.50% for the spectral bands WHT (white light) and 940 nm and an equal error rate EER = 0.05%.These results show that our system is robust against spoofing.

Keywords—*Biometric recognition; palmprint; multispectral images; image fusion; Log-Gabor; KPCA; DWT; CASIA*

I. INTRODUCTION

The person biometric recognition is the measurement of certain biological parameters or physical character specific to each individual. [1, 2, 3, 4].A large number of various biometric modalities identification systems, such as, iris [5], fingerprint [6, 7] and palmprint [8] are unique and allow to identify person with certainty. Counteracting insecurity and external threat, several techniques have been developed. In this context, biometric recognition by multi-spectral palmprint has emerged as a very attractive alternative [8, 9, 10, 11]. In addition, palmprints have the characteristics of stability, uniqueness and scalability, etc. [12]. This modality has a strong discrimination performance as well as an anti-noise capacity [13, 14]. As a result, these biometric systems have predicted their reliability and robustness against spoofing. This may be due to the fact that this modality presents several information such as, ridges, main and fine lines [15]. Furthermore, vein features that are hidden can be easily detected with infrared light [16, 17, 18]. Also, to acquire these visible features, it is necessary to take the images of the palms with several spectral bands. Recently, the multispectral palmprint biometrics have captured attention over the past 15 decades [6, 19, 20, 21, 22].

The palmprint multispectral imaging system uses cameras to represent spectral bands, where each band is represented by a single camera that can acquire palmprint images on different bands [23, 24]. This imaging system offers more advantages because the recording task is not compulsory seen that all the different bands images are procured at the same time. The biometric system is very sensitive to the quality of acquisition [25, 26]. Image acquisition can be done with or without contact. The system of recognition by the palmprint with contact is not acceptable by the users, since it is an environment of diseases transmission, therefore it is not recommended in the market. As well as the contactless biometric system is simple, comfort, less user contribution. So this system seems user-friendly and socially acceptable for person. As a result, it is highly recommended in the market. Despite the advantages of this non-contact biometric system, the acquired images contain several variations of scale between the images as well as the variation due to the translation and the rotation [27]. The multispectral palmprint algorithms focus on different multimodal fusion techniques: image-level fusion, at feature and at score-level. For image-level fusion, the goal is to obtain a single image by combining information from captured images for different spectral bands. For the score-level fusion technique, a feature extraction scheme of captured images at different spectral bands is used to have a comparison score that will be fused using a sum rule on which verification is performed. In literature, this rule has been frequently used, yet there is another alternative for this technique. Thus, the robustness of these multispectral biometric systems that use the image-level fusion approach generally depends on used fusion architecture. For the fusion technique, the use of feature extraction scheme is very interesting to obtain the performances of the biometric system. Several studies have shown that the fusion schemes of image-level are more efficient because these algorithms are robust to different types of noise due to lighting, motion blur, and so on [28,29]. However, the image fusion techniques have shown good extraction of details of the vein patterns and excellent reduction of noise in the image. Also, the technique of FPDCT and MRDCT could enhance the recognition rate [30]. In the work done by J.-G. Wang et al., a novel method of fusion by locality preservation projections (LPP) was developed [31]. This method improves the contrast, preserves the palm vein images and the edges. Also a new representation characteristic "Laplacian palm", which is extracted the features from the

images fused by the technique of locality preservation projections (LPP) is developed [32]. The fusion score level showed good performance compared to that with a single band. Xingpeng et al. have presented a biometric identification system depending on Quarter Principle Component Analysis (QPCA) [33]. This method is presented on different fusion schemes: the Curvelet transform, the morphological pyramid, the wavelet transform, and the gradient pyramid. The results found show that the fusion based on the Curvelet transformation presents more qualitative information. Another feature extraction and image fusion scheme has been developed [34]. This method combines information from the different spectral bands. Using discriminant kernel analysis (KDA) a dimension reduction was performed before classification using a classifier (SRC). A new approach to recognition of palm prints has been presented in [35].This approach is based on AE (an optimized auto-encoder) (AE), RELM (regularized extreme learning machine) and an extraction method (HOG-SGF). The experiments were carried out on three databases (CASIA, MS-PolyU, and Tongji of contact less visible palmprint images). The HOG-SGF approach gives better accuracy results than the HOG descriptor. Therefore, in this paper, a new fusion scheme of a multispectral Palmprint recognition system is developed. First, a new method for decomposing palmprint images of the same right and left spectral band on several spectral bands is proposed. This fusion scheme can more precisely combine information from the same Right and Left spectral band according to the measure of dependence. To obtain a texture characteristic in different sub-bands the discrete wavelets transform 2D (DWT) is carried out in order to acquire this multi-scale decomposition. Works have been developed on the fusion of images of palm prints acquired in two spectral bands for example (460 nm and 940 nm) [34]. On the other hand, our fusion method consists in applying 2D (DWT) on the same right and left spectral band. Our method shows excellent results compared to [34]. Secondly, a measure of the statistical dependence by the mutual information (MI) [36, 37] between the sub-bands has been developed. Finally, a fused palmprint image (left and right) is obtained by performing reverse DWT (IDWT) on fused wavelet coefficients. The accuracy rate

(ACC) of the proposed scheme was assessed using the multispectral CASIA database. In addition, experimental analysis of multispectral images of the palmprint has been performed to show the benefits of band fusion to enhance the recognition accuracy. The rest of the article is structured as follows: Section II presents the feature extraction method and the classification using three techniques: Transformation of Log-Gabor [34, 38] to effectively extract texture feature since our image is a texture image because there are repeating patterns, the Kernel Principal Component Analysis (KPCA) [40] to reduce dimensionality by minimizing redundancy [statistical dependence] and to evaluate the performance of our biometric system an Euclidean distance between the vectors is performed. Section III presents the different experimental results carried out with a comparison between works carried out and our proposed method. Finally, in section IV concludes this work.

II. THE PROPOSED MULTISPECTRAL PALMPRINT SYSTEM

Our developed system is composed of three steps: extraction of the region of interest (ROI), fusion of the (ROI) image followed by the extraction feature and the classification. Fig. 1 presents the proposed flowchart of our method of Multispectral palmprint fusion. Given the images of palm prints obtained are for different spectral bands. Doing the preprocessing of the images permit to extract the ROI firstly. Next, we present a novel multispectral image fusion scheme for the palmprint right and left of the same spectral band. So, that each multispectral image of the palmprint is broken down into two levels to get sub-bands using Haar wavelets. After that, we combine the sub-band information. To do that we use the mutual information MI to measure the dependence .Finally, we perform the feature extraction from the image fused by the Log-Gabor transform and classification to evaluate the system. By using the Log-Gabor we obtain a very high number of feature dimensions. Therefore, we carry out the reduction in dimensionality using the principal kernel component analysis technique (KPCA) before performing the classification using the Euclidean distance between the descriptors derived from (KPCA).
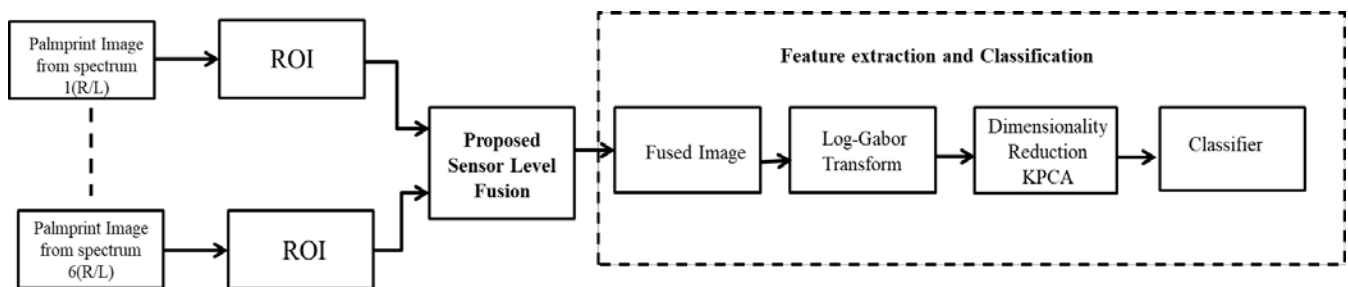

Fig. 1. Proposed Flowchart of our Method of Multispectral Palmprint Fusion.

A. Preprocessing and Palmprint Detection

In this section we will develop the extraction of the region of interest (ROI) and the multi-spectral palmprint image fusion right and left of the same spectral band.

1) (ROI) extraction: The database used in this work is the multispectral palmprint CASIA. All the images of this database are acquired using a contactless sensor. So, there is a large rotational variation, scaling and translation [34, 40, 41, 42]. These images are taken in six spectral bands 700 nm ,460 nm, WHT(White Light), 630 nm, 850 nm and 940 nm [39,43]. This database is very difficult for an accurate extraction of ROI. In this work, we used a ROI extraction scheme proposed by R. Rahavendra and C. Busch [34]. This scheme begins by carrying out the preprocessing phase of extract of the hand limit. The ROI is extracted with accuracy by detecting the benchmarks of the hand. Rahavendra and Busch detected two different landmarks R1 and R2 (see Fig. 2). To identify the binary discontinuities the scanning of the preprocessed image is involved for landmark detection .The choice of this method is based on its robustness since it allows, for different hand position, to estimate the translation and rotation corrections which are extremely hard to get when using contactless palmprint samples. Also this method gives a higher accuracy as indicated in Table I which compares the performance of the

ROI extraction for the two methods used by Rahavendra and Bush [34], and khan et al [44].

Fig. 3 and 4 show the ROI extraction results at 460 nm (hand images Right and Left).

TABLE I. PERFORMANCE OF THE ROI EXTRACTION SCHEME OF RAHAVENDRA AND BUSH

Methods	Accuracy (%)
Rahavendra and Busch [34]	95
Khan et al. [44]	90
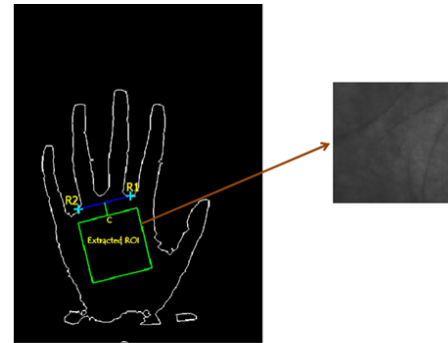
Fig. 2. ROI Extraction of Rahavendra and Busch.

| 001_1_460_01 | 001_1_460_02 | 001_1_460_03 | 001_1_460_04 | 001_1_460_05 | 001_1_460_06 |

| 001_1_460_01 | 001_1_460_02 | 001_1_460_03 | 001_1_460_04 | 001_1_460_05 | 001_1_460_06 |
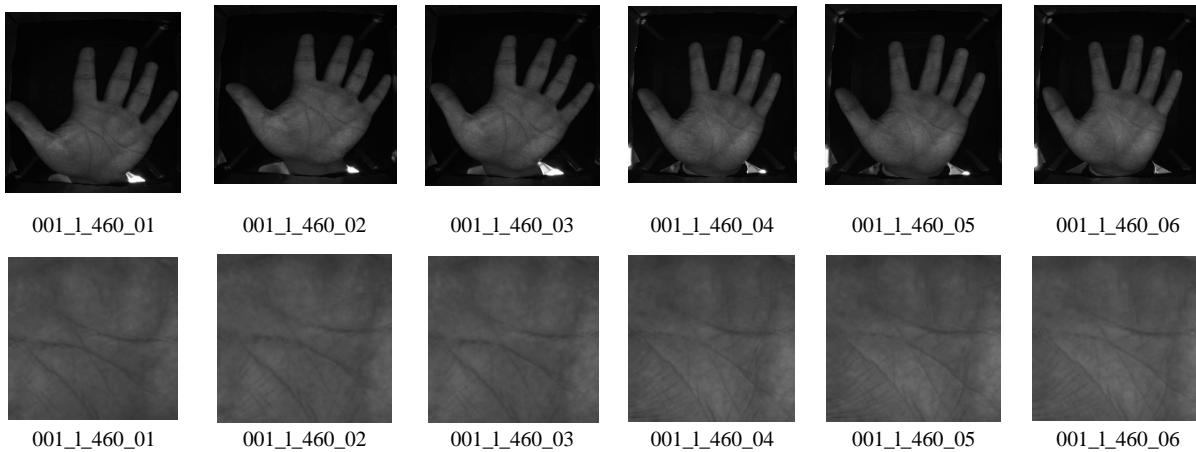
Fig. 3. ROI Extraction Results at 460 nm (Hand Images Left).

| 001_r_460_01 | 001_r_460_02 | 001_r_460_03 | 001_r_460_04 | 001_r_460_05 | 001_r_460_06 |

| 001_r_460_01 | 001_r_460_02 | 001_r_460_03 | 001_r_460_04 | 001_r_460_05 | 001_r_460_06 |
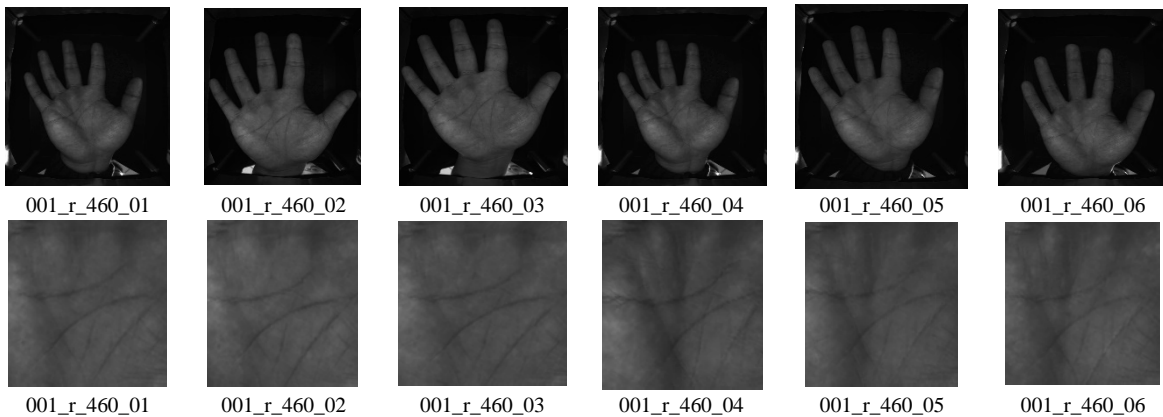
Fig. 4. ROI Extraction Results at 460 nm (Hand Images Right).

2) The proposed fusion scheme: The proposed scheme for multi-spectral palmprint image fusion right and left of the same spectral band is described in this section.This approach is based on three steps. Firstly, using 2D (DWT) with Haar mother wavelet we carry out the multi-scale decomposition of the palmprint images. The decomposition can be carried out by several techniques such as wavelet transformation [45], Laplacian pyramid [43], etc. We have chosen the discrete wavelet transform 2D (DWT) because it is easy to implement and it includes less computational requirements [33, 45, 46, 47]. As a result of this approach we find the wavelet coefficients at sub-bands. Secondly, these sub-bands obtained from palm-print images are combined. Thirdly, to get a fused palmprint left and right the inverse DWT (IDWT) image is performed. By considering the example of images get at two spectral bands right and left of the 460 nm band, we will explain the proposed image fusion and will show this approach can be extended for N different bands. The wavelet coefficients shown in Fig. 6 and which are represented in seven different sub-bands constitutes the result of palmprint image decomposition from the first spectral band using the 2-level DWT. From the first to the sixth sub-bands represent the approximation coefficients (A_1), the horizontal detail coefficients (H_1 and H_2), the vertical detail coefficients (V_1 and V_2) and the diagonal coefficients (D_1 and D_2) respectively.We apply the same procedure to the palmprint image from the Right band. As a result seven sub-bands: H_1' , A_1', H_2' , V_1', V_2', D_1' and D_2' . The degree of dependency between these sub-bands from the Left spectral band and the right spectral band (for example A_1 and A_1') is determined by utilising mutual information in the next step [36]. To quantify the dependence we have taken into account its ability to treat the non-Gaussian distribution and its robustness to the reduced size of the set of observations using mutual information. The mutual information between A_1 and A_1', are defined as follows:

$$M_1(A_1, A_1') = \left[log\left(\frac{f(A_1, A_1')}{f(A_1) f(A_1')} \right) \right] \quad (1)$$

With $A_1 \in S_1$ and $A_1', \in S_2$ the two sub-bands where S_1 represents the image of the left spectral band and S_2 represents the image of the right spectral band. $f(A_1, A_1')$ represents a joint distribution. In equation (1), the mutual information $M_1(A_1, A_1')$ indicates how much the information A_1 is transmitted to A_1'. The value of $M_1(A_1, A_1') = 0$ shows that A_1 and A_1' do not represent any common information. Similarly, when $M_1(A_1, A_1')$ has the highest value, this shows that the two sub-bands are dependent. We have estimated the density functions by the Parzen window with a Gaussian kernel [13].With this approach, we estimate well the distribution of sub-band wavelet coefficients [48].Using Eq. (1) we normalize mutual information values between 0 and 1 with sigmoidal normalization [49]. Then, a threshold (th) is compared with this value. When $M_1(A_1, A_1')$ value is greater than the threshold so it shows that A_1 and A_1' are linked to each other, so they must necessarily have the common information. Therefore, to perform the fusion we make the selection between A_1' and A_1 and. Then we choose between these sub-bands A_1 and A_1' which should be selected. Next, we calculate the energy of the wavelet coefficients in the sub-bands A_1' and A_1 distinctly. The one who has the highest energy is then selected. Instead, if the value $M_1(A_1, A_1')$ is equal or less than the threshold, then fusion between A_1 and A_1' is done, since the two sub-bands have complementary information. We represent the functional relation between the threshold and the mutual information as follows:

$$w_1 = \left(\frac{1 - M_1(A_1, A_1')}{1 + th} \right) \quad (2)$$

$$w_2 = 1 - w_1$$

The higher energy sub-band has the largest weights. We repeat this procedure on the others sub-bands. Then the (IDWT) is performed to obtain the fusion of multispectral palmprint image. Fig. 5 shows the proposed fusion scheme for six spectral bands. We extended this scheme for N bands.
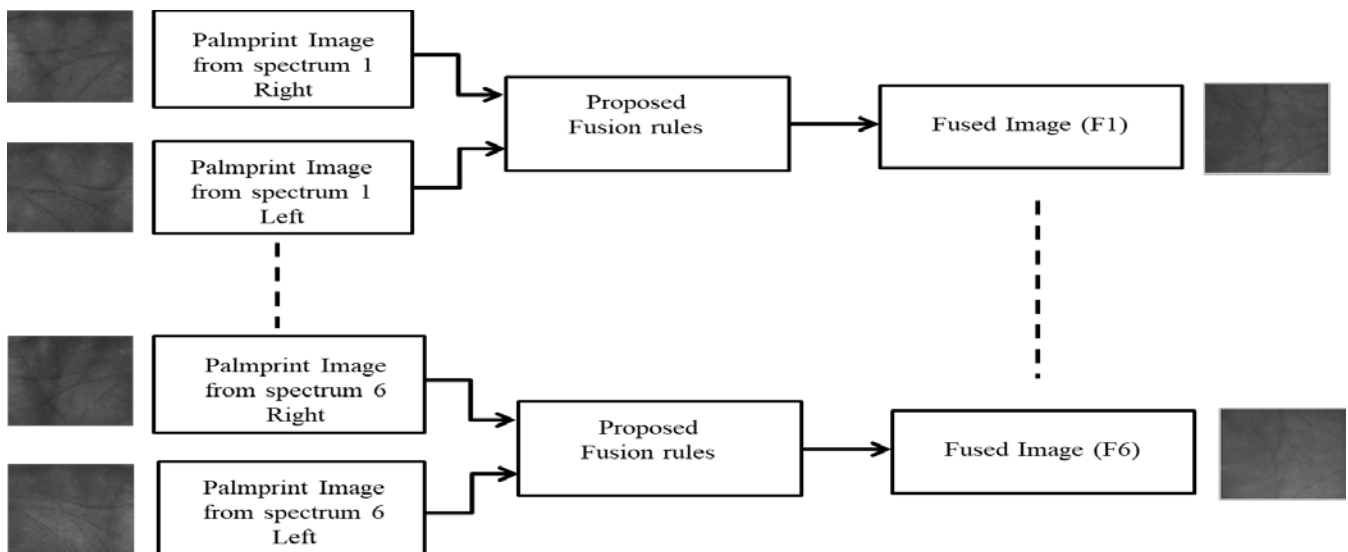

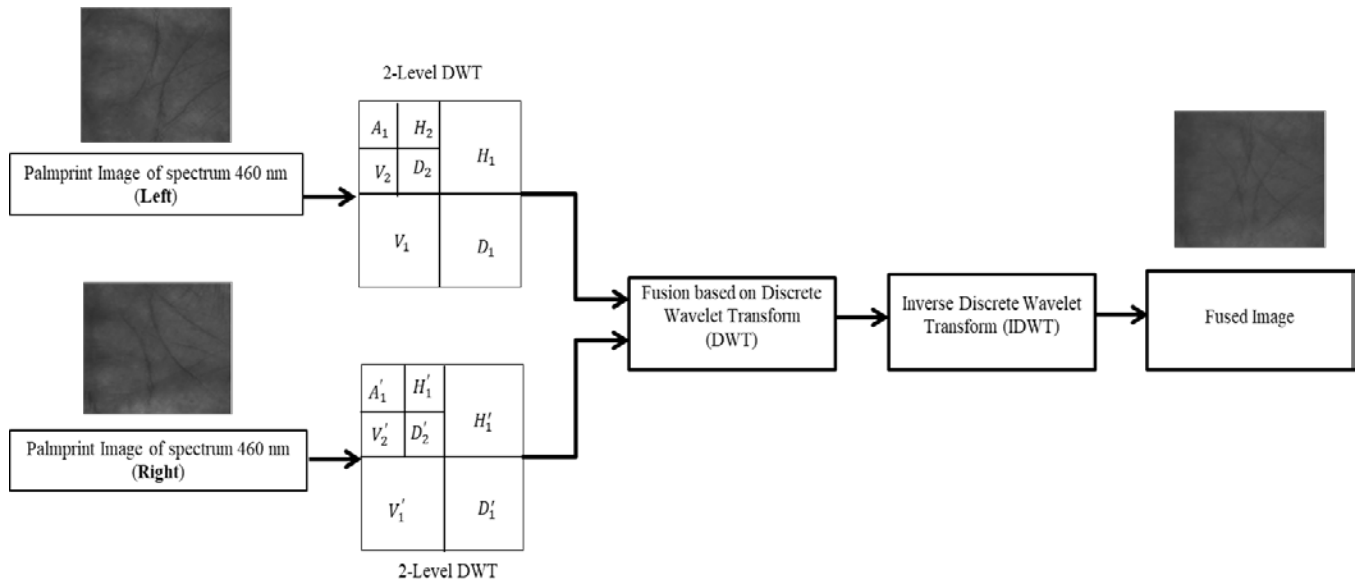Fig. 5. Illustration of the Fusion Scheme for the 6 Bands.

Fig. 6. Illustration of Fused Palmprint Image on Two Spectral Left and Right (460nm) Bands.

B. Feature Extraction and Classification

We get to the main line feature extraction step which is performed using the Log Gabor filter since the fused palmprint image is a texture image. The Log Gabor filter shows an improvement in performance in terms of precision [23] compared to the Gabor filter. In experiments The Log Gabor transform has eight orientations and four different scales. So, we have fixed these values after tests and in accordance with the literature [12]. Considering X_{fp}, the fused palmprint image whose size is 200×200. When we perform the Log Gabor transform X_{fp} we get 32 images. The X_{LGfp} image is the resulting Log Gabor transform image whose size is 420×840 [12]. We then performed a subsampling of the X_{LGfp} image in a ratio equal to 6 to reduce the computation load, and we obtained an important compromise among the computation reduction and the degradation of the performance. Thus, the size of X_{LGfp} is now reduced to 70×140. By concatenating the X_{LGfp} columns we obtain a X_{LGfp} descriptor vector with a size of 20,480 ×1. This procedure is repeated in all the remaining samples.

$$X_{GRef} = (20,480, N_{Ref}) \text{ and } X_{GTes} = (20,480, N_{Tes})$$

N_{Ref} and N_{Tes} are the numbers of reference and probes samples. Then we went to the reduction of dimensions of descriptor vectors to minimize redundancy before performing the classification. The Log Gabor transform each fused image gave us a high dimension of descriptor vector. However, for dimension reduction there are two types of data, non-linear and linear. The linear type has a difficult disadvantage of visually representing the object data in a graph and especially when using big data. In this work we used Kernel Principal Component Analysis (KPCA) as a reduction technique, which is a non-linear projection method and it is successfully used in biometrics [50, 51, 52]. This approach has proven its great capacity for reducing space and its good discriminating quality compared to the other approach. According to Kusban [54, 55],

a kernel can be founded from 'a multiplication matrix with its transposition which can be presented as follows:

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \tag{3}$$

With (x_i, x_j) : For the KPCA process

T: For transpose operation

$\Phi(x_i)$ has a very large value so we have to midpoint of centered features $\widehat{\Phi}(x_i)$. Therefore, equation (3) can be modified as follows:

$$k(x_i, x_j) = \widehat{\Phi}(x_i)^T \widehat{\Phi}(x_j) \tag{4}$$

The equation (4) can be written as :

$$k_c = k - 2lk + lkl \tag{5}$$

With l is [N×N] matrix with the value of 1/N for all elements. During our research on the novelty of the researcher authors, we found that the midpoint determination in the equation (4) is a critical value in KPCA [53]. So, we added several point values which can be presented as follows:

$$k(x_i, x_j) = [\widehat{\Phi}(x_i)^T \widehat{\Phi}(x_j)]^{\sqrt{2}} \tag{6}$$

We can also represent the new form of central nucleus in equation (5) as follows:

$$k_c = k - lk - kl + lkl \tag{7}$$

The k amount data are reduced without losing significant features. The dimension reduction process must be carried out with the diagonal value of Singular Value Decomposition (SVD). To properly characterize the data in the discriminant sub-space of the kernel, the efficiency of the KPCA is related to type of used kernel function. So, in this article, we used the polynomial kernel of degree 2 to characterize the Log-Gabor features. Therefore, the use of KPCA permit to reduce the dimension of X_{GRef} and X_{GTes} by $Xkpca_{Ref} = (838, N_{Ref})$ and $Xkpca_{Tes} = (838, N_{Tes})$.

***Algorithm** Feature extraction*

- *Input: Training Matrix: $TR = [TR_1, TR_2, ..., TR_c] \in R^{m \times n}$ for c classes.*
- *Probe simples: $y \in R^m$*
- *(1) Obtain the samples $TR' = [TR'_1, TR'_2, ..., TR'_c]$ through the Log-Gabor transformation.*
- *(2) Dimension reduction of TR' through KPCA to obtain a training matrix TR'_r.*

Once the step of the dimension reduction is finished, we passed to the step of the performance evaluation on by applying to Euclidean Distance-Based Matching between the vectors resulting from (KPCA).In our work we used 2/3 of data for the train and 1/3 for the test. In general, to evaluate the efficiency of the biometric system we are interested in two metrics of the performance variation FAR and FRR.

III. EXPERIMENTAL RESULTS AND DISCUSSION

We describe in this section the implemented protocol, the development environment, the data base representation and the results of our fusion method. The hardware and software tools used for the implementation of our solution are a personnel computer. Processor: Intel Core i5, CPU 2.4 GHz, Memory: 4 GB and Operating system: Windows 10 (32-bit). Also we use Matlab R2016b as simulation software.

A. Database Description

In this article, the database used is CASIA multispectral palmprint [43]. This database contains 200 hand images (100 right and 100 left). Each hand image has 6 samples and each sample contains six spectral bands (700 nm, 630 nm, 460 nm,940 nm, 850 nm and WHT (white light)),which give a total of 7200 multispectral palmprint images.

B. Motivation and Contributions

The first phase of experiments is the preprocessing phase which consists in detecting the regions of ROI interest of palmprints images from our CASIA -Multispectral database. Then, we performed a fusion at the descriptor level of the right and left image of the same spectral band of the six bands (700 nm, 630 nm, 460 nm,940 nm, 850 nm and WHT (white light)) by 2D discrete wavelet transform technique (DWT). The results of our fusion system showed excellent results compared to [34]. These results are shown in Table II. Then, we have to go to the main line features extraction phase using the Log-Gabor filter which allows to highlight the textures. Our image is an image of texture since it has repeating patterns. The next phase is the descriptor vector dimension reduction phase to minimize redundancy. The approach used for the reduction of dimension is the Kernel Principal Component Analysis technique (KPCA). To evaluate the performance of our biometric system, we applied the Euclidean-distance matching (EuM) technique between the vectors from KPCA.

C. Comparison and Discussion

In our research we have studied the recent methods of multispectral recognition of the palm print proposed in recent years. We have presented in Table II the different methods with their best performance rates for the two phases (verification and / or identification). We have analyzed and compared these rates. The methods studied are the methods of [34], [54], [55], [56] which use the multispectral palmprint CASIA database. The best performance results of these methods are found 95% in terms of Classification Accuracy (ACC) for identification and between [0.02; 3.12] % in terms of EER for the verification mode. The approach proposed in [34] allow to fused images for a multispectral palmprint system by combining information coming from different spectral bands using Kernel Discriminant analysis (KDA). This approach is based on log-Gabor filters for features detection. The classification is done using a sparse representation classifier (SRC). This method allows obtaining an EER rate of 1.64% (fusion of all bands). In other works, the method proposed in [55] shows a new modified radon transformation method as well as the kernel principal component analysis is performed. The experiments used two databases Poly U and CASIA palm print. The EER rate obtained is 0.02% in the CASIA database. Moreover, some methods [57] proposed a palmprint authentication system using Gabor wavelets for extraction coefficients, wavelet decomposition for palm image fusion, and colony optimization system. They use SVM to classify the samples. This approach permit to obtain an EERs of 3.97% and 3.12% on the evaluation sets. Moreover, a bidirectional matching technique has been used in place of a unidirectional matching in [54]. The algorithm used is founded on the strategy a feature-level fusion and sampling. EER rates for this method are 0.16% and 0.73% for two databases. In [56], the method seeks to represent a biometric identification system utilizing PCANet^{S_2} deep learning. This method is done by the 4 classifiers (RFT, SVM, KNN and RBF). The best ERR rate produced by the spectral band (460 nm) is 0.12%.

Evaluating the performance of our system constitute the goal of the last part of this work. After the fusion of Left and Right spectral bands of the same band, a reduction of dimension is carried out by the KPCA. Then, for all the spectral bands, a Euclidean distance between the vectors issuing from KPCA is applied with the best parameters. Table III presents the results for each spectral band, permitting to see the performances of our proposed system. This table shows that the EER reaches the highest value ERR= 0.05% for to the two spectral bands 940 nm and WHT. Also, EER = 1.00% is achieved by the spectral bands 630 nm, 850 nm, 460 nm and an EER = 1.50% for the 700 nm spectral bands. The efficiency of all the spectral bands is illustrated in Fig. 9. This figure shows the Receiver Operating Characteristic (ROC) curves for the different spectral bands .A biometric system is said to be efficient if it has a small EER value. In this table, the best accuracy rate (ACC) in the case of two spectral bands 940nm and WHT = 99.50%. For the 850 nm, 460 nm and 630 nm spectral bands the ACC= 99.00% and the 700 nm band at ACC = 98.50%. Therefore the results show that our Biometric system has reached performance and approached the robustness part comparable to the best of advanced systems like those of [34.56]. Finally, in order to obtain the performance evaluation, an Euclidean distance is applied to each band between the vectors resulting from KPCA. Then the performance of the identification system is evaluated using classifier information.

Table IV shows the accuracy variation for different threshold values (th) for the 940 nm (Right and Left) spectral band. The value of (th) varies from 0 to 1. Fig. 7 shows the illustrating of the Accuracy of the spectral band 940nm for different (th) threshold values. We note that the best Accuracy value is obtained for the optimal decision threshold (0.5). Fig. 8 presents the proposed fusion scheme for six spectral bands. We extend this scheme for N bands.

Taking into account these accuracies, we think that our fusion approach to the level of the left and right image of the same spectral band (technique (DWT) + Kernel PCA) can be sufficient to represent the majority of biometric features within the biometric modality.

TABLE II.   COMPARISON OF THE PERFORMANCE OF OUR METHOD WITH THE STATE OF THE ART.THE RESULTS ARE GIVEN IN TERMS OF EER (EQUAL ERROR RATE) AND CLASSIFICATION ACCURACY (ACC)

Method	Feature extraction	Database	Best recognition performance
Raghavendra et al. (2014) [34]	Log-Gabor filter+KDA	CASIA	Verification: EER = 1.640%(fusion of all bands) Identification: ACC = 95%
Kisku et al. (2012) [55]	Gabor wavelet transform	CASIA	Verification: EER = 0.0248% Identification: /
Kisku et al. (2010) [57]	Wavelet Fusion with ACO	CASIA	Verification: EER = 3.125% Identification: /
Yan et al. (2015) [54]	RootSIFTbased feature fusion	CASIA	Verification: EER = 0.160% Identification: /
Meraoumia et al. (2017) [56]	Two-stage PCANet^{S_2}	CASIA	Verification: EER=0.125%(best 640nm) Identification: /
Proposed Method	Log-Gabor filter+KPCA	CASIA	Verification: EER = 0.05% (best 940 nm and WHT) Identification: 99.5% (best 940 nm and WHT)

TABLE III.   IDENTIFICATION SYSTEM PERFORMANCE OF FUSED BANDS (CASIA DATABASE)

Spectral fused Bands(nm)	EER (%)	Accuracy (%)
630	1.00	99.00
700	1.50	98.50
850	1.00	99.00
460	1.00	99.00
940	0.05	99.50
WHT	0.05	99.50

TABLE IV.   ACCURACY VARIATION FOR DIFFERENT THRESHOLD VALUES (TH) FOR THE 940 NM (RIGHT AND LEFT) SPECTRAL BAND

Threshold (th) values	Accuracy (%)
	Proposed Fusion scheme
0	99.30
0.1	99.19
0.2	99.29
0.3	99.17
0.4	99.49
0.5	**99.50**
0.6	99.41
0.7	99.21
0.8	99.05
0.9	99.04
1	99.21
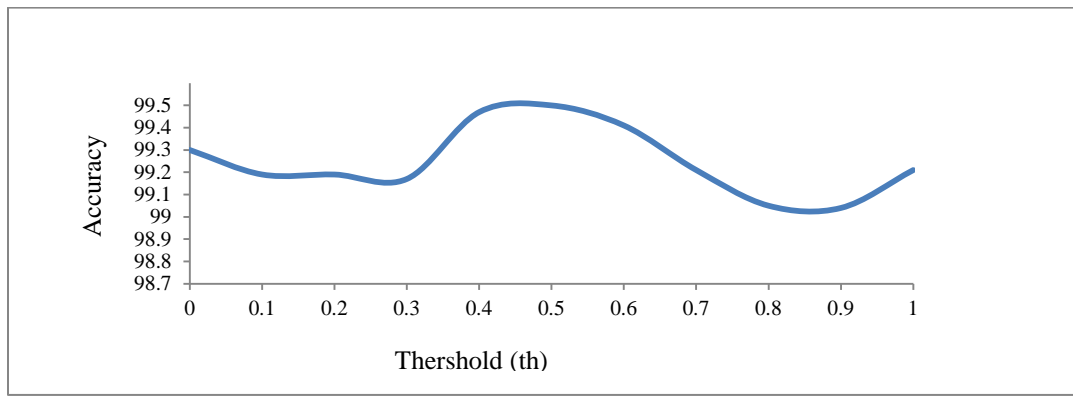
Fig. 7. Illustrating of the Accuracy of the Spectral Band 940nm (Right and Left) for different (th) threshold Values.
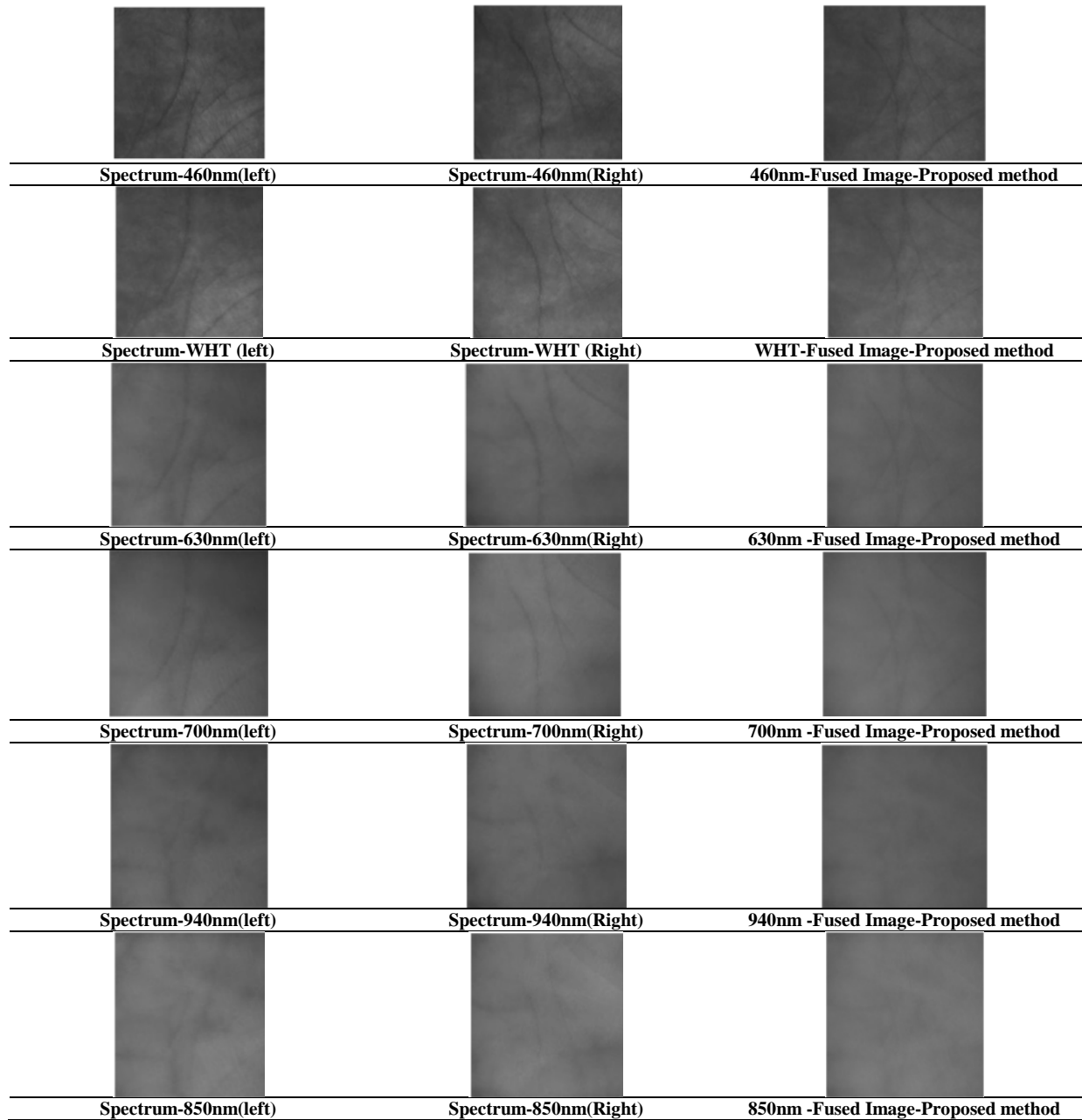
Fig. 8. Results of the Image Fusion Scheme Proposed on the Left and Right Spectral Bands of 6 Bands (700 nm, 460 nm, WHT, 630 nm, 850 nm and 940 nm).
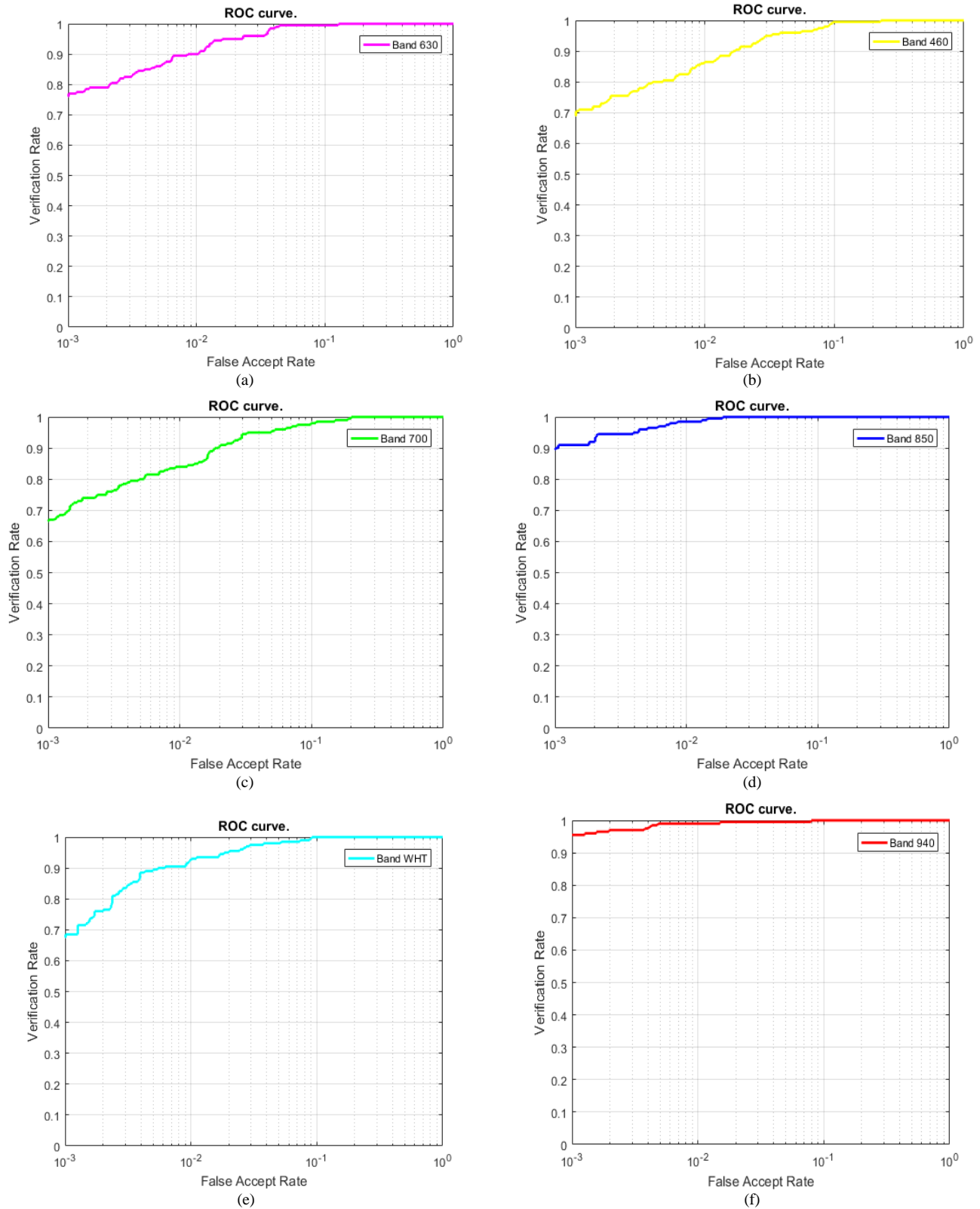
Fig. 9. ROC Curve for the different Spectral Bands: (a) 630nm Band, (b) 460nm Band, (c) 700nm Band, (d) 850nm Band, (e) WHT Band, (f) 940nm Band.

V. CONCLUSION

In this article, a new palmprint biometric system based on the fusion of multispectral palm images at left hand and right hand level was proposed. This approach is performed by the 2D (DWT) technique during the preprocessing phase to keep more information. Multi-spectral palm images are fused at low level using wavelet transform and decomposition. We used in the main line feature extraction step the Log-Gabor filter which corresponds better to natural images than to other filters to highlight textures. We proposed the technique (KPCA) for the step of reduction of dimension to minimize redundancy. This method is capable of adapting effectively to the variation of rotation, to the potential deformation of the image and to the translation variation. We use the multispectral CASIA database, in which the images are captured with a contactless sensor. Evaluating the performance of our biometric system by applying a Euclidean distance between the vectors from KPCA was considered as the last step. We obtained an ACC rate of 99.50% for the spectral bands WHT (white light) and 940 nm. Also, the ACC is equal to 99.00% for the spectral bands 630 nm, 850nm, 460 nm and 98.50% for the band 700nm. The EER rate achieved by our method is 0.05% for the bands (940nm, WHT) and 1.00% for the other spectral bands. These results show that our system is robust against spoofing. Our future work will focus on new classification techniques that are more effective in dealing with shadow and unwanted noise. Deep learning architectures like convolutional neural networks (CNNs) need to be studied.

REFERENCES

[1] X. Yan , W. Kang,"Palm vein recognition based on multi-sampling and feature-level fusion," Neurocomputing, March (2015),P.798–807 Vol. 151, Part 2, 5.

[2] Z.Khan,A.Mian,Y.Hu, "Contour code: robust and efficient multispectral palmprint encoding for human recognition",in:IEEE International Conference on ComputerVision,(2011),pp.1935–1942.

[3] D.Hong,W.Liu,J.Su,Z.Pan,and G.Wang, "A novel hierarchical approach for multispectral palmprint recognition",in Neurocomputing,vol 151,part1,(2015),pp.511-521.

[4] A.K.Jain P.Flynn, and A.A.Ross (Eds.), Handbook of biometrics, Springer, USA (2008).

[5] S.J.Daugmanand, The importance of being random:statistic principales of iris recognition ,Pattern Recognit.36(2),(2003),pp. 279-291.

[6] D. Zhang,L. Zhang ,and G.Zhenhua , "Feature Band Selection for Online Multispectral Palmprint Recognition," in International Publishing Multispectarl Biometric Springer, (2015), pp.153-162.

[7] W.Song,T.Kim, H.C.kim,J.H.Choi, H.-j.Kong,S.-R.Lee, "Afinger-vein verfication system using mean curvature," in Pattern Recognition Letters 32 ,(2011),pp.1541-1547.

[8] X .Guo, W. Zhou and Y.Zhang,Collaborative representation with hm-lbp features for palmprint recognition. Mach Vis Appl 28(3-4): (2017),pp. 283–291

[9] M. Kusban and A, Budiman, "An excellent system in palmprint recognition," in International Conference on Engineering and Applied Technology (ICEAT),(2018),pp.1088-1757.

[10] J. Lu,K.N.Plataniotis A.N.Venetsanopoulos , "Face recognition using kernel direct discriminant an alysis algorithms",in IEEE Trans.Neural Netw.14 (1), (2003),pp.117–126.

[11] D.Hong,W.Liu,J.Su,Z.Pan,and G.Wang, "A novel hierarchical approach for multispectral palmprint recognition",in Neurocomputing,vol 151,part1,(2015),pp.511-521.

[12] R. Raghavendra, B. Dorizzi, A. Rao and G.H. Kumar , "Designing efficient fusion schemes for multimodal biometric system using face and palmprint" ,in Pattern Recognit.44(5) (2011),pp.1076-1088.

[13] M.Christopher Bishop, M. Nasser,and Nasrabadi,Pattern recognition and machine learning. Vol. 1, Springer,(2006) ,New York.

[14] J.Cui, "2D and 3D Palmprint fusion and recognition using PCA plus TPTSR method".in *Neural Comput.Appl. 24*, (2014),pp.497–502.

[15] D.Cheng, X.Zhang,, "An Improved Recognition Approach for Noisy Multispectral Palmprint by Robust L2 Sparse Representation with a Tensor-Based Extreme Learning Machine," in *Sensors* (2019), *19*, 235.

[16] R. Kabaciński and M. Kowalski. " Vein pattern database and benchmark results".In Electronics Letters, (2011),ISSN:pp.1127-1128, vol 47.

[17] A.Al-juboori,1,2 W.Bu, Xi. Wu, and Q.Zhao, "Palm Vein Verification Using Multiple Features and Locality Preaerving Projections", in The Scientific Word Journal Issues, February (2013),Article ID 24603.

[18] A. Al-juboori and W. Bu, "Palm Vein Verification Using Gabor Filter", in ,IJCSI International Journal of Computer Science Issues, January (2013),Vol. 10, Issue 1, No 1.

[19] P.Dubey, T.Kanumuri, "Palmprint Recognition Using Oriented Structural Energy Signature Codes", in Arabian Journal for Science and Engineering. (2019).

[20] M. Tabejamaat, A. Mousavi, Concavity-orientation coding for palmprint recognition, Multimed. Tools Appl., 76 (2017) 9387-940.

[21] L. Fei, G. Lu, Feature extraction methods for palmprint recognition: a survey and evaluation", in IEEE Trans Syst Man Cybern Syst:1–18, (2018).

[22] L. Fei, B. Zhang, Y.Xu,and L Yan, "Palmprint recognition using neighboring direction indicator".in IEEE Trans. Hum. Mach. Syst.46(6), (2016),pp.787–798.

[23] N. Rose, "Facial expression classification using Gabor and Log-Gabor filters", in: 7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006, (2006), pp. 346–350.

[24] Ross A. Abraham, Karthik Nandakumar, "Anil K. Jain, Handbook of multi-biometrics", (2006), vol. 6, Springer, USA.

[25] C. Varon, C. Alzate, and J. A. K. Suykens, "Noise Level Estimation for Model Selection in Kernel PCA Denoising", in IEEE Trans. Neural Networks Learn. Syst., vol. 26, no. 11, (2015), pp. 26502663.

[26] L. Shen, L. Bai,M. Fairhurst, "Gabor wavelets and general discriminant analysis for face identification and verification", Image Vis. Comput. 25 (2007),pp. 552–563.

[27] Y. Hao, Z. Sun, T. Tan and C. Ren, "Multispectral palm image fusion for accurate contact-free palmprint recognition", in: 15th IEEE International Conference on Image Processing, 2008, ICIP 2008, (Oct. 2008), pp. 281–284.

[28] S.Manikanda prab, S.N.Sivanandam, "A Novel Biometric system for Person Recognition Using Palm vein Images" ,in International Journal on Computer Science and Engineering (IJCSE) ISSN,(2013),pp.0975-3397 Vol. 5 No.

[29] A. Al-juboori and W. Bu, "Palm Vein Verification Using Gabor Filter", in ,IJCSI International Journal of Computer Science Issues, January (2013),Vol. 10, Issue 1, No 1.

[30] S. C. Soh, M. Z. Ibrahim, "Image Fusion based Multi Resolution and Frequency Partition Discrete Cosine Transform for Palm Vein Recognition".in IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA),(2019).

[31] J.-G. Wang, W.-Y. Yau, A. Suwandy, E. Sung, "Person recognition by fusing palmprint and palm vein images based on "laplacianpalm" representation",in Pattern Recognit. 41 (May (5)) (2008),pp.1531–1544.

[32] D. Zhang, Z. Guo, G. Lu, L. Zhang and W. Zuo, "An online system of multispectral palmprint verification", in IEEE Trans. Instrum. Meas. 59 (February (2)) (2010),pp.480–490.

[33] X. Xingpeng, Z. Guo, C. Song and Y. Li, "Multispectral Palm-print Recognition Using a Quaternion Matrix," in Sensors,(2012),pp. 4633–4647.

[34] R .Raghavendra, C. Busch, "Novel image fusion scheme based on dependency measure for robust multispectral palmprint recognition", in Pattern Recognition, 47(6), (2014),pp.2205–2221.

[35] A.Gumaei, R. Sammouda, "An Effective Palmprint Recognition Approach for Visible and Multispectral Sensor Images," in Sensors, 18(5), (2018),1575.

[36] M.T. Cover, A.T. Joy, "Elements of Information Theory", John Wiley & Sons, USA, (2012).

[37] K. Huang, S. Aviyente, "Wavelet feature selection for image classification,"in IEEE Trans. Image Process. 17 (9) (2008),pp.1709–1720.

[38] X. Zhitao, G. Chengming, Y. Ming, L. Qiang, "Research on Log Gabor wavelet and its application in image edge detection, " in: Proceedings of 6th International Conference on Signal Processing (ICSP-2002), USA, (2002), pp. 592–595.

[39] M.Kusban,A.Susanto. "Combination a Skeleton Filter and Reduction Dimension of KernelPCA-Based on Palmprint Recognition", in International Journal of Electrical and Computer Engineering (IJECE)Vol. 6, No. 6, December (2016), pp. 3255 – 326.

[40] L. Zhang, L. Li, A. Yang, Y. Shen and M. Yang, "Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach," in Pattern Recognition, 69 (1) (2017),pp.199-212.

[41] Y. Luo, L. Zhao, B. Zhang, W. Jia, F. Xue, J. Lu, Y. Zhu and B. Xu, "Local line directional pattern for palmprint recognition," in Pattern Recognition, 50(2016),pp.26-44.

[42] Y. Hao, Z. Sun, T. Tan, C. Ren, "Multispectral palm image fusion for accurate contact-free palmprint recognition," in: 15th IEEE International Conference on Image Processing, 2008, ICIP 2008, (Oct. 2008), pp. 281–284.

[43] H. Li, B. Manjunath and S. Mitra, "Multisensor image fusion using the wavelet transform," in Gr. Models Image Process. 57 (3) (1995),pp235–245.

[44] Z. Khan, A. Mian and Y. Hu, "Contour code: robust and efficient multispectral palmprint encoding for human recognition," in: IEEE International Conference on Computer Vision, (2011), pp. 1935–1942.

[45] K. Amolins, Y. Zhang and P. Dare, "Wavelet based image fusion techniques—an introduction review and comparison," in ISPRS J. Photogramm. Remote Sens. 62 (4) (2007),pp. 249–263.

[46] R. Raghavendra, B. Dorizzi, A. Rao and G.H. Kumar, "Particle swarm optimization based fusion of near infrared and visible images for improved face verification," in Pattern Recognit. 44 (2) (2011),pp.401–411.

[47] D. Han, Z. Guo and D. Zhang, "Multispectral palmprint recognition using wavelet-based image fusion," in: International Conference on Asian Conference on Computer Vision,(2008), pp. 2074–2077.

[48] K. Huang and S. Aviyente, "Wavelet feature selection for image classification," in IEEE Trans. Image Process. 17 (9) (2008),pp.1709–1720.

[49] O.Richard, Duda, E.H.Peter and G.S.David, Pattern classification, John Wiley & Sons, USA (2012).

[50] N. Japkowicz and M. Shah, "Evaluating Learning Algorithms: A Classification Perspective," in Cambridge University Press, New York, NY, USA,(2011).

[51] M. Kusban.A, Budiman,"An excellent system in palmprint recognition," in International Conference on Engineering and Applied Technology(ICEAT),(2018),pp.1088-1757.

[52] M. Kusban and A. Susanto. Combination a Skeleton Filter and Reduction Dimension of KernelPCA-Based on Palmprint Recognition," in International Journal of Electrical and Computer Engineering (IJECE),(2016),pp.3255-3261.

[53] C. Varon, C. Alzate, and J. A. K. Suykens, "Noise Level Estimation for Model Selection in Kernel PCA Denoising," in IEEE Trans. Neural Networks Learn. Syst., vol. 26, no. 11, (2015),pp. 2650-2663.

[54] X. Yan,W. Kang, F. Deng and Q.Wu, "Palm vein recognition based on multi-sampling and feature level fusion," in Neurocomputing 151(Part 2), (2015),pp.798–807.

[55] D.R. Kisku, A. Rattani, P.Gupta, J.K. Sing, and C.J. Hwang, "Human Identity Verification Using Multispectral Palmprint Fusion," in Journal of Signal and Information Processing, vol. 3, (2012),pp. 263–273.

[56] A.Meraoumia and F. Kadri,"Improving Biometric Identification Performance Using PCANet Deep Learning and Multispectral Palmprint," in Biometric Security and Privacy - Opportunities & Challenges in The Big Data Era,(2016),pp.51-69.

[57] D.R. Kisku, P. Gupta, J.K. Sing and C. Hwang, "Multispectral palm image fusion for person authentication using ant colony optimization," in Proceedings of International Workshop on Emerging Techniques and Challenges for Hand-Based Biometrics (ETCHB) (IEEE, New York, (2010)), pp. 1–7.

A Review on Research Challenges, Limitations and Practical Solutions for Underwater Wireless Power Transfer

Syed Agha Hassnain Mohsan[1]
Optical Communication Laboratory
Ocean College, Zhejiang University
Zhoushan, China

Asad Islam[2]
School of Energy and Power
Engineering
Beihang University
Beijing, China

Mushtaq Ali Khan[3]
Department of Marine Science
Ocean College, Zhejiang University
Zhoushan, China

Arfan Mahmood[4]
Complex Networks and Control Lab
Shanghai Jiao Tong University
Shanghai, China

Laraba Selsabil Rokia[5]
Department of Biomaterials
East China University of Science and
Technology
Shanghai, China

Alireza Mazinani[6]
School of Electronic and Information
Engineering
Beihang University
Beijing, China

Hussain Amjad[7]
Department of Marine Science and Information Technology
Ocean College, Zhejiang University
Zhoushan, China

Abstract—**Wireless power transmission is the process to transmit electrical energy without using wire or any physical link. WPT is mainly used at such places where it is difficult to transfer energy using conventional wires. In this research work, we investigated the need and feasibility of wireless power transmission for underwater applications. This research paper will outline research challenges, limitations and practical consideration for underwater wireless power transfer (UWPT). Recent researchers have focused on WPT in air. However, WPT is still a challenging task in underwater environment. In this study, we have presented a review on previous research works in underwater wireless power transfer (UWPT). We have provided a comparison of different techniques implemented for wireless power transfer. This paper also proposes the idea of MIMO wireless power transmission for AUVs charging. This paper elaborates capabilities and limitations of the wireless power transfer system in underwater media as stochastic nature of ocean is a big challenge in wireless power transmission. We have also addressed design challenges and seawater effects on UWPT system.**

Keywords—*Underwater wireless power transfer; charging; MIMO; AUV*

I. INTRODUCTION

The ongoing research and fast-expanding market of electrical vehicles has reinvigorated the wireless power transfer (WPT) technology. A crucial need of WPT has emerged lately in distributed ocean systems which require excessive power such as ocean monitoring devices, sensor nodes, autonomous underwater vehicle (AUVs). These underwater devices are being used in innumerous underwater applications such as data collection, monitoring and underwater observation. Usually these electronic devices and sensors are deployed in AUVs and ROVs. However, AUVs are advantageous as they do not require any support vessel for ongoing operations. The main disadvantages are its mission time and range which severely limited because of low battery endurance. The typical process to recharge these electronic devices is time-consuming which results into disrupted service and limited operation range. To bridge these gaps, an extensive research work on AUVs is required. Despite of possibility of WPT system for AUVs, still it is bulky in structure which makes it inadequate for smaller AUVs. Nonetheless, overcoming these challenges can unlock a new era of research and new possibilities, offering AUVs to extend operation range and lifetime. A possible solution is to design an underwater docking station. The practical consideration behind this improvement is that the deployed WPT system is also autonomous. In fact, such system has high maintenance periodicity. It also requires high cost for maintenance and useful for near shore applications only. These implementations can be used for specific missions such as SARS operations, oceanography studies, weather reporting and monitoring of water parameters. Here, WPT emerges as a potent solution. Recent scholars are mainly focusing on terrestrial based application of WPT. Several effective techniques have been carried out to maximize WPT system efficiency such as frequency tuning, intermediate field repeaters and impedance matching. In [1], Niu et al. have used metamaterial to enhance WPT system efficiency in air medium. However, WPT technology in underwater

environment, where seawater is a challenging medium, is still looking for interest from research fraternity. We have summarized various challenges for underwater wireless power transfer (UWPT) below.

- What is the effect of high conducting water medium on electrical parameters of WPT system?

- How coil radiation resistance is effected by seawater?

- What are the main losses incurred and how to minimize these losses?

- If any loss is highly dependent of frequency then how to select an operating frequency to achieve an efficient wireless power transmission?

Besides these challenging factors, mobility occurs in WPT system due to dynamic nature of seawater which leads to a varying coupling coefficient. Thus, an optimized design is required to curb this possible variability. Another influential factor to surpass due to high losses in undersea and high electrical permittivity. These challenges pose austere limitations to WPT system. Permittivity and permeability [2-3] parameters for underwater medium are different from air medium. Moreover, the controlling techniques involved in WPT system in air medium require a communication link to keep record of system parameters changes which makes these techniques unviable for underwater wireless power transfer (UWPT). So a controlling mechanism without any communication link is crucially important. It is foreseeable that rapid ongoing research will bring promising revolution in this technology. Furthermore, it is expected that MIMO concept will make it possible to charge multiple devices simultaneously

For an efficient power transmission, WPT should fulfill three conditions: (a) high power, (b) large air gap, (c) high efficiency [4]. The efficiency of WPT system also depends on WPT technique. Transfer efficiencies of different WPT techniques for both near and far region is shown in Fig. 1. 70% to 90% transfer efficiency can be achieved through inductive coupling while magnetic resonance coupling offers a moderate efficiency of 40% to 60%. These transfer efficiencies decay with distance. The efficiency of laser and microwave is less than 30% and it rapidly decrease as the separation distance between transmitter and receiver coils increases.
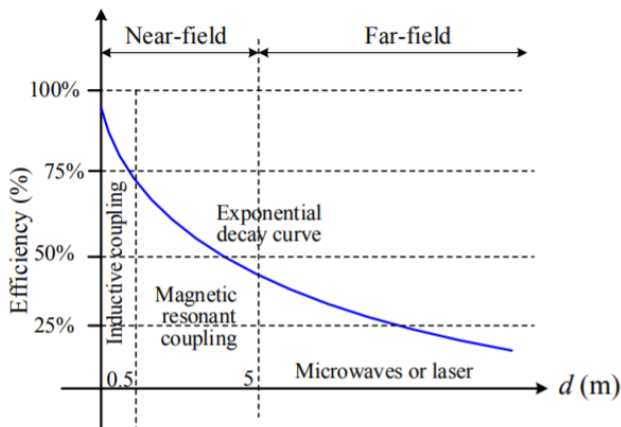
Fig 1.   Efficiency for WPT Techniques [5].

II.   RELATED WORK

Hassnain *et al.* [6] presented a systematic review on UOWC. They have briefly discussed underwater medium parameters and losses. Numerous researchers have made significant contributions to control load power and enhance efficiency in WPT systems. T. Nguyen et al. [7] have provided a comprehensive survey on EM based WPT technologies for UAVs. C. Cai et al. [8] designed a magnetic coupler for wireless charging of AUVs. Eddy current losses and coil structure to reduce these losses are briefly addressed in [9]. Zheng et al. [10] demonstrated a control mechanism for UWPT system by using artificial neural networks. Zhengchao Yan et al. analyzed eddy current losses under different frequencies and misalignments [11]. IF Lopes et al. [12] have discussed a maximum efficiency tracking strategy for UWPT. A systematic review and critical analysis on UWPT is presented [13]. In [14], Tamura et al. have focused on some pivotal elements to improve coupling coefficients for underwater wireless power transfer (UWPT). In-depth insight is given in [15] to design a WPT system for underwater environment. A comprehensive review of docking methods, AUV energy and AUV IWPT system are discussed in [16]. Researchers have given the best practical overview to design an AUV IWPT system in this article. Table I summarizes an yearly periodical overview of wireless power transfer technology implemented by several research groups.

TABLE I.   RESEARCH WORK

Reference #	Focused Key Challenges	Year
Ref. [1]	Frequency Splitting of UWPT	2016
Ref. [4]	UWPT via Electric Coupling	2016
Ref. [6]	EM based WPT Technologies for UAVs	2020
Ref. [8]	Magnetic Coupler for AUVs Charging	2020
Ref. [9]	Coil Structure to Reduce Eddy Current Losses in WPT	2019
Ref. [10]	Maximum Efficiency Tracking Control for UWPT	2020
Ref. [11]	Eddy Current Loss Analysis in UWPT	2018
Ref. [12]	Maximum Effciency Tracking Strategy for UWPT	2020
Ref. [13]	Systematic review and Critical Analysis on Wireless Underwater Power Transfer	2020
Ref. [14]	Lightweight and High Efficiency Coupler for UPWT	2019
Ref. [15]	Maximum Power Efficiency Tracking for UWPT	2018
Ref. [16]	WPT for AUVs	2019
Ref. [17]	Power Transmission Design for UWPT	2019
Ref. [18]	Efficient Modeling of UWPT using Z Parameters	2019
Ref. [19]	Capacitive WPT for Fresh Water Operations	2018
Ref. [20]	Load Modulation for UWPT	2017
Ref. [21]	UWPT with Curly Coil Structure for AUVs	2019
Ref. [22]	Omnidirectional and Positioning Tolerant AUVs Charging Platform	2019

III. WPT METHODS FOR UNDERWATER APPLICATIONS

The main advantages of wireless charging are reliability, convenience, safety, and a complete automated charging process. These advantages can be obtained through using different techniques. These techniques are classified into two categories: radiative and non-radiative technique. Radiative is RF based while non-radiative is coupling based as we can see in Fig. 2. Fig. 2 also shows that these techniques can be sub-divided in four techniques. In radiative technique, power is transferred as an RF wave or optical beam. In non-radiative technique, power is transferred via capacitive charging or EM induction. Furthermore, we have discussed different wireless charging methods for underwater applications.
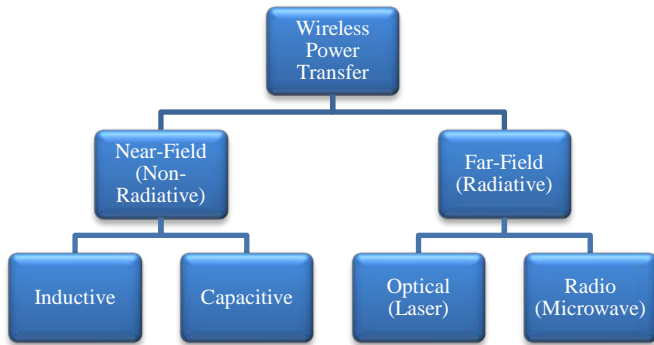
Fig 2. WPT Techniques.

A. Radiative WPT

Radiative WPT, which is commonly referred as far-field WPT, contains WPT techniques based on lasers or microwaves. Microwaves WPT ranges from 300 MHz to 300 GHz, it allows power transfer at large distance up to few meters. Microwaves WPT systems converts microwave power into DC power through energy-radiating antenna and rectenna. Shinohara et al. in [23] used microwaves WPT in near field by using flat beam pattern. Radio waves face high attenuation at high frequencies in seawater. A few researchers made efforts for underwater wireless power transfer (UWPT) through radio waves but resulted in low efficiency [24]. On the other hand, optical WPT makes use of an LED or LD as transmitter and light photons are detected by PD or solar panel to energy photons into electric signals. Researchers in [25] have successfully implemented blue laser diode for underwater communication and reported data transfer up to 12.4 Gb/s. However, laser based WPT is not realistic because of its low efficiency and it is also harmful to interface caused by ambient light and biofouling.

B. Nonradiative WPT

The crucial issue in radiative systems is signal attenuation. The nonradiative WPT systems involve inductive and capacitive power transmission techniques. In this power transmission techniques, the power transfer through magnetic and electric fields is restricted to short distances up to tens of centimeters. The capacitive wireless power transfer contains submerged electrodes which are separated by water medium. In CWPT, water is considered as a lossy dielectric medium. However, CWPT has been neglected due to low coupling capacitance and missing galvanic isolation. A high operation

frequency is required in CWPT. In IWPT, the coil diameter or number of turns should be increased to improve efficiency. IWPT system can involve shield materials e.g. ferrite material for better performance. At present, IWPT is more preferable for power transmission in hovering underwater or data transmission on station in seawater.

C. IWPT

We will discuss more matured IWPT technology here. IWPT, also considered as inductive wireless power transfer, is WPT based on near-field magnetic coupling through coils.

IWPT is more handy for underwater applications as it is efficient, safe and requires low frequencies as compared to CWPT. In 19th century, Nikola Tesla pioneered IWPT by using EM resonance [26]. He generated resonance for higher voltages to implement WPT to longer distances. The emerging WPT systems have proliferated in recent years by providing matured solutions for charging devices. Nowadays, IWPT systems are commercially available for charging biomedical equipments, electronic gadgets and electric vehicles (EVs). IWPT is efficient power transfer system which is less prone to misalignment between coils. IWPT has been used for underwater robotics [27-29] and ships [30], underwater sensors [31] and AUVs [32-33]. IWPTs eliminate biofouling challenges, DECs, corrosion and potential regarding leakage currents. IWPT saves operational cost for AUVs. Therefore, an extensive research has been carried out from different institutes and leading research groups for development of standards for IWPT. These standards include operating frequency ranges, power levels, design, shielding and many other aspects. A commonly known Qi standard operates at frequencies lower than 250 kHz but its power is limited to 15W. The AirFuel Resonant standards operate at 6.78 MHz from medical, scientific and industrial band. However, these standards can still be directly implemented to marine applications with high attenuation at higher frequencies in seawater. We have summarized a comparison of different IWPT systems for various applications in Table II and Table III.

TABLE II. COMPARISON OF VARIOUS IWPT SYSTEMS

Parameter	Resonant IWPT	Loosely Coupled Transformer	LCWT
EMF cancellation	Required	Not required	Not required
Air gap	High	Medium	Very low
Leakage flux	High	Medium	Low
Multiple receiver	Possible	Not possible	Possible
Efficiency	Medium	High	Medium
Coupling coefficient	<0.25	>0.5	~1
Misalignment tolerance	High	Medium	Low
Eddy current losses	High	Considerable	Low
Hysteresis losses	No	Yes	Yes
Copper loss	Low	Low	Very High
Operating frequency	Very high	High	Medium

TABLE III.   COMPARISON OF VARIOUS IWPT USING RESONANT COILS

Ref.	Gap (cm)	Efficiency (%)	Power Level	Coupling Coefficient	Experiment Setup/Application
[15]	4	86	80 W	N/A	Underwater applications
[19]	2	80.7	400 W	N/A	AUVs
[34]	7.5	80	N/A	N/A	Coil submerge in water tank
[35]	20-50	97	1.1 MW	N/A	Charging ferry
[36]	26	80	3 kW	N/A	Halogen lamp
[37]	0.9	84	45 W	0.74	Underwater vehicle docking
[38]	N/A	85	75 W	N/A	AUV charging
[39]	4	82.13	N/A	0.25	Coils submerge in water tank
[40]	N/A	79	200 W	N/A	AUVs
[41]	0.2	70	250 W	N/A	AUVs
[42]	0.5	85	300 W	0.64	Ocean observation
[43]	0.2	90	400 W	0.765	Deep sea applications
[44]	0.6-1	88	500	N/A	AUVs
[45]	N/A	86.19	745	0.16	Leightweight AUVs
[46]	2.1	92.41	1 kW	0.1385	Charging AUVs
[47]	0.6	90	1 kW	0.64	Underwater vehicles
[48]	2.5	70.45	10 kW	0.54	Underwater applications
[49]	0.5	82	N/A	0.43	AUVs
[50]	8	80	600 W	0.49	AUVs

IV. PRACTICAL CONSIDERATIONS AND RESEARCH CHALLENGES

Besides all advantages, many technological difficulties exist in WPT system. Like in terms of system modelling, if we increase number of loads then it causes higher order and variable nonlinearities. It is worth considering, different loads have different power requiremetns. Such as desk lamps, mobile phones and computers which involve same WPT surface but their power requirements will be different. Thus, it is very important to guarantee the power requirements of each load. In addition, WPT system characteristics indicate that a remarkable influence on received power exists due to spatial positions of the electronic devices. Finally, new loads can access during operation of multi-load WPT systems. It will decrease power load stability and its efficiency.

Numerous researchers have made significant contributions to control load power and enhance efficiency in multi-load WPT systems. Researchers have involved dynamic controlling impedance matching to realize tunable allocation of power. In [7], scholars have maximized WPT system efficiency through impedance matching in S–S compensation topology of a multi-load system.
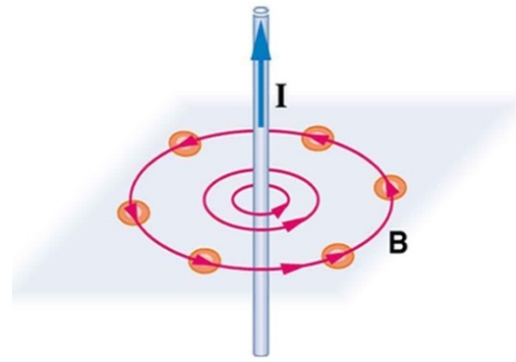
Fig 3.   Electromagnetic Induction.

Another challenging factor in WPT systems appears to be distance between transmitter and receiver component. The distance range of WPT plays an essential role in its applications. Wireless power transmission range can be in *mW* for biomedical implants while for electrical vehicles charging the range is up to *kW*. An intensive research has been carried out to enhance distance range and efficiency of wireless power transfer systems. Our designed experiments follow the theory of electromagnetic induction. When a current *i* flows through any coil then it creates a magnetic field *H* around that coil as shown in Fig. 3.

Commonly magnetic field H is expressed in equation 1. It is obvious from equations 1 that current intensity is proportional to magnetic field while magnetic field is inversally proportional to the distance from conductor.

$$\mathbf{H} = \frac{i}{2\pi r} \tag{1}$$

$$\mathbf{B} = \mu\mathbf{H} \tag{2}$$

We have plotted magnetic flux density *B* in MATLAB for different values of current in space by considering equation 2 as shown in Fig. 4. Graph shows a clear decay in magnetic flux density while increasing distance.
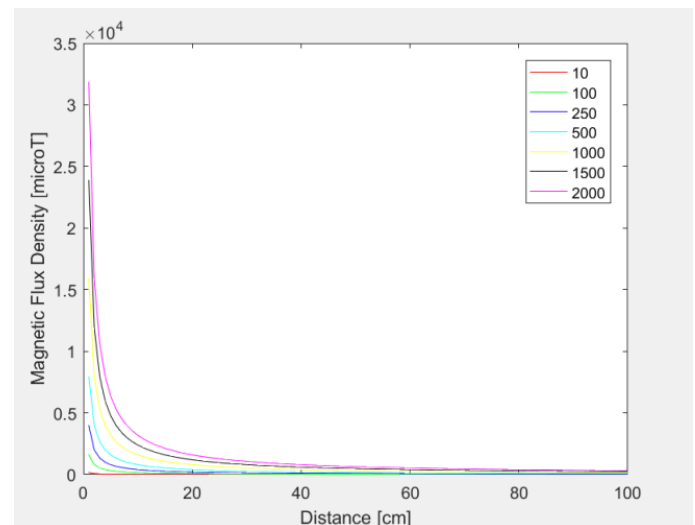
Fig 4.   Magnetic Field against Different Currents.

V. UNDERWATER COILS CHARACTER

We have opted circular coils in our system as it requires least material for design and it occupies lease space. Circular coil has less leakage flux and high misalignment tolerance. On the other hand rectangular coil needs more material for design and has more leakage flux. It shows wide variation in coupling when coils are misaligned. Hence, performance of circular coils is better and it is suitable for WPT applications. Characteristic comparison of both types of coils is shown in Table IV.

We have investigated the parasitic capacitance effect by measuring the impedance of coils through impedance analyzer. We checked the shift of its self resonance and coupling coefficient of coils. The coupling coefficient was 0.44 at 20mm transmission distance of coils with 15 turns with an inner diameter of 40mm. We have noticed that coupling coefficient decreases while increasing transmission distance as provided in Fig. 5.

We can conclude from graphs in Fig. 3 that coupling coefficient increases with increasing area and number of turns. Coil winding plays an effective role in efficient coupling [9]. Commonly two ways are adopted to design any coil; flat coil and solenoid coil as shown in Fig. 6.
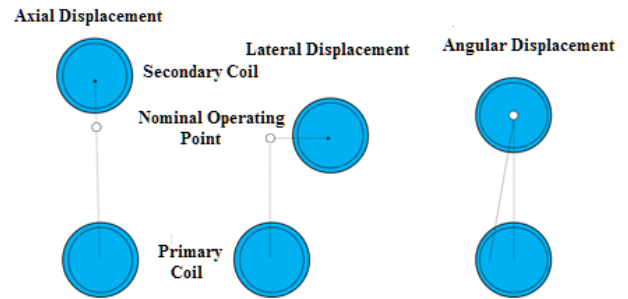
TABLE IV.   COIL CHARACTERISTICS COMPARISON

Parameter	Rectangular Coil	Circular Coil
Required Material	More	Less
Leakage Flux	More	Less
Leakage Flux Direction	Bottom	Horizontal
Horizontal Tolerance	Less	More
Vertical Tolerance	More	Less
Flux Distribution	Non-uniform	Uniform
Polarity	Polarised	Unipolar
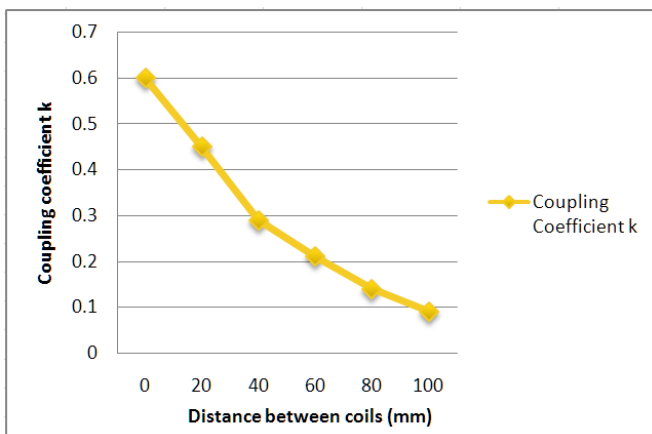
Fig 5.   Coupling Coefficient Variation with Distance.

Fig 6.   Coil Winding.

Fig 7.   Coil Orientation.

We conclude that mutual inductance has more influence on coupling coefficient than coil selection. The coupling coefficient is strongly connected to the relative distance and orientation between the two coils. Fig. 7 illustrates the three spatial degrees of freedom that affects the coupling coefficient of the system. The exact relationship between the coupling coefficient and the displacement is a complex subject that requires an analysis of the magnetic field strength and shape.

D. Coil Topology

It is essential to opt the best suited coil topology for UWPT. In this section we have discussed coil characteristics and design parameters which must be kept into consideration for coil design. We have comparatively investigated the effects of coil shape, size, coupling coefficient, orientation and number of turns. Subsequently, considering both number of turns and change in distance, it is concluded that the coupling coefficient of helix coil is lower than spiral coil. Scholars in [51] have briefly discussed coil topologies. A similar research study in [52] shows that spiral coil gives better efficiency than helix coil as the distance between the coils increases. Surprisingly, the coupling coefficient of spiral coil increases while for helix shape it decreases by increasing number of turns. The coupling coefficient can be referred as an important metric for any coil designing. It shows how much transistor coil flux is captured by the Rx coil. The results presented in Fig. 8 show that the coupling coefficient increases with number of turns. It validates that coil shape has a dramatic effect on coupling. In addition, coil material can also be protected from harsh seawater effects by encapsulating o potting in polyurethane. It has good resistance against seawater effect for a long period of immersion.
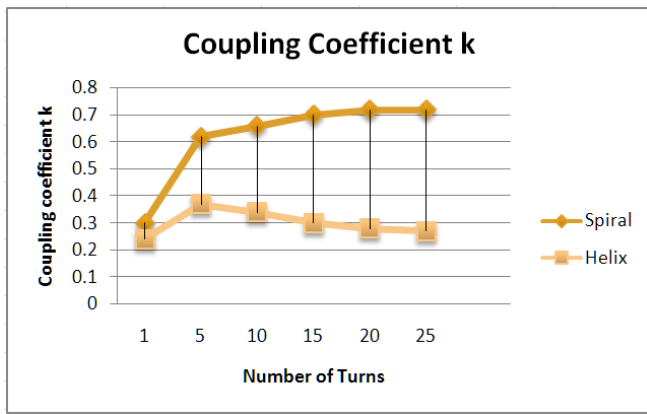
Fig 8.   Coupling Coefficient Variation with Coil.
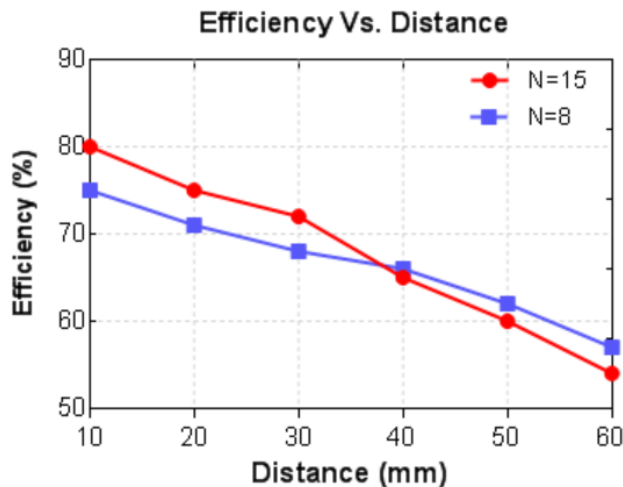
Fig 9.   Coil Effect on Efficiency.

In [53], D. Patil et al. have reported that coil with less number of turns gives better performance at longer distances due to decreased reactance. While coil with more number of turns performs better at shorter distances as shown in Fig. 9. Thus, these trade-offs should be kept into consideration for WPT system design.

VI. SEAWATER EFFECTS ON WPT

We have briefly discussed some effects of seawater on WPT system in this section.

A. Water Conductivity

The salt in seawater increases conductivity which creates eddy currents [54]. These generated eddy currents oppose magnetic field and reduce field strength of WPT system. It results into decrease in overall efficiency. The loss in proportional to the conductivity. It is concluded that high inductance yields high dependence on conductivity as it generates stronger magnetic field. In addition, it also depends on compensation scheme, if power flows in low voltage, it will generate a stronger magnetic field. It will result into high dependence on water conductivity. Dos Santos et al. in [54] have presented a relation between conductivity and efficiency as shown in Fig. 10.
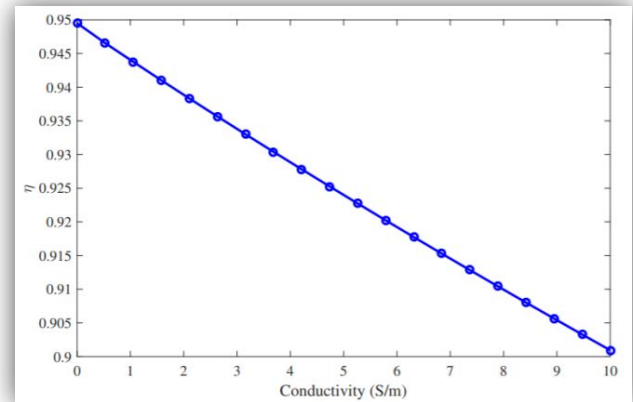
Fig 10.   Conductivity Effect on Efficiency [53].

B. OceanCurrents

The water current can adversely affect WPT system. It can cause misalignments between coils. In an article [47], authors have illustrated that EMF helps to stabilize AUV.

C. Biofouling

Another major problem in IWPT systems is biofouling. It s due to amassed marine microorganism on water surface. It can also cause high gap and misalignment between coils, due to which the WPT system efficiency is decreased. Anderson et al. demonstrated in [55] that heat produced in WPT coil can lower the fouling. Antifoul paints and external heating are significantly effective to control pad fouling.

D. Temperature and Pressure

Seawater is a cooling medium which can enhance the thermal limit of coils. Researchers in [53] demonstrated that coil power was increased from 600 W to 1 kW from air to water medium respectively. The potting material can also effect dissipated heat in coils. High hydrostatic pressure effect on WPT system efficiency is analyzed in [43]. The results show that system efficiency decreases at a depth of 4 km with 40 MPa pressure. It is due to the decrease in permeability under high pressure which decreases magnetic inductance. Magnetic inductance variation is low as the gap increases. However, authors did not include discussion about eddy current losses in seawater.

VII. ENGINEERING CHALLENGES IN WPT SYSTEMS

In this section, we have discussed major challenges for IWPT systems in underwater environment.

A. IWPT System for AUV

The integration of IWPT system on AUV without causing any problem in its shape and hydrodynamic mechanics is essential. The hull, lightweight but mechanically strong, is made of titanium and aluminum. In [37], J. Shi et al. mounted a secondary coil on AUV's hull which generates high magnetic flux. In [50], J.M. Cena et al. placed a secondary coil in AUV's hull which decreased the WPT efficiency. In some earlier works [40-41], [43], authors modified AUV shape which can affect hydrodynamic performance.

B. Pressure-Tolerant Electronics

Electronic equipments in AUV's are enclosed in high compressible material in order to tolerate ambient ocean pressure [56]. The material can be solid like urethane or fluid like insulating oil. Pressure tolerant system design is challenging as it involves an intelligent selection of electronic components. Some research groups [57,58] have provided a comprehensive review on selection of pressure tolerant electronics.

C. Docking Station Stability

Leveling, orientation and positioning should be maintained while deploying the docking station to seafloor. Research in [59] maintained dock leveling by using heavy railroad wheels. They have introduced additional dumping to overcome inadequate angel effect while employing cone docks. When we employ hooking concept, it can happen that whole docking process should be repeated.

D. Battery Charging Rate

The battery operation in dependent on thermal conditions, operation rate and ambient temperature [60]. It is difficult to obtain fast AUV charging behavior as charging and discharging rate is affected by lower ambient temperature. Therefore, commercially available batteries cause delay in mission time. Results discussed in [60] show that increased pressure has a negligible effect on capacity and resistance of battery. While reduced ambient temperature results in decrease in battery capacity and an increase in resistance. Y. Ji et al. [61] have presented a detailed analysis on reduced temperature effect on Li-ion batteries.

VIII. Conclusion

The WPT system submerged in seawater has to deal with challenges like corrosive salts, fluctuation and temperature. Thus, a watertight design is required which can halt intrusion and craggy enough to combat waves motion. This section summarizes the critical challenges for underwater wireless power transfer (UWPT). Seawater possesses inherent properties which creates difficulty in EM waves transmission. In addition, there is a rapid increase in radiation resistance on frequencies above 200 kHz which dominates AC and DC resistances. It limits operation frequency and it must be kept below threshold in order to achieve good efficiency. Another pivotal constraint is physical dimension of transceiver coils. Larger coils can be cumbersome to integrate while smaller coils can limit the throughput power. Researchers need to find an optimized design through proper simulation and experimental trials.

IX. Limitations

One major limitation in UWPT is misalignment between transmitter and receiver coils. Though stagger-tuning technique is implemented in some applications but UWPT still faces limitation due to variation in marine environment. This feature of lateral or angular misalignment between coils can cause in reduction in output voltage. Therefore, optimum frequency for operation is required while achieving a high WPT efficiency. Moreover, research community should investigate optimization

and coordination of WPT system in future. MIMO WPT technology can also necessitate power management.

X. Discussion

The UWPT system has to face several challenging factors like pressure, temperature and ocean currents. Seawater characteristics also put limit over EM waves transfer. In addition, an essential element is WPT system coil which can affect the transfer efficiency. An appropriate coil topology and accurate orientation is required for better performance. Thus, a proper engineering design is required to handle these challenges. Research community should focus on UWPT system design optimization through proper experimental tests.

XI. Future Design Improvements

Underwater environment is dynamic, uncertain and unstructured in nature. As a result, we encounter limitations when we try to deploy any system in this complex environment. Research community is working to tackle these challenges for significant power and data transfer in underwater environment. An adaptive control system can be used to attain maximum efficiency besides frequent change in coupling coefficient due to stochastic nature of seawater. The efficiency of our WPT system can be enhanced by introducing MIMO concept. Introducing MIMO signal processing and optimization in WPT can give promising results. An additional benefit of MIMO will be the distance range and area of charging station. An AUV can approach nearest charging station and can charge itself at any place of charging station as shown in Fig. 11.

In future, we will work to enhance efficiency of our proposed system by using MNZ metamaterials, multiple coil switching techniques and featuring coil size disparity. Analysis of 3D coil and Quasi-Omnidirectional WPT systems are our future research interests.
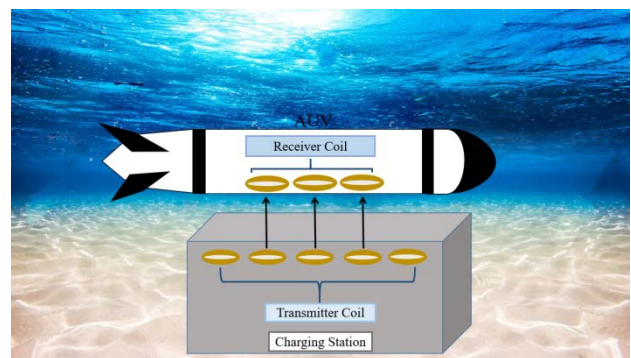

Fig 11. AUVs Self-Charging Station.

REFERENCES

[1] Niu, Wangqiang, et al. "Frequency splitting of underwater wireless power transfer." 2016 IEEE International Workshop on Electromagnetics: Applications and Student Innovation Competition (iWEM). IEEE, 2016.

[2] Hassnain, S. A., M. J. Mughal, and Q. A. Naqvi. "Layered Chiral Spheres with Zero Backscattering." 2019 Photonics & Electromagnetics Research Symposium-Fall (PIERS-Fall). IEEE, 2019.

[3] Hassnain, S. A., M. J. Mughal, and Q. A. Naqvi. "Analysis of Effective Medium Parameters on Polarizability of Homogeneous Chiral Sphere."

2019 Photonics & Electromagnetics Research Symposium-Fall (PIERS-Fall). IEEE, 2019.

[4] Urano, Mitsuhiro, and Akira Takahashi. "Study on underwater wireless power transfer via electric coupling." 2016 IEEE International Meeting for Future of Electron Devices, Kansai (IMFEDK). IEEE, 2016.

[5] Jawad, Aqeel Mahmood, et al. "Opportunities and challenges for near-field wireless power transfer: A review." Energies 10.7 (2017): 1022.

[6] Mehedi, Syed Agha Hassnain Mohsan Md, et al. "A Systematic Review on Practical Considerations, Recent Advances and Research Challenges in Underwater Optical Wireless Communication." IJACSA 11-7 (2020)

[7] T Nguyen, Minh, et al. "Electromagnetic field based wpt technologies for uavs: A comprehensive survey." Electronics 9.3 (2020): 461.

[8] Cai, Chunwei, et al. "A Circumferential Coupled Dipole-Coil Magnetic Coupler for Autonomous Underwater Vehicles Wireless Charging Applications." IEEE Access 8 (2020): 65432-65442.

[9] Zhang, Kehan, et al. "A new coil structure to reduce eddy current loss of WPT systems for underwater vehicles." IEEE Transactions on Vehicular Technology 68.1 (2018): 245-253.

[10] Zheng, Zhongjiu, Ning Wang, and Sara Ahmed. "Maximum efficiency tracking control of underwater wireless power transfer system using artificial neural networks." Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering (2020): 0959651820946510.

[11] Yan, Zhengchao, et al. "Eddy current loss analysis of underwater wireless power transfer systems with misalignments." AIP Advances 8.10 (2018): 101421.

[12] Lopes, Israel Filipe, et al. "Low-Frequency Underwater Wireless Power Transfer: Maximum Efficiency Tracking Strategy." IEEE Latin America Transactions 18.07 (2020): 1200-1208.

[13] Zhou, Jing, et al. "Design considerations for contact-less underwater power delivery: a systematic review and critical analysis." Wireless Power Transfer 7.1 (2020): 76-85.

[14] Tamura, Masaya, Kousuke Murai, and Daiki Fujii. "Lightweight and High-Efficiency Coupler Suitable for Underwater WPT System." 2019 IEEE Asia-Pacific Microwave Conference (APMC). IEEE, 2019.

[15] Orekan, Taofeek, Peng Zhang, and Cyuansi Shih. "Analysis, design, and maximum power-efficiency tracking for undersea wireless power transfer." IEEE Journal of Emerging and Selected Topics in Power Electronics 6.2 (2017): 843-854.

[16] Teeneti, Chakridhar Reddy, et al. "Review of Wireless Charging Systems for Autonomous Underwater Vehicles." IEEE Journal of Oceanic Engineering (2019).

[17] Silva, Miguel, et al. "Power Transmitter Design for Underwater WPT." OCEANS 2019-Marseille. IEEE, 2019.

[18] Kim, Jongwook, et al. "An Efficient Modeling for Underwater Wireless Power Transfer Using Z-Parameters." IEEE Transactions on Electromagnetic Compatibility (2019).

[19] Tamura, Masaya, et al. "Design of a capacitive wireless power transfer system for operation in fresh water." IEEE Transactions on Microwave Theory and Techniques 66.12 (2018): 5873-5884.

[20] Duarte, Candido, et al. "A study on load modulation for underwater wireless power transfer." OCEANS 2017-Aberdeen. IEEE, 2017.

[21] Yan, Zhengchao, et al. "Underwater wireless power transfer system with a curly coil structure for AUVs." IET Power Electronics 12.10 (2019): 2559-2565.

[22] Yang, Canjun, Tianlei Wang, and Yanhu Chen. "Design and analysis of an omnidirectional and positioning tolerant AUV charging platform." IET Power Electronics 12.8 (2019): 2108-2117.

[23] Shinohara, Naoki, and Naoki Kamiyoshikawa. "Study of flat beam in near-field for beam-type wireless power transfer via microwaves." 2017 11th European Conference on Antennas and Propagation (EUCAP). IEEE, 2017.

[24] Naka, Yasumasa, et al. "Improvement in efficiency of underwater wireless power transfer with electric coupling." IEICE Transactions on Electronics 100.10 (2017): 850-857.

[25] Wu, Tsai-Chen, et al. "Blue laser diode enables underwater communication at 12.4 Gbps." Scientific reports 7.1 (2017): 1-10.

[26] Rozman, Matjaz. Inductive wireless power transmission for automotive applications. Diss. Manchester Metropolitan University, 2019.

[27] B. Allotta, L. Pugi, A. Reatti, and F. Corti, "Wireless power recharge for underwater robotics," in Proc. IEEE Int. Conf. Environ. Elect. Eng., IEEE Ind. Commercial Power Syst. Eur., 2017, pp. 1–6.

[28] T. Assaf, C. Stefanini, and P. Dario, "Autonomous underwater biorobots: A wireless system for power transfer," IEEE Robot. Autom. Mag., vol. 20, no. 3, pp. 26–32, Sep. 2013.

[29] H. Jun, A. Asada, T. Ura, Y. Yagita, and Y. Yamauchi, "High speed acoustic network and noncontact power supplier for seafloor geodetic observing robot system," in Proc. OCEANS Conf., 2006, pp. 1–3.

[30] G. Guidi, J. A. Suul, F. Jenset, and I. Sorfonn, "Wireless charging for ships: High-power inductive charging for battery electric and plug-in hybrid vessels," IEEE Electrific. Mag., vol. 5, no. 3, pp. 22–32, Sep. 2017.

[31] D. Yoshioka, H. Sakamoto, T. Ishihara, T. Matsumoto, and F. Timischl, "Power feeding and data-transmission system using magnetic coupling for an ocean observation mooring buoy," IEEE Trans. Magn., vol. 43, no. 6, pp. 2663–2665, Jun. 2007.

[32] M. D. Feezor, F. Yates Sorrell, and P. R. Blankinship, "An interface system for autonomous undersea vehicles," IEEE J. Ocean. Eng., vol. 26, no. 4, pp. 522–525, Oct. 2001.

[33] T. Orekan and P. Zhang, Underwater Wireless Power Transfer Smart Ocean Energy Converters. New York, NY, USA: Springer, 2019.

[34] A. Askari, R. Stark, J. Curran, D. Rule, and K. Lin, "Underwater wireless power transfer," in Proc. IEEE Wireless Power Transf. Conf., 2015, pp. 1–4.

[35] G. Guidi, J. A. Suul, F. Jenset, and I. Sorfonn, "Wireless charging for ships: High-power inductive charging for battery electric and plug-in hybrid vessels," IEEE Electrific. Mag., vol. 5, no. 3, pp. 22–32, Sep. 2017

[36] M. Kesler and C. Mccarthy, "Highly resonant wireless power transfer in subsea applications," WiTricity, Apr. 2015.

[37] J. Shi, D. Li, and C. Yang, "Design and analysis of an underwater inductive coupling power transfer system for autonomous underwater vehicle docking applications," J. Zhejiang Univ. Sci. C, vol. 15, no. 1, pp. 51–62, Jan. 2014.

[38] V. Bana, M. Kerber, G. Anderson, J. D. Rockway, and A. Phipps, "Underwater wireless power transfer for maritime applications," in Proc. IEEE Wireless Power Transf. Conf., 2015, pp. 1–4.

[39] H. M. Santos, M. R. Pereira, L. M. Pessoa, C. Duarte, and H. M. Salgado, "Assessment of design trade-offs for wireless power transfer on seawater," in Proc. OCEANS Conf., 2016, pp. 1–7.

[40] M. D. Feezor, F. Yates Sorrell, and P. R. Blankinship, "An interface system for autonomous undersea vehicles," IEEE J. Ocean. Eng., vol. 26, no. 4, pp. 522–525, Oct. 2001.

[41] T. McGinnis, C. P. Henze, and K. Conroy, "Inductive power system for autonomous underwater vehicles," in Proc. OCEANS Conf., 2007, pp. 1– 5.

[42] J. Zhou, D. Li, and Y. Chen, "Frequency selection of an inductive contactless power transmission system for ocean observing," Ocean Eng., vol. 60, pp. 175–185, Mar. 2013.

[43] Z. Li, D. Li, L. Lin, and Y. Chen, "Design considerations for electromagnetic couplers in contactless power transmission systems for deep-sea applications," J. Zhejiang Univ. Sci. C, vol. 11, no. 10, pp. 824–834, Oct. 2010.

[44] S.Wang, B. Song, G. Duan, and X. Du, "Automatic wireless power supply system to autonomous underwater vehicles by means of electromagnetic coupler," J. Shanghai Jiaotong Univ., vol. 19, no. 1, pp. 110–114, Feb. 2014.

[45] T. Kan, Y. Zhang, Z. Yan, P. Mercier, and C. C. Mi, "A rotation-resilient wireless charging system for lightweight autonomous underwater vehicles," IEEE Trans. Veh. Technol., vol. 67, no. 8, pp. 6935–6942, Aug. 2018.

[46] T. Kan, R. Mai, P. P. Mercier, and C. Mi, "Design and analysis of a threephase wireless charging system for lightweight autonomous underwater vehicles," IEEE Trans. Power Electron., vol. 33, no. 8, pp. 6622–6632, Aug. 2018.

[47] T. Kojiya, F. Sato, H. Matsuki, and T. Sato, "Construction of noncontacting power feeding system to underwater vehicle utilizing electro magnetic induction," in Proc. Eur. OCEANS Conf., 2005, vol. 1, pp. 709–712.

[48] Z. Cheng, Y. Lei, K. Song, and C. Zhu, "Design and Loss analysis of loosely coupled transformer for an underwater high-power inductive power transfer system," IEEE Trans. Magn., vol. 51, no. 7, pp. 1–10, Jul. 2015.

[49] Z. Yan, K. Zhang, H. Wen, and B. Song, "Research on characteristics of contactless power transmission device for autonomous underwater vehicle," in Proc. OCEANS Conf., 2016, pp. 1–5.

[50] J. M. Cena, "Power transfer efficiency of mutually coupled coils in an aluminum AUV hull," M.S. thesis, , Nav. Postgrad. School, Monterey, CA, USA, 2013.

[51] P. Hadadtehrani, P. Kamalinejad, R. Molavi, and S. Mirabbasi, "On the use of conical helix inductors in wireless power transfer systems," in Proc. IEEE Can. Conf. Elect. Comp. Eng. (CCECE), May 2016, pp. 1–4.

[52] X. Shi et al., "Effects of coil shapes on wireless power transfer via magnetic resonance coupling," J. Electromagn. Waves Appl., vol. 28, no. 11, pp. 1316–1324, 2014.

[53] D. Patil, M. K. McDonough, J. M. Miller, B. Fahimi, and P. T. Balsara, "Wireless power transfer for vehicular applications: Overview and challenges," IEEE Trans. Transp. Electrific., vol. 4, no. 1, pp. 3–37, Mar. 2018.

[54] dos Santos, Hugo Miguel Guedes Pereira. "Underwater Wireless Power Transfer." (2016).

[55] G. Anderson, V. Bana, M. Kerber, A. Phipps, and J. D. Rockway, "Marine fouling and thermal dissipation of undersea wireless power transfer," SPAWAR Syst. Center Pac., San Diego, CA, USA, Tech. Rep. 2056, 2014.

[56] M. Luck et al., "Pressure tolerant systems for deep sea applications," in Proc. OCEANS Conf., 2010, pp. 1–4.

[57] R. P. Granger, C. M. Baer, N. H. Gabriel, J. J. Labosky, and T. C. Galford, "Non-contact wet mateable connectors for power and data transmission," in Proc. OCEANS Conf., 2013, pp. 1–4.

[58] "Seatooth Connect—Wireless Connector," 2015. [Online]. Available: http://www.wfs-tech.com/wp-content/uploads/2015/06/SeatoothConnect-15-1.0.pdf

[59] B. W. Hobson et al., "The development and ocean testing of an AUV docking station for a 21" AUV," in Proc. OCEANS Conf., 2007, pp. 1–6.

[60] K. Rutherford and D. Doerffel, "Performance of lithium-polymer cells at high hydrostatic pressure," in Proc. Unmanned Untethered Submersible Technol., 2005.

[61] Y. Ji, Y. Zhang, and C.-Y. Wang, "Li-ion cell operation at low temperatures," J. Electrochem. Soc., vol. 160, no. 4, pp. A636–A649, Feb. 2013.

An Intermediate Representation-based Approach for Query Translation using a Syntax-Directed Method

Hassana NASSIRI[1], Mustapha MACHKOUR[2], Mohamed HACHIMI[3]

Laboratory of the Computing Systems and Vision, Laboratory of Engineering Sciences
University Ibn Zohr, Agadir, Morocco

Abstract—We aspire to make one query reasonably sufficient to extract data regardless of the data model used in our research. In such a way, users can freely use any query language they master to interrogate the heterogeneous database, not necessarily the query language associated with the model. Thus, overcoming the needing to deal with multiple query languages, which is, usually, an unwelcome matter for non-expert users and even for the expert ones. To do so, we proposed a new translation approach, relying on an intermediate query language to convert the user query into a suitable query language, according to the nature of data interrogated. Which is more beneficial rather than repeat the whole process for each new query submission. On the other hand, this empowers the system to be modular and divided into multiple, more flexible, and less complicated components. Therefore, it increases possibilities to make independent transformations and to switch between several query languages efficiently. By using our system, querying each data model with the corresponding query language is no longer bothersome. As a start, we are covering the eXtensible Markup Language (XML) and relational data models, whether native or hybrid. Users can retrieve data sources over these models using just one query, expressed with either the XML Path Language (XPath) or the Structured Query Language (SQL).

Keywords—*Data Model; Relational Database; eXtensible Markup Language (XML); translation; model integration; intermediate representation; ANTLR (ANother Tool for Language Recognition)*

I. INTRODUCTION

The relational database has been the most data model used in most organizations to store and manage data. Likewise, the XML (eXtensible Markup Language) is progressively utilized as a universal solution to exchange data over the internet. At which point, many projects, and studies have been interested in integrating them and find means to interrogate both data. Some researchers focused on storing and querying XML data using a relational database system [1] [2]. Some others attempt to create general systems to manage XML, among other data formats [3]. However, approaches mentioned above have considerable advantages indeed but, along with limitations too, to some degree [4].

Nevertheless, by exploring some other orientations to query heterogeneous databases, especially those based on query translation, we perceived some related aspects to our intentions. Accordingly, adopting a translation tool can efficiently meet our aim, and using a syntax-directed approach would be a correct solution. To empower the process, we generate an intermediate query language that reflects the

logical interpretation of the query. We called it the universal query language (UQL); a transitional phase that provides an intermediate representation to switch between steps accurately, instead of converting the source query language directly to the target query language. The system is capable of performing queries against XML and relational databases and against hybrid ones too.

Henceforth, there is no need to be familiar with the many query languages to access data from variant data models, nor to express queries with precisely the suitable query language that corresponds to the data model used to structure that data. One query represented with either the Structured Query Language (SQL) or The XML Path Language (XPath) is moderately enough [5].

We are relying on the syntax-directed translation method, in which the parser drives the source query language translation. Therefore, semantic analysis and interpretation are performed based on the syntax structure. For the hands-on part in building language processing tools, handwriting the parser may work, but it is obviously not the best approach in complex cases. Alternately, using a powerful parser generator can save us time, effort, and resources as it is capable of automating momentous phases along the process. For that, to implement the parser, we are using ANTLR (ANother Tool for Language Recognition). It takes as input a grammar that specifies a language and generates as output source code for a recognizer for that language. A language is specified using a Context-Free Grammar (CFG), expressed using Extended Backus–Naur Form (EBNF).

The paper is organized as follows: This introduction introduces the general context of the project. Section 2 brings in some preliminaries and terminologies. Section 3 presents our objects and summarizes the mechanisms of the overall system and the translation process. Section 4 explores the language recognition and processing phase. Section 5 presents the intermediate representation phase. Section 6 discusses the data extraction phase and the nature of the database understudy that can be handled by the system. Finally, Section 7 concludes.

II. PRELIMINARIES

A. Describing a Language using a Grammar

A regular expression is quite useful but also leave little to no room for extension. Not all patterns can be described using regular expressions. The most obvious limitation is the lack of recursion. Statements can quickly turn out messy and hard to maintain [6] [7]. Thus, regular expressions are not quite

enough. Instead, CFGs, the type-2 grammar in Formal Grammar Hierarchy classification as known as Chomsky Hierarchy [8], would be a great deal to define the syntax of a language.

Formally, A CFG [9] is a 4-tuple (N; ∑; S; P) where:

- N is a finite set of variables called nonterminals;

- ∑ is a finite set of terminals;

- S: An axiom is it the start nonterminal

- P is a finite set of productions (rewrite rules).

Each production has the form $N \rightarrow (N \cup \sum) *$

The head consists of a single nonterminal, and the body is a sequence of terminals and nonterminals.

We use the CFG to replace nonterminals by a string of nonterminals and terminals. The language of grammar is the set of strings it generates. A grammar could tell us the valid options to put together a piece of code for a given language and help us recognizing and identifying typical portions structures quickly.

B. Grammar Notation

There are many ways to describe a grammar, but we are using EBNF [10]. It is an extended version of the BNF (Backus-Naur form), an unambiguous, formal and mathematical way to specify CFGs. It is more concise and widely used as a formalism to describe a formal language grammar with a precise structure. It can be considered a metalanguage as it is a powerful way to define other languages. An EBNF grammar of a language consists of a set of terminal symbols and a set of productions for nonterminals, which shows the way terminal symbols are combined into a proper sequence.

C. Another Tool for Language Recognition (ANTLR)

It is possible to handwrite a parser from scratch, but this process can be complex, error-prone, and hard to change. Instead, there are many parser generators like Bison and Yacc [11] that take a grammar expressed in a domain-specific way, and generate code to parse that language. We are using ANTLR [12] [13] [14], It is a parser generator that uses LL(*) parsing [15] [16]. It takes a grammar as input and generates parsers that can build and walk parse trees and generate abstract syntax trees that can be further processed with tree parsers. From antlr.org, ANTLR is a powerful parser generator for reading, processing, executing, or translating structured text or binary files. Also, ANTLRWorks [17] is a great ANTLR grammar development environment.

ANTLR is used by several popular frameworks, products, and projects, like:

- Apple, Oracle, NetBeans IDE, Eclipse projects (e.g., XText).

- Hive and Pig Languages use it to parse Hadoop queries,

- Twitter uses it to parse queries.

- Hibernate, Drools, JBoss, Groovy, Jython.

III. AIMS AND MECHANISMS

A database is a set of information stored by a tool according to a data model, a defined structure. To extract and manipulate this information, we need a query language. Sources can be stored according to any model; this means that it can be heterogeneous. Now, let us assume that we have faced one of these scenarios: (1) A user who has some data sources in XML, and knows only SQL. (2) A user who masters XPath and wants to access relational data. The common point between these two use cases is that the user query language does not match the interrogated data model. It is not easy to retrieve data because the appropriate query language is needed, which is XPath, for example, for the first case, and SQL for the second one. Besides, most of the time, users cannot master all of these query languages all at once. Each query language has a particular specification and probably challenging to learn. It is where our proposed system comes in. To overcome the dependencies between the data model and the query language, we develop a system to extract data regardless of the nature of the model used (XML or relational). Using only one single query posed freely with any query languages (SQL or XPath), as explained in Fig. 1.

Fig. 2 depicts the principle of the translation proposed in our approach based on an intermediate representation in place of a direct translation, which is beneficial for other interpretations and independent transformations. It strengthens the system to be more independent and modular.

As shown in Fig. 3, it all begins from the user, who is free to choose between two different query languages, SQL, or XPath, to express the query and submit it to the reader. The latter provides a uniform interface between users and the system, and read their queries. At the outset, we are dealing with characters, but we aspire to get an abstract syntax tree that enables us to perform other actions for analysis. That is where the language recognizer phase (Section 4) takes action. It consists of two parsers, one to parse SQL queries and the other to parse XPath queries. For that, we developed a lexer grammar and a parser grammar for each query language. At the end of this stage, the output is a parse tree that will be fed to the Analyzer, where the abstract syntax tree is built and processed. Then, selecting only the relevant information to develop our universal query language in the UQL Builder phase using the mapping rules module to map each part of the query with a suitable part in the UQL. After that comes the role of the translator in translating it into the target language. Then, the converted query is executed in the data extraction phase.
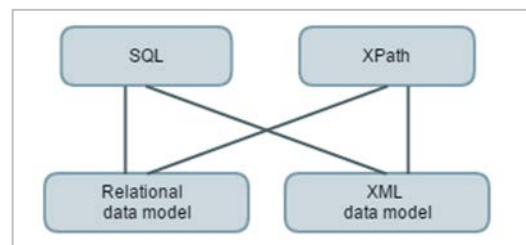

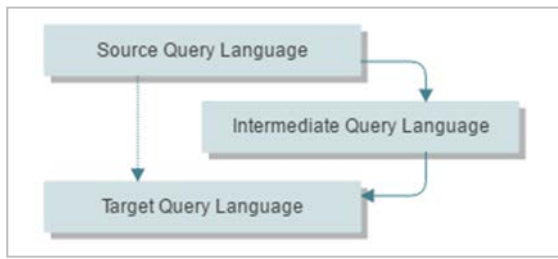Fig. 1. One Query to Retrieve Data from XML and/or Relational Data Models [18].

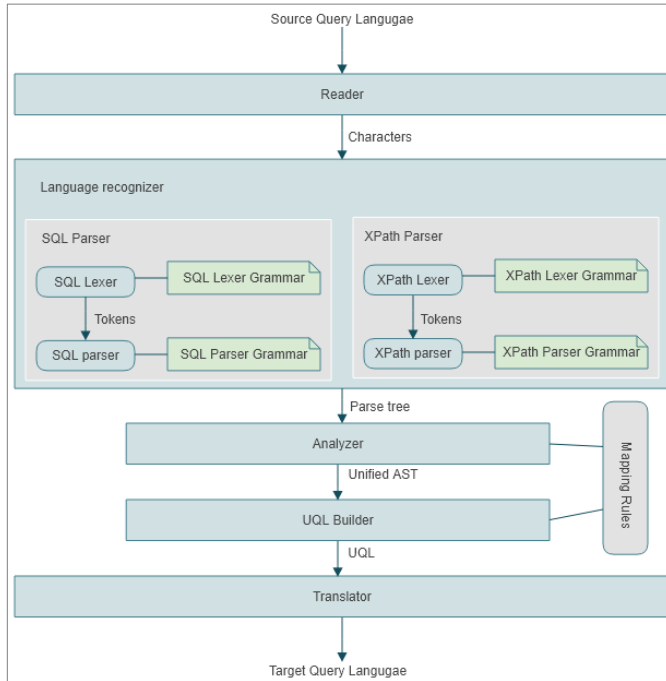Fig. 2. The Principal of the Translation [19].

Fig. 3. The Query Translation Process.

IV. LANGUAGE RECOGNITION AND PROCESSING

The previous section presents the overview of the translation principle briefly, and the phases pursued to convert the source query into the target query. This section describes the several steps in the language recognition and processing phase.

ANTLR admit three variants of grammar specifications: lexers, parsers, and tree walkers or tree-parsers, as shown in Fig. 4. All of them are alike, and the generated files behave the same way because ANTLR uses LL(k) analysis for all of them.

The lexer reads the input character by character and translates it into a sequence of syntactical units called Tokens. Then, fed to the parser, which takes a stream of tokens and produces a parse tree according to the grammar rules. Afterward, the tree walker process the Parse Tree produced.

A. Query Language Specifications

The first step is the grammar. Because we are covering XML and relational database models, we need to define the grammar of their query languages, namely XPath and SQL.

SQL is a powerful query language for managing and manipulating data and can fit almost every interaction aspect.

However, as the objective herein is interrogating data, we are focusing specifically on the select command, whose syntax is as follows in Fig. 5. Similarly, we are focusing on the most critical construct of XPath, the location path. Fig. 6 illustrates its EBNF notation.

B. Diagrammatic Form

ANTLR uses a simple EBNF-like syntax to define the grammar. For example, a column's syntax in a select clause can be written, as shown in Fig. 7, and presented in Fig. 8 using the railroad diagram.

Lexer rules start with an uppercase letter, and parser rules start with a lowercase letter.

Each rule has one or more patterns that it matches.

The K_AS? Means matches zero or one occurrence of K_AS.

'|' mean alternative patterns for the rule.

C. Parsing Queries

Parsing has the following phases: lexical analysis, syntactic analysis, semantic analysis. In the lexical Analysis (Tokenization), the lexer split up the user query into tokens and defines precisely how these tokens can be recognized. It reads a character stream as an input and generates a token stream as an output. Some tokens can be discarded like whitespaces; they are ignored during parsing. In the syntactic analysis, the parser figures out the relationship between the tokens that the lexer has produced to generate a parse tree, a data structure that reflects the input query's syntactic structure. In the Semantic analysis, the parse tree is checked for invalid semantic.
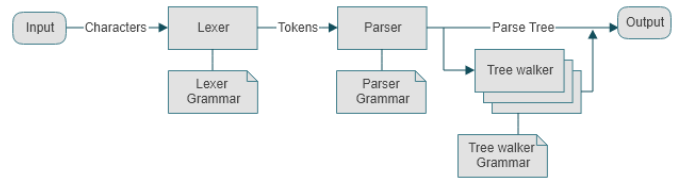
Fig. 4. Parsing Workflow.

```
select_clause ::=
        SELECT ( ALL | DISTINCT )? ( <star> | ( <select sublist> (
<comma> <select sublist> )* ) )
from_clause ::=
        FROM ( <table reference> ( <comma> <table reference> )* )
where_clause ::=
        WHERE <condition>
group by clause ::=
        GROUP BY ( ROLLUP <lparen> <expression list> <rparen> |
<expression list> )
having clause ::=
        HAVING <condition>
order by clause ::=
        ORDER BY <sort specification> ( <comma> <sort
specification> )*
```

Fig. 5. EBNF for SQL Select Grammar [20].

```
LocationPath ::=
        RelativeLocationPath
        | AbsoluteLocationPath
AbsoluteLocationPath ::=
        '/' RelativeLocationPath?
        | AbbreviatedAbsoluteLocationPath
RelativeLocationPath::=
        Step
        | RelativeLocationPath '/' Step
        | AbbreviatedRelativeLocationPath
Step ::=
        AxisSpecifier NodeTest Predicate*
        | AbbreviatedStep
AxisSpecifier ::= AxisName '::'
        | AbbreviatedAxisSpecifier
AxisName ::=
        'ancestor'
        | 'ancestor-or-self'
        | 'attribute'
        | 'child'
        | 'descendant'
        | 'descendant-or-self'
        | 'following'
        | 'following-sibling'
        | 'namespace'
        | 'parent'
        | 'preceding'
        | "preceding-sibling'
        | 'self'
NodeTest ::=
        NameTest
        | NodeType '(' ')'
        | "processing-instruction' '(' Literal ')'
Predicate ::=
        '[' PredicateExpr ']'
PredicateExpr ::=
        Expr
AbbreviatedAbsoluteLocationPath ::=
        '//' RelativeLocationPath
AbbreviatedRelativeLocationPath ::=
        RelativeLocationPath '//' Step
AbbreviatedStep        ::=
        '.'
        | '..'
AbbreviatedAxisSpecifier        ::=
        '@'?
```

Fig. 6.   EBNF for XPath Location Path [21].

The next section explores the process of generating the intermediate query language after parsing the source query, which is an in-between phase that helps generates the target query quickly. Section 6 provides further details regarding how the extraction works.

```
column
    : '*'
    | table_name '.' '*'
    | expression ( K_AS? column_alias )?
    ;
```

Fig. 7.   ANTLR EBNF Syntax for a column

Fig. 8.   Syntax Diagram for Column.

V.   INTERMEDIATE REPRESENTATION

The process of building the UQL starts from the output of the previous phase: the language recognizer. Then more steps have to be proceeded to get brief details that are needed, short, and to the point to efficiently generate the UQL. Besides, the parser generator builds a Concrete Syntax Tree (CST), not an Abstract Syntax Tree (AST). The CST reflects exactly the form of the grammar, every detail described in the syntax. It is like another representation of the grammar. That may seem easy to create but difficult to analyze and performed further interpretation with it. Whereas the AST contains only the mandatory elements needed and discard irrelevant details and extra information. It is more clear, compact, and easy to process than a parse tree. It is almost a direct translation of the grammar. We can get the abstract syntax from concrete syntax [22] [23]. After extracting the AST, all we need is to unify the ASTs and finally use the mapping rules to map every part of the unified AST and easily generates the UQL.

A.   Parse Tree

A parse tree or derivation tree is a data structure that matches the syntactic structure of the input. For example, the SQL select query: "*select first_name, last_name from Employee where id = 1;*" has the following parse tree (Fig.9), presented in tree form in Fig. 10.

B.   Abstract Syntax Tree

An AST is a variant of parse tree where we eliminate extra information and discard irrelevant details. Fig. 11 shows the AST for the query.

This example illustrates the output from a console, but we developed an interface ([5]) to present the analysis better.

```
select first_name, last_name from Employee where id = 1;
-- Reading data...
Your Query Language is: SQL
Columns: first_name last_name
Tables: Employee
Conditions : id=1
```

(select_statement (select_core (selectClause select (list_columns (column (expression (column_name (any_name first_name)))) , (column (expression (column_name (any_name last_name)))))) (fromClause from (list_tables (table_or_subquery (table_name (any_name Employee))))) (whereClause where (list_conditions (expression (expression (column_name (any_name id))) (comp_operator =) (expression (literal_value 1)))))) ;))

Fig. 9. Parse Tree for "Select First_Name, Last_Name from Employee where id = 1;".

Fig. 10. Parse Tree in Tree Form.

```
'- select_statement
  |- selectClause
  | |- TOKEN[type: 63, text: select]
  | '- columns
  |    |- anColumn
  |    | '- TOKEN[type: 67, text: first_name]
  |    |- TOKEN[type: 2, text: ,]
  |    '- anColumn
  |       '- TOKEN[type: 67, text: last_name]
  |- fromClause
  | |- TOKEN[type: 41, text: from]
  | '- tables
  |    '- TOKEN[type: 67, text: Employee]
  |- whereClause
  | |- TOKEN[type: 66, text: where]
  | '- conditions
  |    |- condition
  |    | '- TOKEN[type: 67, text: id]
  |    |- comp_operator
  |    | '- TOKEN[type: 6, text: =]
  |    '- lit_value
  |       '- TOKEN[type: 68, text: 1]
  '- TOKEN[type: 1, text: ;]
```

Fig. 11. Abstract Syntax Tree.

We get the following query after applying the unification principle illustrated in Fig. 12, along with the mapping rules to generate the UQL. We will take the same example from [5].

```xml
<?xml version="1.0" encoding="UTF-8"?>
<UQLroot>
  <Object>
    <ObjectName>emps</ObjectName>
    <Properties>
      <Property>nom</Property>
    </Properties>
  </Object>
</UQLroot>
```


Fig. 12. The unified Abstract Syntax Tree.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified">
  <xs:element name="UQLroot">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Object"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="Object">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="ObjectName"/>
        <xs:element ref="Properties"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="ObjectName" type=" xs:string "/>
  <xs:element name="Properties">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Property"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="Property" type=" xs:string "/>
</xs:schema>
```

Fig. 13. XML Schema.

Now just one more step to complete this phase is the XML document validation. A well-formed XML document is one that conforms to the syntactic rules of the XML language. When an XML document has an associated DTD (Document Type Definition) or XSD (XML Schema Definition) and respects it, it is said to be valid. Validation is a way to verify that the document conforms to a grammar. We use the XML Schema depicted in Fig. 13 to describe the structure of our XML document.

VI. Data Extraction

The system can access data from heterogeneous data models, namely relational and XML, since the relational database has been the popular option to store and manage data since 1970 [24]. It is still the most data model used in most organizations and powerful database systems [25]. Likewise, XML is widely used as a standard to exchange data over the internet, and the Native XML Database tends to be a practical solution for variable data [26] and provides full support for XML query languages such as XPath or XQuery [27]. The system can also access data from a hybrid database, as major relational database management systems (DBMS) are appealing for hybrid engines so that they fit XML into a relational database environment [3], for instance, Oracle [28] [29] [30], IBM [31] and Microsoft. Furthermore, the extension to SQL for XML, SQL/XML, is making good advancements [32] [33].

AS shown in Fig. 14, after executing the query against the suitable database, we perform other transformations to determine what is to be done with the data, and how to go about doing it. Lastly, format the answer according to the user preference, if the choice is indicated. If it is not the case, we apply the obvious choice, a tree form for XML sources, and tabular form for relational and hybrid databases. The tabular layout is the selected format by default.

Fig. 14. Data Extraction Phase.

VII. Conclusion and Outlook

This paper presents our intermediate representation-based approach for translating queries, relying on the syntax-directed translation technique, to access data from heterogeneous sources and get the most out of each technology. The intermediate transition aids in empowering the system feature, especially in terms of independence. So that matching the data model with the query language corresponding remains no longer bothersome or a burden. Herein, we covered XML and relational data models, whether native or hybrid and hoping to incorporate others in future contributions.

References

[1] Florescu and D. Kossmann, "Storing and Querying XML Data using an RDMBS," IEEE Data Eng. Bull., vol. 3, pp. 27–34, 1999, [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Storing+and+Querying+XML+Data+using+an+RDMBS#0.

[2] I. Tatarinov, S. D. Viglas, K. Beyer, J. Shanmugasundaram, E. Shekita, and C. Zhang, "Storing and querying ordered XML using a relational database system," 2002 ACM SIGMOD Int. Conf. Manag. Data, SIGMOD'02, no. October, pp. 204–215, 2002, doi: 10.1145/564691.564715.

[3] M. Rys, D. Chamberlin, and D. Florescu, "XML and Relational Database Management Systems : the Inside Story," SIGMOD '05 Proc. 2005 ACM SIGMOD Int. Conf. Manag. data, pp. 945–947, 2005.

[4] J. Shanmugasundaram et al., "Relational databases for querying XML documents: Limitations and opportunities.," Proc. 25th VLDB Conf., pp. 302-314. Morgan Kaufmann Publishers Inc., 1999, doi: 10.1016/j.acalib.2005.12.008.

[5] H. Nassiri, M. Machkour, and M. Hachimi, "One query to retrieve XML and Relational Data," in Procedia Computer Science, 2018, vol. 134, pp. 340–345, doi: 10.1016/j.procs.2018.07.201.

[6] L. G. Michael, J. Donohue, J. C. Davis, D. Lee, and F. Servant, "Regexes are Hard: Decision-Making, Difficulties, and Risks in Programming Regular Expressions," pp. 415–426, 2020, doi: 10.1109/ase.2019.00047.

[7] J. C. Davis, L. G. Michael, C. A. Coghlan, F. Servant, and D. Lee, "Why arent regular expressions a lingua franca? An empirical study on the re-use and portability of regular expressions," ESEC/FSE 2019 - Proc. 2019 27th ACM Jt. Meet. Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng., pp. 443–454, 2019, doi: 10.1145/3338906.3338909.

[8] N. Chomsky, "Three models for the description of language," IRE Trans. Inf. Theory, vol. 2, no. 3, pp. 113–124, 1956, doi: 10.1109/TIT.1956.1056813.

[9] A. Cremers and S. Ginsburg, "Context-free grammar forms," J. Comput. Syst. Sci., vol. 11, no. 1, pp. 86–117, 1975, doi: 10.1016/S0022-0000(75)80051-1.

[10] ISO/IEC 14977, "Information technology - Syntactic metalanguage - Extended BNF," Int. Stand. 149771996(E), vol. 1996, pp. 1–24, 1996, doi: 10.1002/(SICI)1099-1670(199603)2:1<35::AID-SPIP29>3.0.CO;2-3.

[11] S. C. Johnson, "Yacc : Yet Another Compiler-Compiler," Comput. Sci. Tech. Rep. No. 32, p. 33, 1975.

[12] T. Parr, "The Definitive ANTLR 4 Reference," Anim. Behav., vol. 67, no. 4, pp. 627–636, Apr. 2004, doi: 10.1016/j.anbehav.2003.06.004.

[13] T. Parr, The Definitive ANTLR Reference - Building Domain Specific Languages. 2007.

[14] T. Parr and K. Fisher, "LL(*): The foundation of the ANTLR parser generator," Proc. ACM SIGPLAN Conf. Program. Lang. Des. Implement., pp. 425–436, 2011, doi: 10.1145/1993498.1993548.

[15] T. J. Parr and R. W. Quong, "ANTLR : A Predicated- LL (k) Parser Generator," pp. 1–21, 1994.

[16] T. Parr, S. Harwell, and K. Fisher, "Adaptive LL(*) parsing," ACM SIGPLAN Not., vol. 49, no. 10, pp. 579–598, 2014, doi: 10.1145/2714064.2660202.

[17] Y. H. Wang and I. C. Wu, "ANTLRWorks: an ANTLR grammar development environment," Softw. - Pract. Exp., vol. 39, no. 7, pp. 701–736, 2009, doi: 10.1002/spe.

[18] H. Nassiri, M. Machkour, and M. Hachimi, "Integrating XML and Relational Data," in Procedia Computer Science, 2017, vol. 110, pp. 422–427, doi: 10.1016/j.procs.2017.06.107.

[19] H. Nassiri, M. Machkour, and M. Hachimi, "Querying XML and Relational Data," Int. J. New Comput. Archit. their Appl., vol. 7, no. 2, pp. 50–55, 2017, doi: http://dx.doi.org/10.17781/P002328.

[20] "BNF for SQL Grammar." https://docs.jboss.org (accessed Aug. 27, 2020).

[21] "xpath." https://www.w3.org/TR/xpath/ (accessed Aug. 27, 2020).

[22] D. S. Wile, "Abstract syntax from concrete syntax," pp. 472–480, 1997, doi: 10.1145/253228.253388.

[23] I. Ráth, A. Ökrös, and D. Varró, "Synchronization of abstract and concrete syntax in domain-specific modeling languages: By mapping models and live transformations," Softw. Syst. Model., vol. 9, no. 4, pp. 453–471, 2010, doi: 10.1007/s10270-009-0122-7.

[24] E. F. Codd and S. Jose, "A Relational Model of Data for Large Shared Data Banks," vol. 13, no. 6, 1970.

[25] Y. Bassil, "A Comparative Study on the Performance of the Top DBMS Systems," arXiv Prepr. arXiv1205.2889, pp. 20–31, 2012, [Online]. Available: http://arxiv.org/abs/1205.2889.

[26] G. Pavlovic-Lazetic, "Native XML databases vs. relational databases in dealing with XML documents," Kragujev. J. Math., vol. 30, pp. 181–199, 2007, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.7634&rep=rep1&type=pdf.

[27] W. C. W. D. March, XQuery 1 . 0 Formal Semantics, no. March. 2002.

[28] R. Murthy and S. Banerjee, "Xml schemas in Oracle XML DB," Proc. 29th Int. Conf. Very large databases, vol. 29, pp. 1009–1018, 2003, doi: 10.1016/B978-012722442-8/50094-X.

[29] S. Banerjee, V. Krishnamurthy, M. Krishnaprasad, R. Murthy, and O. Corporation, "Oracle8 i - The XML Enabled Data Management System Oracle Corporation for XML," vol. 2, no. 100, 2000.

[30] M. Krishnaprasad, Z. H. Liu, A. Manikutty, J. W. Warner, V. Arora, and S. Kotsovolos, "Query Rewrite for XML in Oracle XML DB," Data Base, 2004.

[31] F. Ozcan, D. Chamberlin, K. Kulkarni, and J. E. Michels, "Integration of SQL and XQuery in IBM DB2," Ibm Syst. J., vol. 45, no. 2, pp. 245–270, 2006, doi: 10.1147/sj.452.0245.

[32] A. Eisenberg, J. Melton, and O. Corp, "Advancements in SQL/XML," SIGMOD Rec., vol. 33, no. 3, pp. 79–86, 2004, doi: 10.1145/1031570.1031588.

[33] A. Eisenberg and J. Melton, "SQL/XML is making good progress," ACM SIGMOD Rec., vol. 31, no. 2, p. 101, 2002, doi: 10.1145/565117.565141.

An Adaptive Quality Switch-aware Framework for Optimal Bitrate Video Streaming Delivery

Wafa A. Alqhtani[1], Maazen S. Alsabaan[3]

Department of Computer Engineering
King Saud University
Riyadh, CO 11543 Saudi Arabia

Ashraf A. Taha[2]

Department of Computer Networks
City of Scientific Research and Technological Applications
SRTA-CITY, Alexandria CO 21934 Egypt

Abstract—**Given a large number of online video viewers, video streaming, over various networks, is important communication technology. The multitude of viewers makes it challenging for service providers to provide a good viewing experience for subscribers. Video streaming capabilities are designed based on concepts including quality, viewing flexibility, changing network conditions, and specifications for different customer devices. Adjusting the quality levels, and controlling various relevant parameters to stream the video content with good quality and without interruption is vital. This paper proposes an adaptive framework to balance the average video bitrates with respect to appropriate quality switches, making the transition to higher switches more seamless. The quality adaptation scheme increases the bitrates to the maximum value at their current quality switch before shifting to a higher level. This reduced switching times between levels and guarantees the stability of viewing and avoids interruptions. The use of a dynamic system ensures optimal performance, by controlling system parameters and making the algorithm more tunable. We built the system using an open-source DASH library (Libdash) with QuickTime player, studied the video load changes on two performance parameters, Central Processing Unit and Memory usages that have a high impact on multimedia quality. Consequently, the values of parameters that affected the performance of video streaming could be decreased, permitting users to regulate the parameters according to their preferences. Further, reducing the switching levels will reduce the overloads that occur while transferring from one level to another.**

Keywords—*Adaptive video streaming; average bit rate; mobile devices; modeling; quality of experience; quality switches; wireless networks*

I. INTRODUCTION

With the rapid growth of technology and wireless devices, the necessity for supporting various applications in the future has increased with the quality of service (QoS) requirements. A vital aspect of communication over various networks is video streaming, given the increase in the number of viewers of videos over the internet. There are several challenges that must be addressed to achieve seamless video streaming, such as avoiding interruption of playback, increasing video quality, reducing the initialization time, and reducing the number of video level switches. These can be addressed by adaptive video streaming, with control algorithms to deliver video with appropriate quality and with changes in network parameters. Applications of video streaming include live streaming, video on demand (VoD) services, and mobile applications. VoD

delivers video over the internet by dividing the video into parts called fragments, transmitting these parts, and enabling the receiver to decode and playback the video. This service allows for smooth streaming without having to wait for the entire video to be delivered, and allows the user to view the video at any time. Live streaming (real-time) transmits the contents to all users simultaneously, so that the fragments are transmitted at the same time as they are viewed by the users. With mobile applications, users operate mobile devices to download videos from online sources, such as YouTube, Vimeo, LiveTV, and PPStream. Several mobile applications have been available to enable users to stream videos online. There are three techniques for streaming video data. Using the first technique, referred to as progressive download, the server sends the video through the hypertext transfer protocol (HTTP) and at the receiver, the device downloads the file and can run the file after the buffering. This method of the basic version uses HTTP over the Transmission Control Protocol (TCP) so that the file is downloaded sequentially, where the user can watch after downloading part of the whole file. The second technique uses specialized protocols for streaming, real-time messaging protocol (RTMP) and real-time streaming protocol (RTSP), by sending chunks of data continuously to the media, which is displayed without a buffering or local caching. This technique is known as RTMP/RTSP streaming. The most popular technique, adaptive video streaming, collects the segments, encodes then indexes them, and subsequently determines their location (references) by using a profile file from the HTTP server. Table I provides a comparison of these three streaming technologies.

The adaptive streaming technique allows for the optimal video streaming experience for a range of dissimilar devices over a wide range of connection speeds. Generally, in an adaptive video streaming system (Fig. 1), a client plays a video that is received from a server. The client uses control algorithms to dynamically select the optimal video switch level for each downloaded segment and the period of idle times introduced to shape the received rate. In addition, the client uses a buffer to perform the synchronization inside the contents of the video, because the bit rate and the bandwidth available to the network cannot be predicted. Adaptive video streaming technologies face a variety of issues that affect performance including bandwidth, error rate, delay and jitter, synchronization, heterogeneity, and the user interface [2].

TABLE I.   COMPARISON OF STREAMING TECHNOLOGIES [1]

	Progressive Download	Streaming	Adaptive Streaming
Basic Principle	Client requests for file using HTTP GET method and server sends the entire file over HTTP.	Server sends fragments of data based on client request. Just in time transfer of data.	Content is encoded at multiple bit rates. A manifest file maintains the details of the fragments and their location. Client requests best suited fragments from the list.
Transport Protocol	HTTP over TCP	RTMP/RTSP over TCP/UDP	Simple HTTP server over TCP
Bandwidth usage	Less efficient and wastage of bandwidth as the entire file may not be played.	More efficient as only part of the file is downloaded being played.	Fragments can be cached and reused, thus saving bandwidth.
Content Security	Stored locally. Less secure.	No temporary storage. More secure.	Digital rights management (DRM) integration possible for specific adaptive streaming technology.
Advantages	Easy to setup. No special licenses required.	Can access any part of the video without waiting for an entire download.	High flexibility to change video quality.
Disadvantages	Bandwidth is wasted on data which is downloaded but not watched.	Adds significant cost and complexity to the setup and operations require special network configuration for port enabling.	Requirement to have multiple encoded version requiring additional content processing and storage.
Example of online video platforms	YouTube, Vimeo	Hulu	Network Television BBC, Netflix

Fig. 1.   Adaptive Video Streaming System [2].

Dynamic adaptive streaming over HTTP (DASH), also known as MPEG-DASH, provides high quality video streaming over the internet from an HTTP server. Fig. 2 illustrates a model of the MPEG-DASH setup. First, the multimedia content is captured and stored on an HTTP server and sent by HTTP. There are two types of content on the server. The Media Presentation Description (MPD) describes a manifest of the available content, its various alternatives, URL addresses, and other characteristics. Segments contain the real multimedia bitstreams in the form of fragments, in single or multiple files. The DASH client plays the content by parsing the MPD; therefore, the DASH client has information about the program timing, media content availability, media types, resolutions, minimum and maximum bandwidths, and the existence of various encoded alternatives of multimedia components, accessibility features and required digital rights management (DRM), media-component locations on the network, and other content characteristics. The DASH client uses this information to select the appropriate encoded alternative and to start streaming the content by fetching segments, using HTTP GET requests. After appropriate buffering to allow for network throughput variations, the client continues fetching the subsequent segments and monitors the network bandwidth fluctuations. Subsequently, the client decides how to adapt to the available bandwidth by fetching segments of different alternatives (with lower or higher bitrates) to maintain an adequate buffer, depending on its measurements. The MPEG-DASH specification only defines the MPD and segment formats [3].

Media content has several components (audio, video, and text), with each component having multiple characteristics. In MPEG-DASH, these characteristics are described in the MPD, in an XML format. Fig. 3 illustrates the MPD hierarchical data model. The MPD consists of one or multiple periods; each period has a starting time and duration, and consists of one or multiple adaptation sets. An adaptation set provides information about one or multiple media components and its various encoded alternatives. Each adaptation set usually includes multiple representations. A representation is an encoded alternative of the same media component, varying from other representations by bitrate, resolution, number of channels, or other characteristics. Each representation consists of one or multiple segments; media stream fragments in temporal sequence. Each segment has a URL that is an addressable location on a server that can be downloaded using HTTP GET or HTTP GET with byte ranges. DASH client first parses the MPD XML document. The client selects the set of representations it will use based on descriptive elements in the MPD, the client's capabilities, and the user's choices. The client then builds a timeline and starts playing the multimedia content by requesting appropriate media segments. Each representation's description includes information on its segments, which enables requests for each segment to be formulated in terms of the HTTP URL and byte-range [3].

The main contributions of this paper are as follows:

- Demonstrating some of the challenges that video streaming faces and how it affects video quality and investigating video streaming with a control algorithm to deliver video inappropriate quality with respect to network parameters changes.

- Designing an adaptive framework to balance the average video bitrate with respect to the appropriate quality switches and making the transition to higher switches more seamless.

- Proposing a method to decrease the impact of the parameter values on the performance of video streaming.
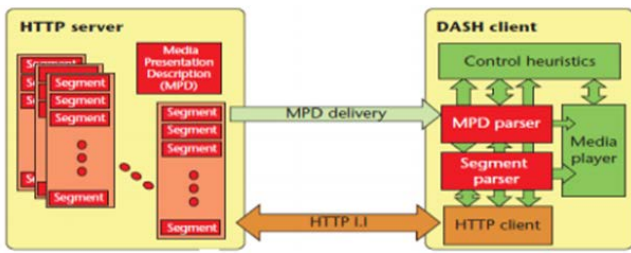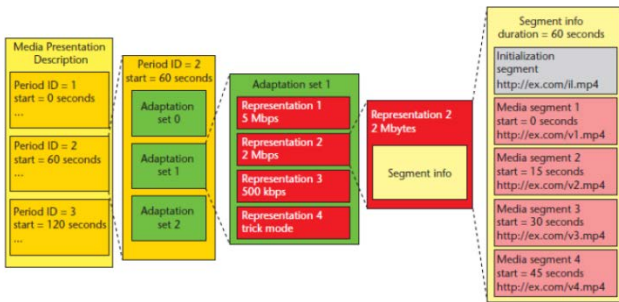
Fig. 2.   MPEG-DASH System Model [3].

Fig. 3.   MPD Hierarchical Data Model [3].

Our research can avoid missing live moments and adjust the video quality switch to the optimal bitrate level. In our proposal, we used two performance parameters: central processing unit (CPU) usage and memory usage. CPU usage limitation is about 25%, whereas memory usage limitation depends on the memory size and the registration of memory space during the streaming video, or when it exceeds the limit, which appears in red line. Imposing these conditions reduces switching between quality levels, and causes bitrates to balance with quality switching.

The remainder of this paper is organized as follows. Section II describes Literature Review. In Section III, the proposed quality adaptation framework is introduced and its functionality is explained. Section IV describes the datasets used and details of the results and discussion. Section V concludes the paper with a final review and presents future work.

II.   Literature Review

The major challenge associated with video streaming is delivering seamless video with maximum quality of experience (QoE). To this end, changes are made adaptively by the design control algorithm. Stream switching is a common control algorithm. The server encodes video content in different bitrates, while the control algorithm at the client side selects the appropriate video level. This approach is used by two main standards: MPEG-DASH and HTTP live streaming (HLS). Riad et al. [4] proposed a quality scheme to obtain a balance between the number of quality switches and the average bitrates at certain cutoff points. They achieved this by measuring the variation of bandwidth values, calculating the throughput change between pairs of consecutive results, and then making quality selection decisions by matching the channel alteration to the threshold. They evaluated their proposal by testing on actual datasets, comparing them with Liu's and Adobe algorithms [4]. These results showed their

scheme was successful in minimizing the number of qualities switching decisions, whilst keeping high average bitrates. In general, the proposed algorithm attains a good trade-off point between the number of quality switches and the average bitrates. Fig. 4 illustrates the influence of the cutoff on the average bitrate and over switches quality on a certain stream.

Fig. 4.   Effect of a Cutoff on Average Bitrate and Quality Switches for Specific Stream [5].

In [5], Xiang et al. designed a rate adaptation algorithm to find an ideal streaming strategy in a user- aware QoS, playback breaks, average playback quality, and playback smoothness. They formulated the rate adaptation problem as a finite Markov Decision Process (MDP) using dynamic programming. The optimal strategy requires the offered bandwidth statistics and a large number of states; therefore, it is hard to obtain the optimal solution in real-time, making it difficult to create an optimum streaming policy. To counter this, they produced an online algorithm that accumulates bandwidth statistics. The online algorithm also makes streaming decisions in real-time, using a reward parameter to ensure a good balance between average playback quality and playback smoothness. The experimental results presented showed the proposed algorithm is possible; however, several issues were encountered, which required further investigation. Improving quality control and adaptation algorithms has a noticeable effect on video streaming, there are recent studies that improve these algorithms based on the video quality scale. In [6], authors depend their studies on the P.1203 series of standards proposed by ITU-T is one such example for bit stream-based models. This series consists of three main parts: Pv: short term video quality prediction,

Pa: audio short term quality, Pq: overall integration of quality. This paper focus on extending the existing mode 0 model to support the aforementioned newer codecs and higher resolutions and frame rate. They propose correction mapping for new codecs, resolutions and frame rates and not retrain the existing model. As a result, we use the unmodified mode 0 predictions from the existing model and then do a correction on this prediction for the newer use cases. This approach of just using a correction mapping and not re-training ensures that we can rely on the well-developed P.1203 models. To ensure that the proposed correction reflects quality ratings by humans, two subjective tests were conducted. They first test considered, H.264, H.265 and VP9 with resolutions up to 4K, 60 fps as frame rate and realistic bitrate settings , a second test, included H.265 and AV1 as codec . These tests give a good example of using a simple correction chart. In thesis [7], Huang et al. designed a buffer-based algorithm to adaptation video rate by using the buffer to select a video bitrate, then request when

capacity estimation is required. This approach has two phases of the process. In the steady-state phase, when the buffer encodes appropriate information, the algorithm selects the video rate depending on the playback buffer. In the startup phase, when the buffer holds few information, we expand the buffer-based design with capacity estimation. Huang et al. revealed that this approach led to a reduction in the re-buffered rate by 10–20% compared to Netflix's default available bit rate (ABR) algorithm, while refining the steady-state video bitrate. The main reason is that DASH is solely a client-side standard. A DASH client is the only agent that manage the video streaming process despite (i) its limited information about the network and (ii) being unaware of actions taken by the other clients. in [8] the authors propose to maximize fairness and efficiency of end-users' QoE by achieving a level of cooperation between clients and servers without requiring any modification on the client-side, by using Dec-POMDP model and use RL to train two neural networks to find an optimal solution to the fairness problem. This optimal solution is then enforced, through client and server cooperation, to make their system fully compatible with the DASH standard. The experimental results proved that algorithm outperformed the state-of-the-art algorithm. In [9], they propose a novel algorithm for video rate adaptation in HTTP Adaptive Streaming (HAS), based on online learning, named Learn2Adapt (L2A), is to provide a robust rate adaptation strategy which, unlike most of the state-of-the-art techniques, does not require parameter tuning, channel model assumptions or application-specific adjustments. Simulations show that L2A improves on the overall Quality of Experience (QoE) and in particular the average streaming rate. The robustness property of L2A allows it to be classified in the small set of rate adaptation algorithms for video streaming, that mitigate the main limitation of existing mobile HAS approaches. Learning-based Adaptive Bit Rate (ABR) is approaches to learn outstanding strategies without any presumptions, has become one of the research hotspots for adaptive streaming. However, it typically suffers from several issues, i.e., low sample efficiency and lack of awareness of the video quality information. In [10], they propose Comyco, a learning-based ABR system which aim to thoroughly improve the performance of learning-based algorithm. To overcome the sample inefficiency problem, they leverage imitation learning method to guide the algorithm to explore and exploit the better policy rather than stochastic sampling, also including its NN architectures, datasets and QoE metrics. With trace-driven emulation and real-world deployment, the results of Comyco significantly improves the performance and effectively accelerates the training process. Joseph and De Veciana [11], established the online algorithm NOVA to optimize video delivery that supports DASH-based clients. NOVA is asynchronous, distributing the tasks of resource allocation to the network controller, and quality adaptation to respective video clients using minimal communication. In [12], Mao et al. proposed the Pensieve system. This system automatically learns algorithms without any predefined control guidelines or assumptions about the operating environment. This is achieved by using modern strengthening learning systems to learn the strategy of controlling adaptive bitrate through reinforcement. This enhancement is in the form of reflective traffics QoE for

previous video resolutions. This system takes special information about the real performance of the previous decisions to improve the control policy in the form of a neural network, so that the observations are used to decide the bitrate of the next fragment. The authors proposed learning-based approaches to producing ABR algorithms that rely on an effort to learn ABR policy from observations, especially as this method depends on learning enhancement. The aim of reinforcement learning (RL) is to increase the predictable cumulative discounted reward. Fig. 5 shows how RL is able to achieve bitrate adaptation. The ABR agent collects the metric information (bandwidth, bitrate of previous fragment, buffer occupancy) and applies it in a neural network model in the form of actions. The result is a bitrate decision; this QoE result is returned to the ABR agent as a reward. The ABR agent uses the reward for the training and development of the neural network model to improve performance. After applying the Pensieve system, experiments revealed that the proposed system deviated from ABR algorithms by 12% to 25%.

Cofano and De Cicco [13] proposed rules to guide the controller design, by designing a model to control the level based on a hybrid dynamic system. Based on this model, they derived a relationship between minimum switching frequency and control system parameters. They also proposed a methodology to adjust the lowest playout buffer that must be guaranteed to prevent rebuffing events as shown in Fig. 6. The general goal of the proposed control algorithm ([13]) is maximizing QoE for users in available bandwidth. To achieve this goal, they designed an official model of the closed-loop system by using a level-based hysteresis controller. Fig. 7 presents a comparison between the numerical simulations and the experimental results. The model in [13] fits the performance of the real system with good precision, and is able to expect system performance in terms of video level switching frequency and no rebuffing probability.
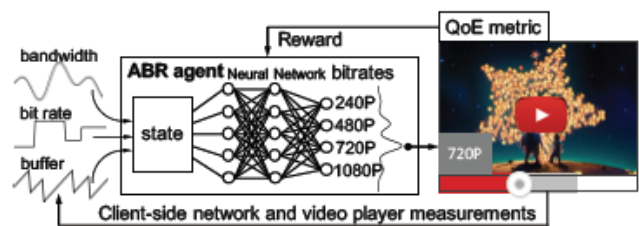

Fig. 5.   Applying Reinforcement Learning to Bitrate Adaptation [10].
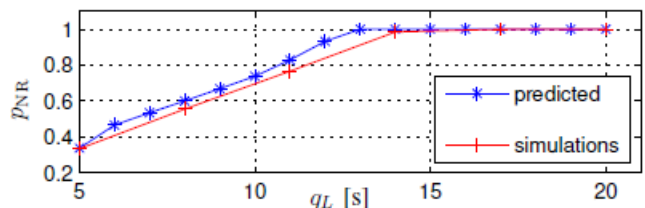

Fig. 6.   No Rebuffering Probability PNR Function of q_L, the Lower Playout Buffering Threshold [13].
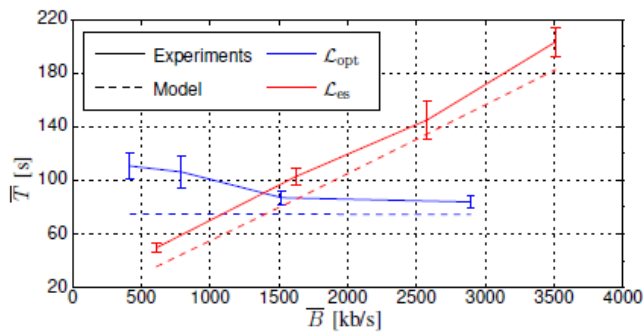
Fig. 7. Comparison between the Video Level Switching Period Obtained with Simulation Model and Experimental Result [13], between Two Sets of Data; the Optimal Level Set (Blue) and the Equally Spaced Level Set (Red).

Adaptive streaming is a technology that has to adapt video playback according to the network conditions. It is achieved by switching between representations of different bitrates and resolutions. These resolution changes affect the users' perceived quality. In [14], Asan et al. proposed a method to analyze resolution changes and their impact on QoE, as well as investigate adaptive patterns with respect to their mean opinion score (MOS). The results of this study are still inconclusive concerning single impairment factors that are typical for HTTP adaptive streaming (HAS) services. Their method attempted to predict the effect of a specific switch in terms of MOS degradation. This experiment is just one of the series of tests that they will conduct. In [15], Rodríguez et al. determined that changes in video quality level (VQL) had an effect on the user QoE. They proposed a DASH algorithm, including a decision parameter named the switching degradation factor (SDF) that captured a correlation between the QoE and VQL switching types, the frequency of VQL switching events and their temporal locations. The DASH algorithm was improved by performing VQL switching depending on SDF values. Parameter testing was done on the SDF model, alongside testing to assess the performance of the quality prediction in the MOS model. After analyzing and verifying the results, it was revealed that the MOS provided by the monitors had improved, through the incorporation of the SDF with the DASH algorithm.

In previous studies, researchers relied on various proposed methods and analyzed the results based on some of the parameters and algorithms, like measuring the variation of bandwidth pairs in consecutive results or using the MDP. Other studies depending on measuring buffer occupancy, throughput, average playback quality, and decision parameter to analyze their results. Another study testing their proposed algorithm with other algorithms by using H.264 codec. Our proposed method depends on some of the above previous studies by using H.264 codec, makes decision parameters based on preferences to improve the quality, and views stability and reduces switching levels, using average playback quality with evaluating network parameter values.

III. PROPOSED FRAMEWORK

The proposed method is to design an adaptive framework to balance the average video bit rates with respect to appropriate quality switches and make the transition to higher switches seamless. The quality adaptation scheme increases the bitrates to their maximum value corresponding to the current quality switch, before shifting to the higher switch. This will help in reducing the number of switching levels and reducing switching times between levels to guarantee viewing stability and avoiding interruptions. A dynamic system is required to achieve optimal performance by controlling system parameters and making the algorithm more tunable, allowing each user to regulate the parameters with respect to their own preferences. Further, reducing the switching levels will reduce the overloads that occur because of transferring from one level to another. Our research can avoid missing live moments and reduce the video interruptions by adjusting the video quality switch to the optimal bitrate level. The system model is designed as proposed in Fig. 8. The green parts (server) are standardized and contain the MPD and segment formats. The delivery of the MPD, DASH streaming control, media player, and segment parser, are depicted in blue. These parts are not standardized, allowing developers to modify or add features to improve their performance. The open-source (Libdash) is depicted at the client, containing the MPD parsing and HTTP module that is responsible for HTTP download. Therefore, the library provides interfaces for these modules to access the MPD and the downloadable media segments. The DASH streaming control is responsible for downloading the order of media segments. The DASH server provides segments in several bitrates and resolutions (MPD files). The client initially receives the MPD through Libdash; the MPD contains the temporal relationships for the various qualities and segments. Based on that information, the client can download individual media segments through Libdash at any point in time. Therefore, varying bandwidth conditions can be handled by switching to the corresponding quality level at segment boundaries, in order to provide a smooth streaming experience. This adaptation is not part of the Libdash and MPEG-DASH standard and we will implement it in our system to obtain our goals.
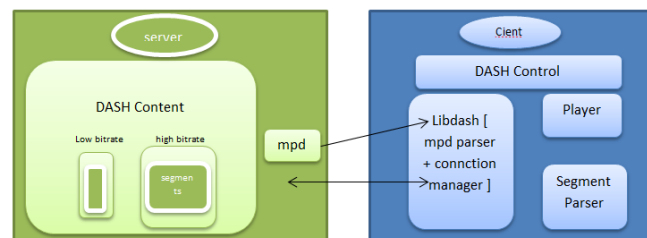

Fig. 8. System Model.

A. MPEG-DASH

Providing video content over the internet faces many challenges. Initially, the Real-Time Transport Protocol (RTP) was designed to define packet formats for audio and video content. However, the protocol performance is poor because it is used in internet protocol (IP) networks rather than content delivery networks (CDNs), and in the firewall most RTP packages are not supported. Therefore, the HTTP appeared to deliver the media content through it which is good with the firewall and uses streaming and smooth and dynamic streaming. But each of the streaming protocol has to deliver its own manifest and segment formats, so the content received

from the device must support the client protocol. MPEG-DASH is a technology that can provide interoperability between various servers and clients' devices. MPEG-DASH delivers a multimedia file to the client by using HTTP protocol and an MPEG coder. MPEG is the standard digital content format for transmission and storage of audio and video. The segment information called (MPD) Media Presentation Description.

After describing the details of the MPEG-DASH system, a description of the building of a system including client-side and server-side is provided. The server-side obtains MPD files and the client-side uses the Libdash open-source tool and DASH control. We installed the Libdash Library and opened it in visual studio 2015, with MPEG coder and QuickTime (QT) 5.11.2 tool to support the QT sample player. Libdash is the official reference software of the ISO/IEC MPEG-DASH standard, and is an open-source library that provides an object-oriented (OO) interface to the MPEG-DASH standard, developed by Bitmovin [16].

B. Performance Parameters

The dynamic systems have multiple parameters to measure the performance; in our system, we have used the common performance parameters, namely its CPU usage and memory usage. The CPU time is the amount of time for which the CPU was used for processing the instructions of a computer program or operating system. The CPU time is measured in clock ticks or seconds; it is useful to measure CPU time as a percentage of the CPU's capacity, called CPU usage. Now will explain how using Libdash Player and the details adaptation sets to balance between selecting the best level of quality and CPU usage. Starting performance profiling and selecting CPU usage, using first representation 320 x 240 (47 kbps), showed the CPU usage at less than 20%. By selecting another representation 1280 x 720, better quality and an increase in CPU usage to 39% were attained. Therefore, a high-resolution selection increases the amount of work for the processor. Referring back to the many studies of the maximum limitation of CPU usage in DASH streaming, it is expected than the upper limit of CPU usage on DASH streaming did not exceed 25% [17]. The diagnostic report is shown in Fig. 9, sorted by total CPU from highest to lowest in the selected time range. For the total CPU, the milliseconds and CPU percentage used by calls to the function in the selected time range include functions called by the function. This is different from the CPU utilization timeline graph, which compares the total CPU activity in a time range to the total available CPU [18].

$$TotalCPU\% = \frac{TotalMethodActivity}{AppaActivity} \times 100 \tag{1}$$

Self CPU unit refers to the time (in milliseconds) and CPU percentage used by calls to the function in the selected time range exclude functions called by the function [18].

$$SelfCPU\% = \frac{SelfMethodActivity}{AppActivity} \times 100 \tag{2}$$

Fig. 9.   CPU usage Report.

Here, (1) and (2) demonstrate the calculations of the CPU usage during the video flow and the average results.

Memory usage is the amount of memory currently in use by all applications. In the memory usage result, we obtained some snapshots from the memory by using a diagnostic tool to read the quantity of the data on the memory, including one or more snapshots of the managed and native memory heaps. A single snapshot can be analyzed to understand the relative impact of the object types on memory use, and to determine the code in the app that uses memory inefficiently. Two snapshots of an app can be compared to determine the areas in the code that cause the memory used to increase over time. A live graph of memory application is shown Fig. 10. The byte counter increases with time. The red dotted bar displayed over the graph is the memory threshold (Memory Limitation) [18]; the maximum limit does not exceed 1400 MB. To analyze memory usage, detailed report of memory usages can be obtained. The details of the difference between the current snapshot and the previous snapshot, such as the increase or decrease in memory usage, can also be obtained. The numbers in the snapshot panes indicate the bytes and objects in memory when each snapshot was taken, as well as the difference between the current snapshot and the previous one. We analyzed the numbers of Fig. 10 in depth as.

1) The total number of bytes in memory when the snapshot was acquired; a snapshot details report sorted by the total size of the type instances is displayed.

2) The total number of objects in memory when the snapshot was acquired; a snapshot details report sorted by the count of instances of the types is displayed.

3) The difference between the total size of memory objects in this snapshot and the previous snapshot (a positive number means the memory size of this snapshot is larger than the previous one, a negative number means the size is smaller); a snapshot difference report sorted by the difference in the total size of instances of the types is displayed.

4) The difference between the total number of memory objects in this snapshot and the previous snapshot; a snapshot difference report sorted by the difference in the total count of instances of the types is displayed.
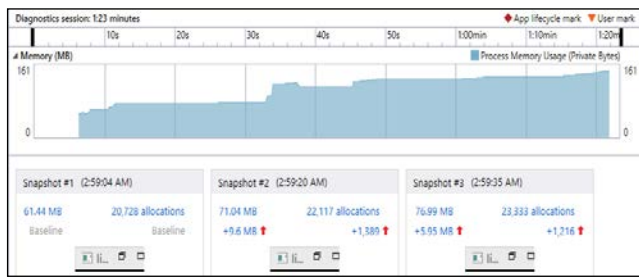
Fig. 10. Memory usage.

IV. RESULTS AND DISCUSSION

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Simulation and Test Scenarios

Our objectives were to balance the average video bit rates with the appropriate quality switches and make the transition to higher switches more seamless, by reducing the number of switching levels to reduce the overloads on the network while transferring from one level to another. The goal were achieved by studying the video load changes on the parameters of performance, such as CPU and memory usages, that have a high impact on multimedia quality. Maximizing the average bit rates for the current quality switch before shifting the switching level to the higher one enabled control over system parameters, thereby making the algorithm more tunable. We have divided the adaptation set into three groups of levels; every level has a set of converged qualities and data rates of CPU and memory usages. The transmission among levels is based on high or low quality and its effect on CPU and memory usages.

Table II shows the results of CPU and memory usage averages of all adaptation sets (according to the bitrates in the MPD file) excluding the parameters that have the biggest result from the limitation of CPU and memory usages. The average results were recorded because the results changed over time; memory usage value increases with time. The resolution value has multiple data rates; every data rate has value that effects CPU and memory usage. 320 x 240 resolution takes up approximately 9% of CPU usage and consumes between 100 to 120 MB of memory; however, at a higher resolution of 480 x 360, we observed that less than 10% of CPU usage and up to 137 MB of memory were consumed. Therefore, a switch to higher resolution consumes a larger amount of CPU and memory use, thus affecting the performance of video streaming. Further, as shown in Table II, we can observe that an 854 x 480 resolution has a CPU limitation and memory limitation, which means that other higher resolutions such as 1280 x 720 and 1920 x 1080 violate the conditions of CPU and memory limits. Therefore, we applied the reduction of switching quality levels only on resolutions from 320 x 240 to 854 x 480. Fig. 11 to 17 illustrate the results with curves of CPU usage and memory usage associated with the streaming

data, and the effect of multiple instances of switching quality on these performance parameters.

TABLE II. ADAPTATION SET

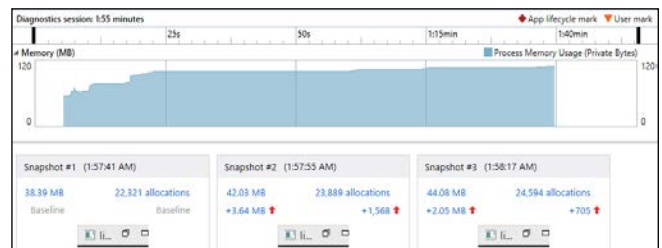| Resolutions	Data rates	CPU Usage	Memory Usage
320 x 240	47 kbps	8 %	120 MB
	92 kbps	8 %	126 MB
	135 kbps	9 %	1126 MB
480 x 360	182 kbps	12%	117 MB
	226 kbps	12%	117 MB
	270 kbps	13%	125 MB
	353 kbps	13%	127 MB
	425 kbps	13%	127 MB
854 x 480	538 kbps	19%	132 MB
	621 kbps	20%	143 MB
1280 x 720	808 kbps	39%	161 MB
	101, 103, 17 Mbps	up 42%	Up 210 MB
1920 x 1080	202, 206, 303, 308, 402, 407 Mbps	Up 60%	Up 230 MB

(a) CPU usage 320 x 240, 47kbps Bitrate.

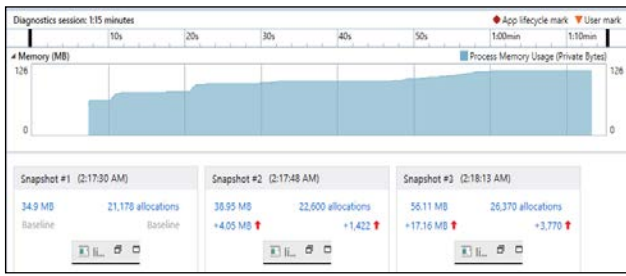(b) CPU usage 320 x 240, 92kbps Bitrate.

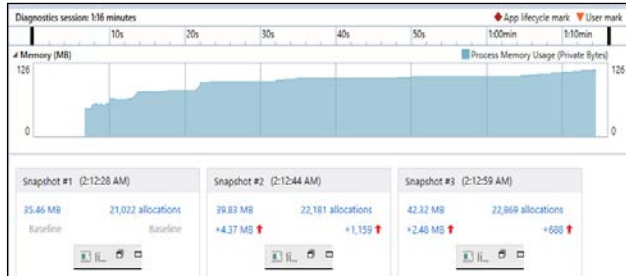
(c) CPU usage 320 x 240, 135kbps Bitrate.

Fig. 11. CPU usage of 320 x 240 for Bitrates 47, 92, 135kbps.

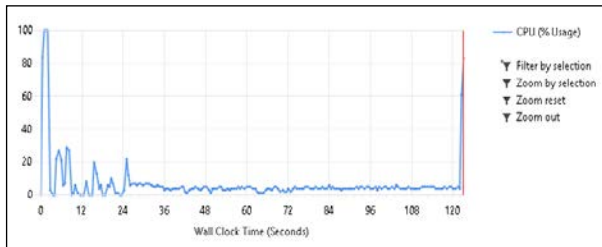(a) Memory usage 320 x 240, 47kbps bitrate.
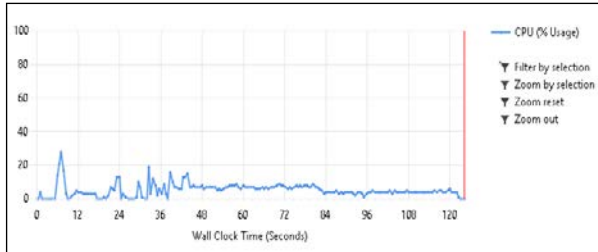
(b) Memory usage 320 x 240, 92kbps bitrate.

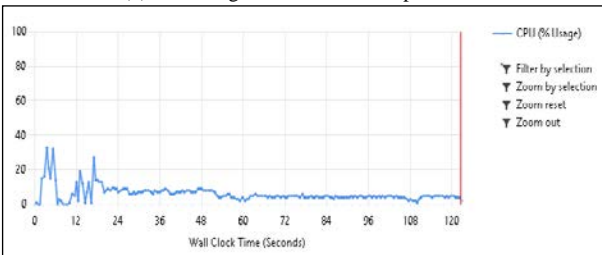(c) Memory usage 320 x 240, 135kbps bitrate.

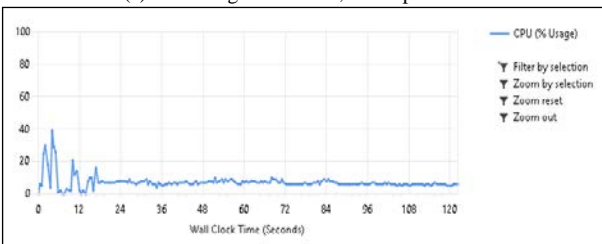Fig. 12. Memory usage of 320 x 240 for bitrates 47, 92, 135kbps.

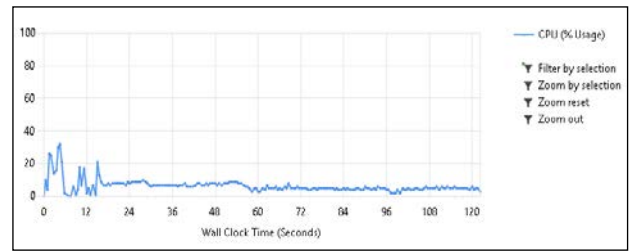(a) CPU usage 480 x 360, 182kbps bitrate.

(b) CPU usage 480 x 360, 226kbps bitrate.
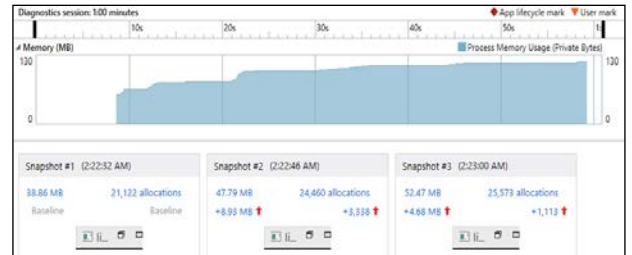

(c) CPU usage 480 x 360, 270kbps bitrate.

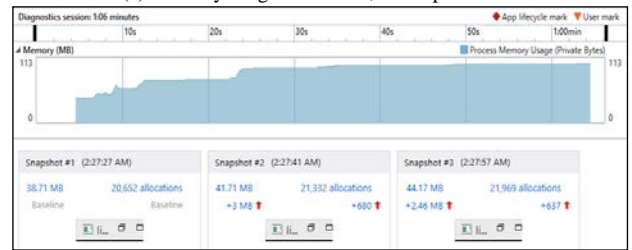(d) CPU usage 480 x 360, 353kbps bitrate.
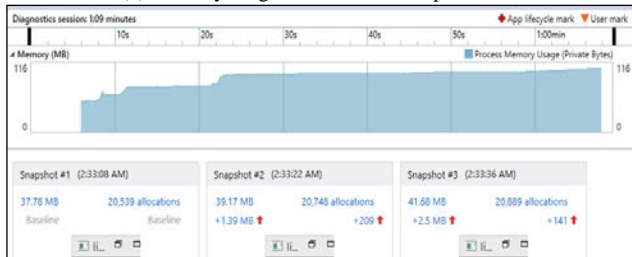

(e) CPU usage 480 x 360, 425kbps bitrate.

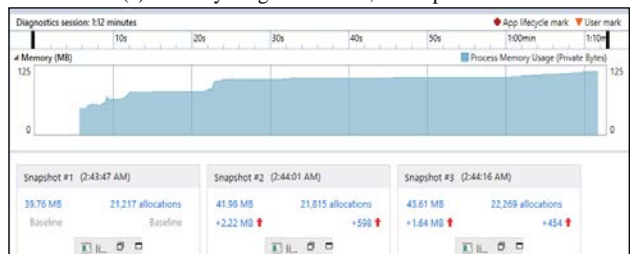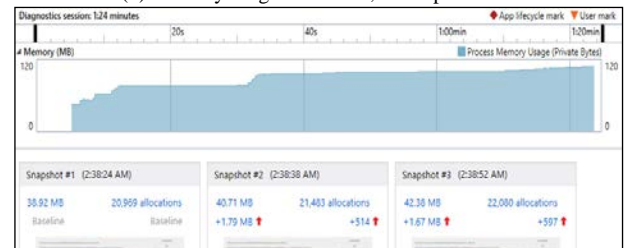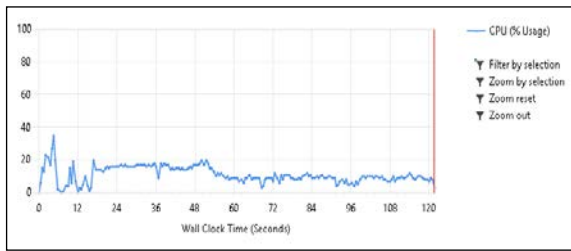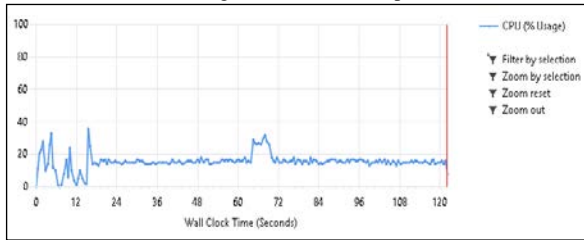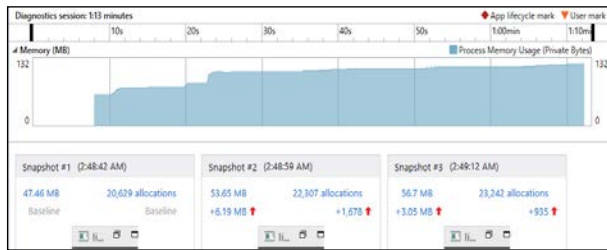Fig. 13. CPU usage 480 x 360 for Bitrates 182, 226, 353, 425kbps.

(a) Memory usage 480 x 360, 182kbps bitrate.

(b) Memory usage 480 x 360, 226kbps bitrate.

(c) Memory usage 480 x 360, 270kbps bitrate.

(d) Memory usage 480 x 360, 353kbps bitrate.

(e) Memory usage 480 x 360, 425kbps bitrate.

Fig. 14. Memory usage of 480 x 360 for bitrates 182, 226, 270, 353, 425kbps.

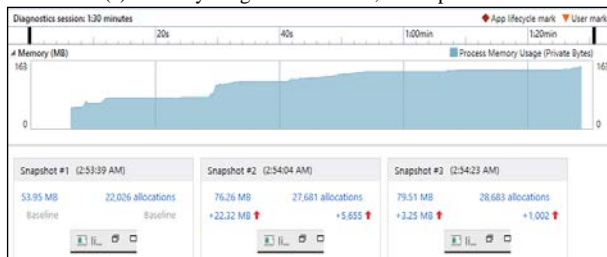(a) CPU usage 854 x 480, 538kbps bitrate.

(b) CPU usage 854 x 480, 621kbps bitrate.

Fig. 15. CPU usage of 854 x 480 for bitrates 538, 621 kbps.
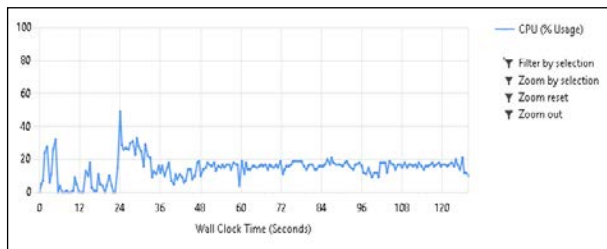

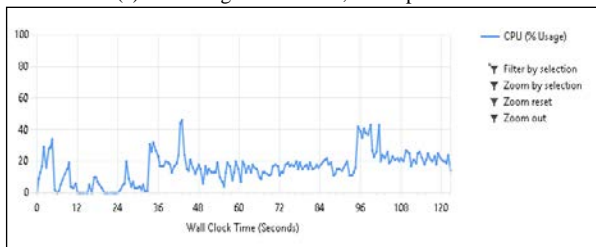(a) Memory usage of 854 x 480, 538kbps bitrate

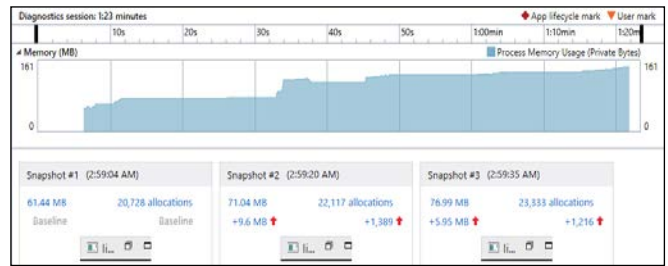(b) Memory usage 854 x 480, 621kbps bitrate.

Fig. 16. Memory usage of 854 x 480 for bitrates 538, 621 kbps.

(a) CPU usage 1280 x 720, 808kbps bitrate.

(b): CPU usage 1280 x 720, 1.7Mbps bitrate.

(c): Memory usage 1280 x 720, 808kbps bitrate.

Fig. 17. CPU usages of 1280 x 720 for Bitrates 808kbps and 1.7Mbps and Memory usages for Bitrates 808kbps.

B. Results and Discussion

Our contribution herein is the design of an adaptive framework to balance the average video bit rates with respect to appropriate quality switches and make the transition to higher switches more seamless. The quality adaptation scheme increased the bitrates to the maximum value corresponding to the current quality switch, before shifting to the higher level. This helped reduce the number of switching levels, and hence reduce switching times between levels to guarantee stable viewing and avoid interruptions. A dynamic system was required to achieve optimal performance by controlling system parameters (CPU and memory). This dynamic system was also required to make the algorithm more tunable, permitting each user to regulate the parameters with respect to their own personal preferences. Further, reducing the switching levels reduced the overloads that occurred because of transferring from one level to another. In this study, we analyzed the results of highest and lowest value of each data rate in one level of resolution, and reduced these levels by combining the closest results to closest level. Trading off between data rates and quality to make the stream video more seamless, see Fig. 18. which illustrates the processes of adaptive framework flowchart. Table III illustrates the difference between bitrates before and after the study of minimizing levels of quality to reduce switching between levels. We divided the adaptation set onto three levels only; every group of quality and data rates has close results on parameters of performance such as (CPU and memory usages). After searching and studying the results of CPU usage and memory usage, Fig. 11 to 17, it is revealed we can have only three levels on the adaptation set, namely high level (HL), middle level (ML), and low level (LL), as organized in Table IV.

TABLE III. ADAPTATION SET BEFORE PROPOSED METHOD

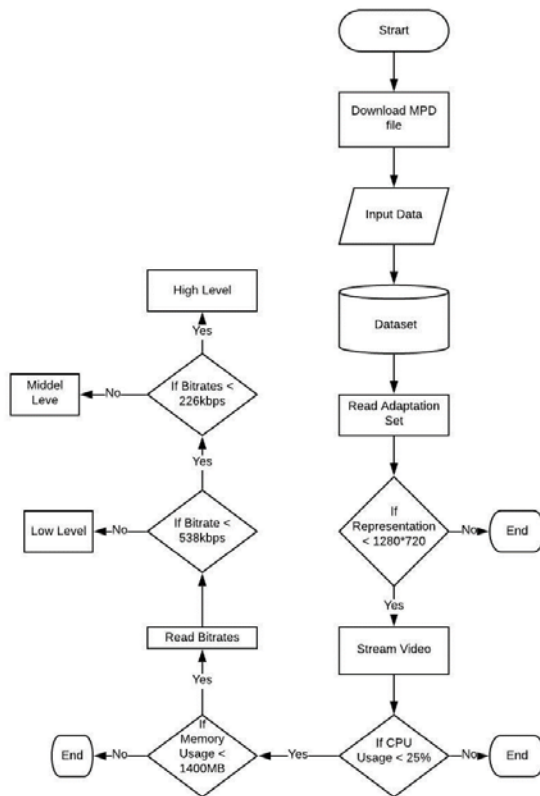| Resolution	Bit Rates (Before)	Bit Rates (After)
320 x 240	47 kbps	47 to 135 kbps 480x360, 182 kbps
	92 kbps	
	135 kbps	
480 x 360	182 kbps	480x360, 226 to 425 kbps
	226 kbps	
	270 kbps	
	353 kbps	
	425 kbps	
854 x 480	538 kbps	854x480, 538 to 621 kbps
	621 kbps	

Fig. 18. Adaptive Framework Flowchart.

TABLE IV. ADAPTATION SET AFTER PROPOSED METHOD

Levels	Adaptation Set
High Level (HL)	320 x 240, 47 to 135 kbps 480 x 360, 182 kbps
Middle Level (ML)	480 x 360, 226 to 425 kbps
Low Level (LL)	854 x 480, 538 to 621 kbps

V. CONCLUSION

In this study, we have investigated video streaming with a control algorithm to deliver video in appropriate quality with respect to network parameters changes. We designed an adaptive framework to balance the average video bitrate with respect to the appropriate quality switches and made the transition to higher switches more seamless. We used a dynamic system (Libdash) to achieve optimal performance by controlling system parameters (CPU usage and memory usage) to make the algorithm more tunable. After analyzing the results, we minimized the level of quality switches to have only three levels in the adaptation set. So, this study was able to decrease the values of parameters that affected the performance of video streaming.

In future research, we will expand this study to include other influences that affect video streaming performance and quality.

ACKNOWLEDGMENT

REFERENCES

[1] Suzen Saju Kallungal., "A Survey on Adaptive Video Streaming Technologies," IJARCET, Advanced Research in Computer Engineering & Technology, vol. 6, issue 3, pp. 350–352, March 2017.

[2] G. Cofano, L. De Cicco and S. Mascolo, "A Hybrid Model of Adaptive Video Streaming Control Systems," in Proc. 55th IEEE CDC, Las Vegas, NV, USA, Dec. 12–14, 2016.

[3] R. G. Asir, K. Kumar, and P. K. Reddy, "MPEG-DASH Enhanced Multimedia Streaming," IJARCSSE, vol. 4, issue 3, pp 848–851, March 2014.

[4] M. Riad, H. Abu-Zeid, H. S. Hassanein, M. Tayel, and A. A. Taha, "A Channel Variation-aware Algorithm for Enhanced Video Streaming Quality," in Proc. 4th IEEE LCN Workshops, Clearwater Beach, FL, USA, Oct. 26-29, 2015, pp. 893–898.

[5] S. Xiang, M. Xing, L. Cai, J. Pan, "Dynamic rate adaptation for adaptive video streaming in wireless networks," Signal Process. Image Commun., vol. 39, pp. 305–315, November 2015.

[6] R.R.Rao, S. Goring, P. Vogel, N. Pachatz, J. J Villamar, W. Robitza, P. List, B. Feitin, A. Raake, " Adaptive video streaming with current codecs and formats: Extensions to parametric video quality model ITU-T P.1203" in Proc. of Conference: Electronic Imaging, At Burlingame, California , January 2019.

[7] T-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A Buffer-Based Approach to Video Rate Adaptation," Stanford University, 2014.

[8] S. Altamimi, S. Shirmohammadi, " Client-Server Cooperative and Fair DASH Video Streaming" In Proc. Of Conference: the 29th ACM Workshop, June 21, 2019, Amherst, MA, USA.

[9] T. Karagkioules, G. S. Paschos, N. Liakopoulos, A. Fiandrotti, D. Tsilimantos, M. Cagnazzo, " Optmizing Adaptive Video Streaming in Mobile Networks via Online Learning " IEEE Transactions on Multimedia, 28 May 2019.

[10] T. Huang, C. Zhou, R. Zhang, C. Wu, X. Yao, L. Sun, " Comyco: Quality-Aware Adaptive Video Streaming via Imitation Learning" Proceedings of the 27th ACM International Conference on Multimedia, October 21–25, 2019, Nice, France.

[11] V. Joseph and G. de Veciana, "NOVA: QoE-driven optimization of DASH-based video delivery in networks", in Proc. IEEE INFOCOM'14 Conference of Computer Communication, April 2014, pp. 82–90.

[12] H. Mao, R. Netravali and M. Alizadeh, "Neural Adaptive Video Streaming with Pensieve," in Proc. Conference of the ACM Special Interest Group, Los Angeles, CA, USA, August 2017, pp. 4503–4653.

[13] G. Cofano and L. De Cicco, "Modeling and Design of Adaptive Video Streaming Control Systems," IEEE Transactions on Control of Network Systems, vol. 99, pp. 1-1, 22 November 2016.

[14] A. Asan et al., "Impact of Video Resolution Changes on QoE For Adaptive Video Streaming," in Proc. 18th IEEE International Conference on Multimedia and Expo., July 2017.

[15] D. Z. Rodríguez, Z. Wang, R. L. Rosa, G. Bressan, "The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over HTTP," EURASIP Journal on Wireless Communications and Networking, vol. 216, Dec. 2014.

[16] Bitmovin, July 2013. [Online]. Available: https://github.com/bitmovin/libdash.

[17] A. Wiersma, "Determining meaningful metrics for Adaptive Bit-rate Streaming HTTP video delivery," UVA University van Amsterdam, 15th June 2016.

[18] Visual studio 2019. [Online]. Available: https://docs.microsoft.com/en-us/visualstudio/profiling/cpu-usage?view=vs-2019. Microsoft.

A New Clustering Algorithm for Live Road Surveillance on Highways based on DBSCAN and Fuzzy Logic

Hasanain Alabbas[1], Árpád Huszák[2]

Department of Networked Systems and Services
Budapest University of Technology and Economics
H-1117 Budapest, Hungary[1, 2]
Al-Qasim Green University, Babylon, Iraq[1]

Abstract—**Video streaming over Vehicular Ad Hoc Networks is a promising technique (VANETs), and it has gained great importance in the last few years. The highly dynamic topology of VANETs makes high-quality video streaming very challenging. In order to provide the most useful camera views to the vehicles, clustering and cluster head selection techniques are used. Too frequent camera view changes can be annoying; therefore, we propose a new stable clustering algorithm to ensure a stable live road surveillance service without interruptions for vehicles that do not have enough vision area. In the proposed solution, we integrated Density-Based Spatial Clustering of Applications with Noise (DBSCAN) with Fuzzy Logic Control (FLC). DBSCAN is used to form the clusters, while FLC is used to find the best cluster head for the cluster. Different parameters are utilized like density parameters for DBSCAN, and relative speed, acceleration, leadership degree and vision area for fuzzy logic. Our proposed algorithm showed better results in terms of cluster lifetime and vehicle status change. Our proposed algorithm has been compared with another clustering scheme to prove the effectiveness of our proposed algorithm.**

Keywords—*Vehicular ad hoc networks (VANETs); V2V; intelligent transportation systems; clustering algorithms; road surveillance; DBSCAN algorithm; fuzzy logic control*

I. INTRODUCTION

Recently, Vehicular Ad Hoc Network (VANET) is deemed to be one of the most vital subjects which have gained noteworthy research consideration [1]. It is part of Intelligent Transportation Systems (ITS), which is intended to provide reliable communication between vehicles and fixed roadside units and among vehicles themselves by enabling them to make contact with each other directly [2]. VANET applications can be classified into safety, traffic efficiency, and luxury applications [3].

The communication entities, Road Side Unit (RSU) and On-Board Unit (OBU) form the backbone of VANET infrastructure. OBU is a communication device fixed in the vehicle, while RSU is a fixed unit distributed on the side of the roads or near the traffic signals [4]. In general, two connection types can be set up by this communication equipment, Vehicle-to-Vehicle (V2V), which enables the vehicular nodes to contact each other directly and Vehicle-to-Infrastructure (V2I), in which vehicles can communicate with

RSUs [5]. The standard 802.11p applied in VANETs is a branch of the standard IEEE 802.11, which represents a suitable solution for vehicular communications [6]. It has been mainly designed to provide a short-to-medium transmission range for high-speed vehicles up to (300 m - 1000 m) and to meet every V2V and V2I application [7].

VANETs inherit the main characteristics of Mobile Ad Hoc Networks (MANETs), in which the vehicles play the role of mobile nodes supplied with wireless communications [8]. Many applications are designed for safety and non-safety purposes, and one of these applications is the video streaming, which is now the focus of research community attention because it enriches the drivers and passengers with valuable information in comparison with textual messages [9]–[11]. Many challenges face video streaming service, which is considered highly-bandwidth-demanded like network congestion and the highly dynamic topology of VANETs. The network congestion happens if there are many vehicles that broadcast this service at the same time [12]. These factors have a significant impact on video streaming services.

Clustering strategies are considered one of the most effective solutions that were applied in VANETs to enhance the performance of the system, increase the scalability of the network and provide good management by grouping the vehicles into groups depending on some metrics [13].

Although VANET is a sub-class of MANETs, the conventional clustering strategies designated for MANET is unworkable to VANET directly because VANET has different challenges. The nodes in VANET can be connected or disconnected to the network very quickly, causing a highly dynamic topology and a frequently disconnected network. They are also constrained by the route of the roads and traffic regulation [15]. These parameters have a massive influence on the communication stability. A broad range of clustering algorithms was presented for VANET, and many issues are addressed by them. One of the problems, which is considered essential for clustering design in VANETs, is how to improve cluster stability [16]–[18]. In general, the VANET clustering algorithm divides the vehicular nodes into virtual sets named clusters. Each group elects a cluster head (CH) according to some rules set, while the other nodes in the group join the cluster as cluster members (CM). The CH is responsible for

cluster maintenance and coordinating the transmission among CMs in the same way as an infrastructure wireless access point.

In this paper, we propose a new algorithm aiming to achieve stable clusters and find a suitable CH for vehicles that tend to get the road conditions via video surveillance service. DBSCAN technique is used to configure the clusters, while Fuzzy Logic Control (FLC) is used to select the best CH. The elected CH will be responsible for providing video surveillance on the conditions of the road to all CMs in the same cluster, depending on the on-board camera substantiated inside the vehicle. Our proposed scheme has been compared with the Effective-Vision-Area-Based Clustering Algorithm with the Adaptive Video-Streaming Technique (EVAC-AV) algorithm and showed an effective result in increasing the cluster lifetime.

This paper is organized as follows. Section II presents the literature review. Section III describes the proposed clustering approach. In Section IV, the simulation environment and the methodology is shown. The performance evaluation and results are introduced in Section V. Finally; Section VI concludes the paper.

II. RELATED WORK

The clustering mechanism is an effective technique, which is used to streamline some critical functions like media access management, quality of service achievement, and bandwidth allocation, etc. In general, the nodes in the clustering algorithms have three states: CH state, normal state (NS), and CM state. These terms may vary in some articles, but they have the same notions. CH is the focal point of the cluster, which is elected to coordinate the cluster, while NS represents the state of a node that does not belong to any cluster. When it joins a cluster, it becomes a CM. Fig. 1 shows the topology of three clusters, in which each cluster elects a single CH. It clarifies how the different nodes are formed and grouped.

Due to the significance of the issues that clustering addresses, many clustering methods have been proposed lately in the context of VANETs. Most of them aim to achieve network constancy.
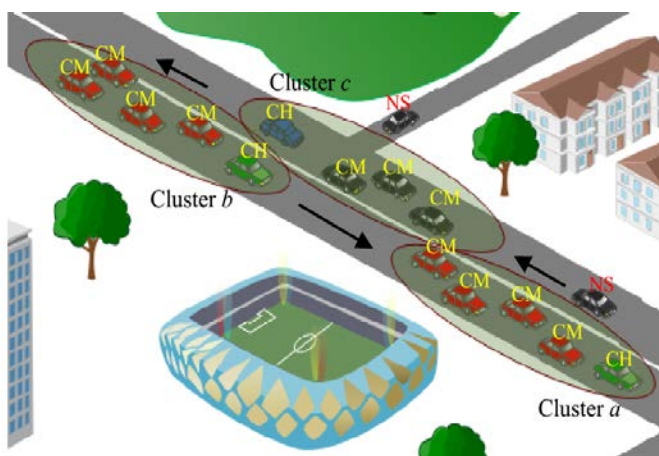

Fig. 1. An Example of a Cluster Network Topology.

Amjad Mehmood and et al. [19] have employed the flow of traffic knowledge in addition to using several metrics, like the degree of connectivity, the node position, direction, and speed variation to form stable clusters. The naïve Bayesian probabilistic estimation technique is used to enhance cluster stability and increase the CH lifetime. The proposed technique was compared with other algorithms and showed improvements in cluster and CH lifetime. Regardless of the efficiency of the ANTSC algorithm in selecting the CH and increasing the cluster lifetime, it is used for a particular scenario, so it was unclear whether it could be used in different scenarios. Moreover, the naïve Bayesian network probabilistic estimation requires real datasets for each zone, which makes it inapplicable in case lack of dataset.

The authors in [20], proposed a new clustering algorithm to select the most suitable CH based on FLC. A blend of several metrics was considered as inputs of the proposed cluster head selection algorithm, such as speed, distance, acceleration, and direction. The results showed that developed fuzzy logic (FL) based Cluster Head Selection Algorithm (CHSA) has increased the stability of CH and improved generally the performance in various scenarios in VANETs.

The Fuzzy-Based Cluster-Management System (FBCMS) has been proposed in [21]. Two models of this system have been created of this system, where each model has different parameters to select the most appropriate cluster head. The first model utilizes three parameters, which are the group speed, relative acceleration, and security as inputs of fuzzy logic, while the second model uses four metrics. Three of them are the same as the first model in addition to the degree of connectivity as the fourth parameter. However, using the location of vehicles in relation to a fixed RSU as one of the parameters to determine the CH in a highly dynamic environment like VANET could have a negative impact on the stability of clusters and may lead to frequent network disconnection, especially on highways.

The authors of [22] proposed a novel clustering scheme, which depends on the average speed of vehicles and standard deviation to increase the cluster lifetime. Two clustering patterns have been introduced which rely mainly on the principle of the normal (Gaussian) distribution and the relative speed. The calculated residence time of vehicular nodes in a cluster is used as a stability criterion. The first pattern represented a very high stable cluster in which the vehicles having speeds within the range of mean and standard deviation are used to configure this cluster (i.e., only 68% of the vehicles permitted to form this cluster). The election of the cluster head is carried out from the vehicles having speeds close to the average of cluster speed. The second pattern aims to group about 95% of the vehicles by selecting only the vehicles having speeds with a deviation lower than the double of the average standard deviations ($\sigma \leq 2\overline{\sigma}$) in one cluster. The analytical analysis showed that the second pattern has less stability than the first one. These two metrics (average speed and standard deviation) alone are not enough to establish stable clusters and select the optimal cluster heads for them, particularly as many parameters should have been taken like acceleration and position.

About video streaming and live road surveillance, the EVAC-AV [23] has been proposed as a solution for this kind of clustering. The cluster is initiated when a vehicle disseminates a request to join a cluster for having a live road surveillance service, so the vehicles which are ahead of it will be triggered to calculate their vision area. If their vision area is larger than a predefined vision area threshold, they will be deliberated as candidate CHs. Using the largest vision area as a single parameter to determine the best cluster head is not enough to create a stable cluster especially since the video streaming service is the most affected by the changes and re-clustering furthermore, the other algorithms which aim to provide stable clusters depend mainly on RSUs as a key parameter which makes it difficult to apply these algorithms in highways environments lacking to V2I technologies, therefore, this paper proposes a new stable V2V clustering algorithm entitled "A New Clustering Algorithm for Live Road Surveillance on Highways Based on DBSCAN and Fuzzy Logic". This algorithm aims to create stable clusters based on the density of vehicles on the street by using DBSCAN to form clusters and select the optimal CH based on FLC. DBSCAN ensures constructing the clusters without having to rely on RSUs as other algorithms do while FLC is used to select the best cluster head.

III. PROPOSED ALGORITHM

This paper presents a new algorithm that has the ability to detect and form a cluster automatically when the density of vehicles increases as well as selects the optimal CH. The strength of our algorithm is derived from the integration of the DBSCAN algorithm with fuzzy logic control. In our assumption, all vehicles are fitted with OBUs to be able to handle the IEEE802.11p as a Dedicated Short Range Communications (DSRC) system. Each vehicle broadcasts its information by sending Cooperative Awareness Message (CAM), which is a single hope broadcast communication. This CAM message is sent periodically at regular time intervals called T_{update}. Based on these messages, each vehicle will sense its current neighboring vehicles and update its *Neighbor Table* through exchanging the speed, position, and direction. Our proposed algorithm aims to provide a stable density-based clustering technique on a highway that consists of two phases cluster configuration phase and cluster head selection phase. Fig. 2 depicts our proposed algorithm.

A. Overview of DBSCAN

Clustering represents the most commonly used and more powerful unsupervised learning mechanism in machine learning. It is a useful tool that aims to classify the input data into sets depending on some similarity calculations. These algorithms are categorized into groups like partitional algorithms, density-based algorithms, hierarchical algorithms, etc. [24]. Among them, DBSCAN has been selected in our proposed algorithm because it has many features that make it more suitable than other clustering techniques. DBSCAN is an effective density-based clustering algorithm for spatial data systems due to its ability to discover clusters with arbitrary shapes in one scan, not like, for example, K-mean, which needs many iterations to find out the clusters. It is characterized by its capability to detect outliers as well as it

does not need to predetermine the number of clusters. In DBSCAN, the distance of two points is determined by a distance metric, such as the Euclidean distance. However, there are two parameters in DBSCAN which are required to be specified, Epsilon (ε) and Minimum Points (*MinPts*). ε represents the maximum distance between two points, which means that if the distance between two points is lower or equal ε, these points are considered neighbors. *MinPts* represents the minimum number of points counted neighbors for that point. It is used to identify if the point is a core point, border point, or noise point [25].

B. Vehicle Vision Area

Vision area plays an essential role in defining the cluster topology and electing the CH because the CH is responsible for providing live video surveillance to all vehicles located behind it that do not have enough vision area. No vehicle can be nominated to be a CH if it does not have a sufficient vision area. Therefore, the vehicles will be classified into two classes: (i) vehicles that have vision area (V_{vision}), which can be potential CHs; (ii) vehicles that do not have vision area $V_{no-vision}$, which can be possible CMs. We assume each vehicle has a camera mounted on the vehicle dashboard to capture live road conditions. The Distance Threshold (D_{th}) is used to define the vision area. We can say any vehicle is a V_{vision} if the distance between it and the adjacent front vehicle on the same lane is less than the D_{th}, but if the distance is less than D_{th}, it is considered a $V_{no-vision}$. It is worth noting that the D_{th} value is the same as the value of ε parameter used in the DBSCAN algorithm, which represents the safety distance that gives the driver the sufficient time for appropriate decision in case if he decides to overtake or in case of any emergency like a sudden incident.
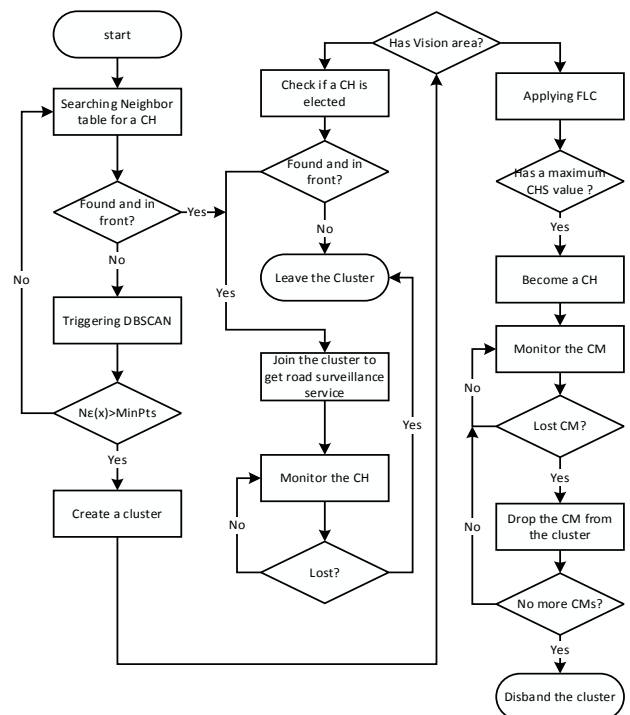

Fig. 2. Flowchart of Our Proposed Algorithm.

C. Cluster Configuration

At the beginning of clustering, each vehicle initializes its state to NS state, which means it does not belong to any cluster yet. Then, each vehicle shares its information by sending CAM message, which contains the speed, position, direction, and vision area. The *Neighbor Table* is updated periodically in a time interval called T_{update}. Each vehicle will filter the *Neighbor Table* and select only the vehicles in the same direction to prevent the vehicles on the reverse side or in a different street from shortly participating in the cluster. Referring to Fig. 2 DBSCAN will establish the cluster when it detects enough nearby vehicles close to each other based on ε and *MinPts*. Let us suppose the vehicle x moving on the highway does not have enough vision area. It will search in its *Neighbor Table* about a suitable CH. The suitable CH should be in front of it and should have enough vision area to provide efficient live video streaming about road conditions. If vehicle x does not find a suitable CH, it will filter the *Neighbor Table* and only collects the positions of vehicles in an NS, which have the same direction to form a dataset and then trigger the DBSCAN algorithm.

In DBSCAN, the vehicles are adjacent to the vehicle x, denoted by $N_\varepsilon(x)$ is defined by.

$$N_\varepsilon(x) = \{y \in D \mid \text{distance}(x, y) \le \varepsilon\}, \qquad (1)$$

Where y any vehicle in the dataset of vehicle x and D represents the DBSCAN dataset. Three types of nodes are defined in DBSCAN:

1) Core Node: The vehicle x is considered a core node if $|N_\varepsilon(x)| \ge MinPts$.

2) Border Node: The vehicle x is considered a border node if $|N_\varepsilon(x)| < MinPts$, but one of the $N_\varepsilon(x)$ is a core node.

3) Noise Node: The vehicle (x) is considered a noise node if $|N_\varepsilon(x)| < MinPts$ and no one of $N_\varepsilon(x)$ is a core node.

After DBSCAN has found a core node, the remaining adjacent vehicular nodes are checked consecutively to identify the next core node. If another nearby node becomes a core point, the cluster domain is extended. DBSCAN continues this process until no more core points can be found. Fig. 3 shows how the cluster is established after being originated by vehicle x. All the vehicles are included in the group, except the vehicle considered as a noise node. It should be noted that the vehicle in the center of the blue circle is a core node, and the vehicle in the center of the yellow circle is a border node, while the vehicle in the center of the red circle is a noise node. In case of existing a suitable CH, it will send a request to join the cluster and become a CM. In the case of more than one CH in front of it, it will select the closest one.

D. Cluster Head Selection (CHS) Phase

Cluster Head Selection (CHS) plays a significant role in cluster stability, which in turn represents one of the performance criteria in VANETs. The CHS process starts after cluster creation in which only V_{vision} vehicles in the cluster will enter to CHS phase as Candidate Cluster Heads (CCHs). FLC is the technique used to find the most suitable CH in the cluster. Fuzzy logic is an effective multi-characteristic decision technique because of its ability to a trade-off between significance and precision. Three parameters are considered in the CHS phase: Cluster Speed (*CS*), Vehicle Acceleration (*VAcc*), and Leadership Degree (*LD*). *CS* is determined by calculating the average speed of the vehicles in the clusters. *LD* is a value between 0 and 1 which shows if the CCH is eligible to be a CH or not, where 0 means that the CCH is in the back of the cluster (all potential CMs are in front of it), which means it is not eligible to be a CH while 1 represents that the CCH is at the front of the cluster and has the highest degree of eligibility to be a CH. The *LD* metric is calculated for each CCH so

$$LD(CCH) = \frac{\sum_{i=1}^{N} V_B}{\sum_{i=1}^{N} V_{no-vision}}$$

$$(2)$$

Where N is the number of vehicular nodes in the cluster, V_B is the number of $V_{no-vision}$ behind the CCH in the cluster. These three metrics are fuzzified using the fuzzy logic system. Fig. 4 shows our CHS System. As shown in this figure, three parameters (*CS, VAcc, LD*) are considered as an input of the fuzzification system. The function of the fuzzification system is to convert the actual values of the input parameters into fuzzy sets by using membership functions. There are many types of membership functions.

In our CHS system, we have utilized triangular and trapezoidal membership functions, as shown in Fig. 5 because they are more efficient in real-time applications [26]. The term sets of *CS*, *VAcc*, and *LD* are defined respectively as:
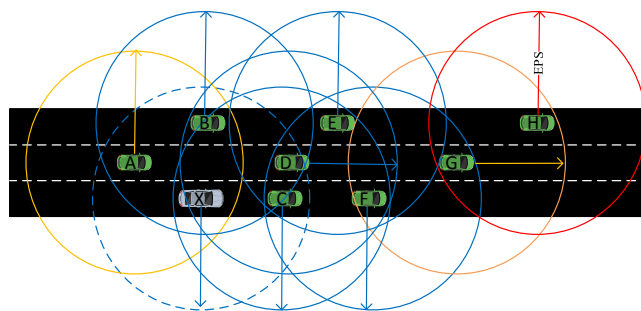

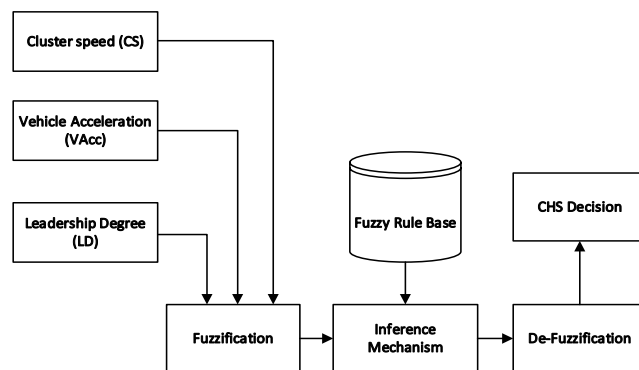Fig. 3. DBSCAN Algorithm Configuration.

Fig. 4. CHS System.

CS = {Low Speed, Normal Speed, High Speed}

= {LS, NS, HS}

VAcc = {Deceleration, Same Speed, Acceleration}

= {Dec, SS, Acc}

LD = {Low Degree, Normal Degree, High Degree}

= {LD, ND, HD}

CHS = {Very low, Low, Medium, High, Very High}

= {VL, L, M, H, VH}

The Fuzzy Rule Base (FRB) is built based on these membership functions. Since we have three inputs, and each input has three values, then we need 27 rules to cover all possible inputs.

$$X* = \frac{\sum_{j=1}^{27} x_j \left(\prod_{i=1}^{3} \mu(p_i) \right)}{\sum_{j=1}^{27} \left(\prod_{i=1}^{3} \mu(p_i) \right)}$$
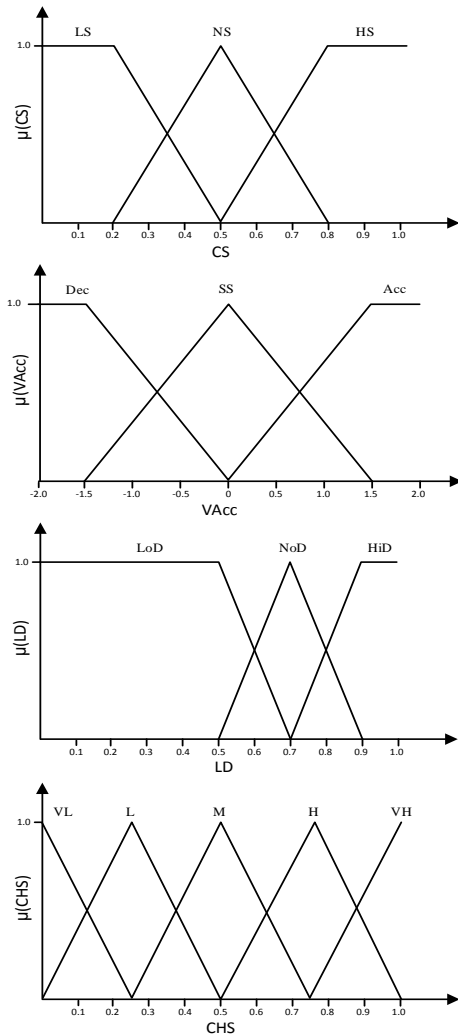
(3)

Fig. 5. Membership Functions of CHS System.

TABLE I. FUZZY RULE BASE OF CHS SYSTEM

Rule No.	*CS*	*VAcc*	*LD*	*CHS*
1	LS	Dec	LoD	VL
2	LS	Dec	NoD	VL
3	LS	Dec	HiD	L
4	LS	SS	LoD	VL
5	LS	SS	NoD	L
6	LS	SS	HiD	M
7	LS	Acc	LoD	VL
8	LS	Acc	NoD	M
9	LS	Acc	HiD	H
10	NS	Dec	LoD	VH
11	NS	Dec	NoD	L
12	NS	Dec	HiD	M
13	NS	SS	LoD	VL
14	NS	SS	NoD	HP
15	NS	SS	HiD	VH
16	NS	Acc	LoD	VL
17	NS	Acc	NoD	H
18	NS	Acc	HiD	M
19	HS	Dec	LoD	VL
20	HS	Dec	NoD	H
21	HS	Dec	HiD	VH
22	HS	SS	LoD	VL
23	HS	SS	NoD	M
24	HS	SS	HiD	H
25	HS	Acc	LoD	VL
26	HS	Acc	NoD	L
27	HS	Acc	HiD	VL

The CHS value ($X*$) is ranging between 0 and 1. Table I shows the Fuzzy rule base of the CHS system. After determining the CHS value ($X*$) to all CCHs in the cluster, the CCH with maximum CHS value will declare itself as a CH. All the vehicles behind the CH will send a *Request* message to get a video streaming service. When the CH receives the *Request* message, it sends *Accept* message back to them. After the vehicles receive *Accept* message, they will change their status from NS to CM. All the vehicles in front of the CH will leave the cluster.

E. Disbanding the Cluster

Referencing to Fig. 2, the CH will disband the cluster in two cases:

Case 1: When CH is blocked by other vehicles, so its vision area is less than the D_{th}, it will disband the cluster.

Case 2: When the CH discovers that all its CMs have left the cluster, it will disband the cluster.

F. Leaving the Cluster

Two conditions can cause the CMs to leave the cluster. Each CM monitors its link to its CH every T_{update}. If it has not received the CAM message from the CH for two T_{update}, the vehicle's link to its CH fails, and in this case, it will leave the cluster and change its state from CM to NS. The CM also

leaves the cluster when it overtakes the CH and becomes in front of it, so the video streaming service from the CH becomes useless. In this case, it will send a *Leave* message to its CH.

IV. Tools and Methodology

Our proposed clustering algorithm has been evaluated using MATLAB R2017b, while the mobility of vehicles has been simulated by the Simulation of Urban Mobility (SUMO). Our algorithm is designed for highway, so a 5 km highway is modeled in SUMO. The road consists of six lanes, three lanes for each direction in which 50 vehicles moving in the same direction with different speeds were deployed. The speed of vehicles is ranging from 60-120 km/h. The standard D_{th} is 50 m, which is the distance approved as safety distance between the vehicles when the speed of vehicles is 100 km/h [23]. In our work, we use different D_{th} to confirm the efficiency of our proposed algorithm. Concerning the DBSCAN parameters, we considered the number of the lane in the same direction represents the *Minpts*, and the safety distance represents the value of ε, so if the D_{th} is 50, then ε is 50 m. The main parameters applied in the simulation are mentioned in Table II. MATLAB and SUMO blocks have been connected together by TraCI (Traffic Control Interface). TraCI creates a TCP connection to make a connection between MATLAB and SUMO. SUMO acts as a server (TraCI-Server) and MATLAB as a client (TraCI-Client).

TABLE II. System Simulation Parameters

Parameters	Values
Highway Distance	5 Km
Simulation Time	100 sec
Number of vehicles	50
Distance threshold (D_{th})	(50,70) m
OBU transmitting range	300 m
Vehicles speed	60, 80,100,120 Km/h

V. Results

The performance evaluation of the proposed clustering algorithm was done by comparing our results with EVAC-AV algorithm results. It should be noted that we chose EVAC-AV introduced in the related work section as a benchmark algorithm because it is the only clustering algorithm based on vision area estimations and use V2V mode. The rest of the other algorithms have been excluded from the comparison because they differ in terms of purpose, parameters, and calibration. Our aim is to improve the performance and increase the stability of CHs as well as decreasing the number of vehicles status change. The following two performance metrics were used:

- **Average cluster lifetime**: It is defined as how long each cluster will last continuously. A higher value of this measure denotes a better stability.

- **Vehicles status change**: Vehicle status change is defined as the number of status changes per vehicle during its lifetime.

Fig. 6 and Fig. 7 demonstrate the distribution of CHs, CMs, and NSs during the simulation time when the D_{th} =50 m and 70 m, respectively.

They show a high number of NSs in each time step. The reason for this is that these vehicles still in NS because they have enough vision area and do not need to join to any cluster as well as they are moving in an area with a low density of vehicles, so they are detected as noise nodes (NSs) during the forming the cluster by DBSCAN algorithm. This is considered a virtue because our proposed algorithm aims to form clusters and provides stable CHs for only the $V_{no\text{-}vision}$ that need video streaming to know the road conditions; therefore we classified the NSs into $NS_{no\text{-}vision}$ and NS_{vision} to calculate the exact number of vehicles that do not have enough vision area in NS and do not find CH. Fig. 8 shows the number of NS_{vision} and $NS_{no\text{-}vision}$ during the simulation time. The results showed the percentage of remaining $NS_{no\text{-}vision}$ is 2% during the simulation time when the D_{th}=50 m (Fig. 8 (a)) while it is up to 8% when D_{th}=70 m (Fig. 8 (b)).

Our proposed algorithm is compared with EVAC-AV in the term of cluster lifetime and number of cluster. As shown in Fig. 9, the cluster lifetime was increased in comparison to the EVAC-AV algorithm when we applied different D_{th}s. The average cluster lifetime was (51 sec) while the average cluster lifetime of EVAC-AV is (18.84 sec) at D_{th}= 50 m during the simulation time. At D_{th} = 70m, the cluster lifetime of our algorithm is (50.57 sec) whereas, the EVAC_AV is (26.52 sec). Also, the number of clusters has been more than halved compared to the EVAC-AV algorithm at different D_{th}s, as shown in Fig. 10.

Regarding the vehicle's status change, it was measured by calculating the rate of total status changes of the vehicles and applying this step by using a different D_{th}s. The results have been compared with EVAC-AV under the same condition. As displayed in Fig. 11 and Fig. 12, our proposed algorithm shows better results by decreasing the rate of vehicle status change.
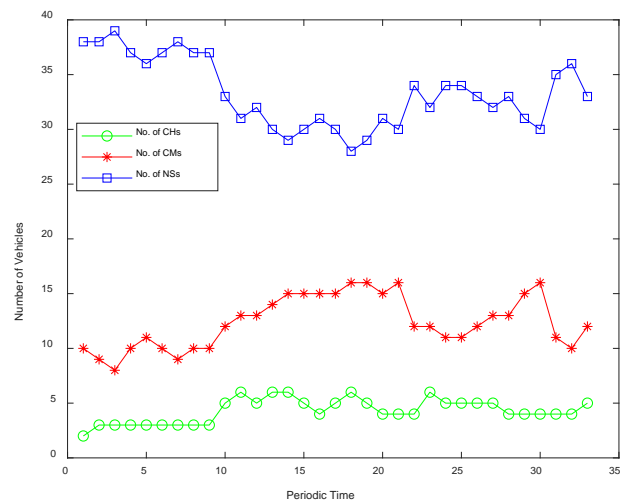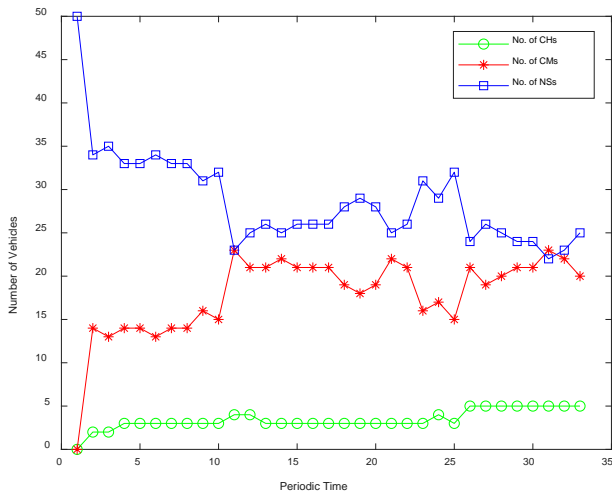

Fig. 6. Vehicles States at D_{th} = 50 m

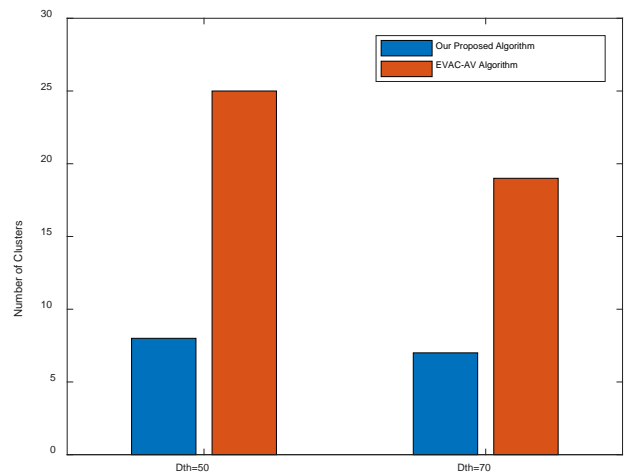Fig. 7.   Vehicles States at D_{th} = 70 m.
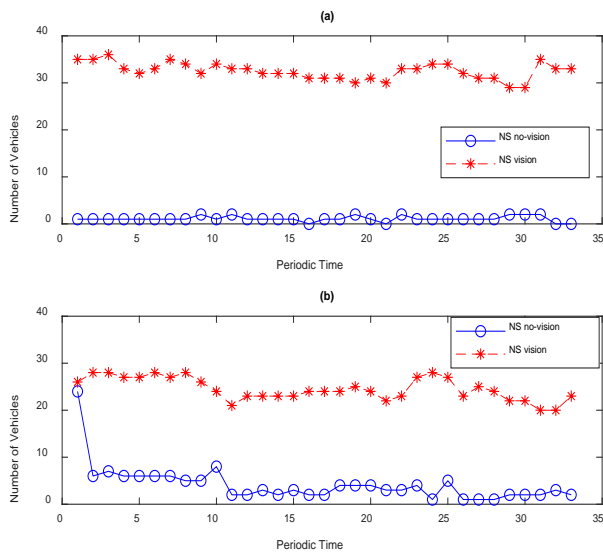

Fig. 8.   Distribution of NSs Vehicles According to Vision Area.

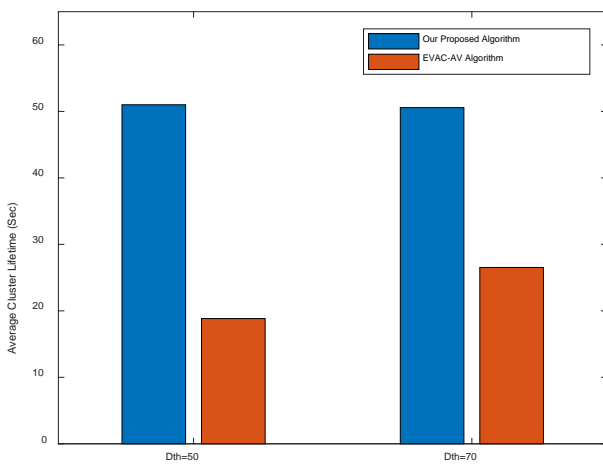Fig. 9.   Average Cluster Lifetime at different D_{th}s.

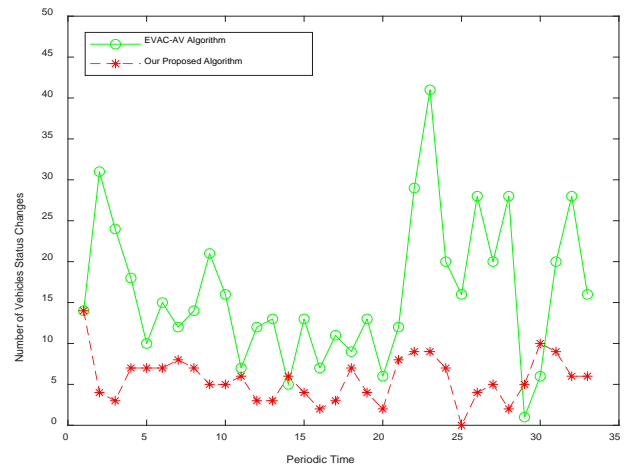Fig. 10. Number of Clusters at different D_{th}s.

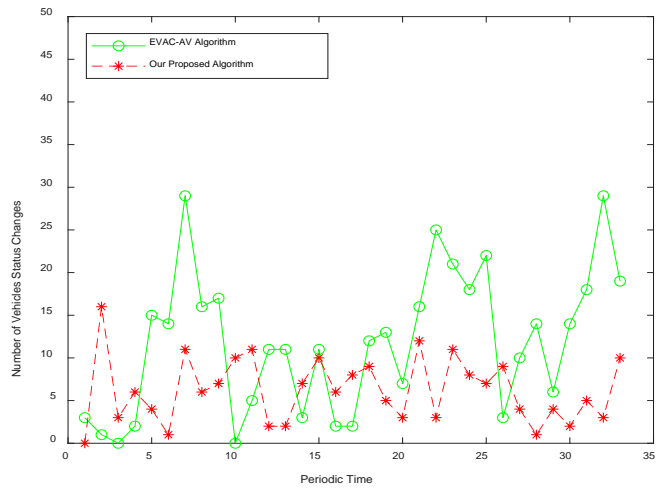Fig. 11. Vehicles Status Change at D_{th} = 50 m.

Fig. 12. Vehicles Status Change at D_{th} = 70 m.

VI. CONCLUSION

Video streaming enriches drivers with substantial information for safety, emergency, and entertainment. Clustering algorithms can be used as effective methods to improve and organize the work of the network. In this paper, a new clustering algorithm is proposed for road surveillance. It is characterized not only by the ability to detect and form the clusters when the density increases on the highway but also by finding the optimal CH for each cluster. The merits of this algorithm come from merging DBSCAN with FLC. DBSCAN algorithm is responsible for forming the cluster when it is triggered by a vehicle need video streaming to know its road condition. The CH is responsible for providing live road surveillance to all vehicles in the cluster. Fuzzy logic control is used to select the CH based on the metrics, which are cluster speed, acceleration, and leadership degree. Our proposed algorithm uses the vision area as a crucial parameter, so no vehicle can be nominated to be a CH if it does not have a sufficient vision area. Simulation results showed that our proposed algorithm provides a lower number of CHs and clusters and less variation of vehicle status. Additionally, our proposed algorithm increases the cluster lifetime in comparison with the EVAC-AV algorithm, which means the road surveillance service will be more efficient and more stable. In future work, we would like to add some additional metrics like link connectivity duration.

REFERENCES

[1] S. Al-Sultan, M. M. Al-Doori, A. H. Al-Bayatti, and H. Zedan, "A comprehensive survey on vehicular Ad Hoc network," J. Netw. Comput. Appl., vol. 37, no. 1, pp. 380–392, 2014.

[2] A. Tomatis, H. Menouar, and K. Roscher, "Forwarding in VANETs: GeoNetworking," in Vehicular ad hoc Networks: Standards, Solutions and Research, R. Scopigno, C. Campolo, and A. Molinaro, Eds. Springer, 2015, pp. 221–251.

[3] R. Ghebleh, "A comparative classification of information dissemination approaches in vehicular ad hoc networks from distinctive viewpoints: A survey," Comput. Networks, vol. 131, pp. 15–37, 2018.

[4] K. S. Eunice and I. Juvanna, "Secured Multi-Hop Clustering Protocol for Location-based Routing in VANETs," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 4, pp. 121–126, 2019.

[5] A. Alharbi, "Survey on Homomorphic Encryption and Address of New Trend," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 7, pp. 618–626, 2020.

[6] I. T. Abdel-Halim and H. M. A. Fahmy, "Prediction-based protocols for vehicular Ad Hoc Networks: Survey and taxonomy," Comput. Networks, vol. 130, pp. 34–50, 2018.

[7] F. Cunha et al., "Data communication in VANETs: Protocols, applications, and challenges," Ad Hoc Networks, vol. 44, pp. 90–103, 2016.

[8] P. Y. Lai, C. R. Dow, and Y. Y. Chang, "Rapid-Response Framework for Defensive Driving Based on Internet of Vehicles Using Message-Oriented Middleware," IEEE Access, vol. 6, pp. 18548–18560, 2018.

[9] T. Phakathi, F. Lugayizi, B. Isong, and N. Gasela, "Quality of Service of Video Streaming in Vehicular Adhoc Networks: Performance Analysis," Proc. - 2016 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2016, pp. 886–891, 2017.

[10] R. An, Z. Liu, and Y. Ji, "SVC-based cooperative video streaming in highway vehicular networks," Proc. - 31st IEEE Int. Conf. Adv. Inf. Netw. Appl. Work. WAINA 2017, pp. 216–221, 2017.

[11] L. Cai, H. He, L. Sun, A. Huang, and H. Shan, "Channel Allocation for Adaptive Video Streaming in Vehicular Networks," IEEE Trans. Veh. Technol., vol. 66, no. 1, pp. 1–1, 2016.

[12] E. B. Smida, S. G. Fantar, and H. Youssef, "Video streaming challenges over vehicular ad-hoc networks in smart cities," 2017 Int. Conf. Smart, Monit. Control. Cities, SM2C 2017, pp. 12–16, 2017.

[13] C. Cooper, D. Franklin, M. Ros, F. Safaei, and M. Abolhasan, "A Comparative Survey of VANET Clustering Techniques," IEEE Commun. Surv. Tutorials, vol. 19, no. 1, pp. 657–681, 2017.

[14] A. H. Abbas, M. I. Habelalmateen, L. Audah, and N. A. M. Alduais, "A Novel Intelligent Cluster-Head (ICH) to Mitigate the Handover Problem of Clustering in VANETs," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 194–203, 2019.

[15] W. Qi, Q. Song, X. Wang, and Z. Ning, "SDN-Enabled Social-Aware Clustering in 5G-VANET Systems," IEEE Access, vol. 6, 2018.

[16] B. Hassanabadi, C. Shea, L. Zhang, and S. Valaee, "Clustering in Vehicular Ad Hoc Networks using Affinity Propagation," Ad Hoc Networks, vol. 13, no. PART B, pp. 535–548, 2014.

[17] F. Yang, Z. Lin, and Y. Tang, "A Traffic Flow Based Clustering Scheme for VANETs," Sensors & Transducers, vol. 180, no. 10, pp. 110–116, 2014.

[18] A. A. Khan, M. Abolhasan, and W. Ni, "An evolutionary game theoretic approach for stable and optimized clustering in vanets," IEEE Trans. Veh. Technol., vol. 67, no. 5, pp. 4501–4513, 2018.

[19] A. Mehmood et al., "ANTSC: An Intelligent Naïve Bayesian Probabilistic Estimation Practice for Traffic Flow to Form Stable Clustering in VANET," IEEE Access, vol. 6, pp. 4452–4461, 2017.

[20] A. ÇALHAN, "A fuzzy logic based clustering strategy for improving vehicular ad-hoc network performance," Proc. Eng. Sci., vol. 40, no. 2, pp. 351–367, 2015.

[21] K. Ozera, K. Bylykbashi, Y. Liu, and L. Barolli, "A fuzzy-based approach for cluster management in VANETs: Performance evaluation for two fuzzy-based systems," Internet of Things, vol. 3–4, pp. 120–133, 2018.

[22] M. S. Talib, A. Hassan, B. Hussin, Z. A. Abas, Z. Saad, and Z. Sabah, "A Novel Stable Clustering Approach based on Gaussian Distribution and Relative Velocity in VANETs," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 4, pp. 216–220, 2018.

[23] C. M. Huang, H. L. Wang, H. Zhou, S. Xu, and D. Ren, "EVAC-AV: The live road surveillance control scheme using an effective-vision-area-based clustering algorithm with the adaptive video-streaming technique," IEEE Syst. J., vol. 11, no. 3, pp. 1228–1238, 2017.

[24] M. Ester, "Density-Based Clustering," in Data Clustering Algorithms and Applications, Minnesota: Crc, Hall, 2014, pp. 111–145.

[25] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in Kdd, 1996, vol. 96, no. 34, pp. 226–231.

[26] R. A. Shaikh, "Fuzzy Risk-based Decision Method for Vehicular Ad Hoc Networks," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 9, pp. 54–62, 2016.

Towards an Integrated Model of Data Governance and Integration for the Implementation of Digital Transformation Processes in the Saudi Universities

Abdulfattah Omar[1] (✉)

Department of English
College of Sciences & Humanities
Prince Sattam Bin Abdulaziz University

Ahmed Almaghthawi[2]

Department of Computer Science and Artificial Intelligence
University of Jeddah
Jeddah, Saudi Arabia

Abstract—In the face of the challenges of the Digital Age and the outbreak of the pandemic COVID-19 which have changed higher education institutions remarkably, universities are urgently required to speed up their digitalization initiatives and cope up with the global digital developments to survive. For the successful implementation of digital transformation, however, data governance should be considered. Despite the extensive literature on data governance and digital transformation, there is no focus on the issue in the Saudi Higher Education institutions. In the face of this, the current study investigates data governance policies and practices in the Saudi Universities. This study is built on a case study design. Nine universities in Saudi Arabia were selected for the purpose of the study. These include public, community, and private universities that make up the Higher Education system in Saudi Arabia. For data collection purposes, three methods were selected. These were the survey, focus group discussions and in-depth interviews. The findings of this study indicate that data governance is an effective tool in the implementation of digital transformation processes in higher education institutions and thus should be embedded into strategies of the universities for using digital technologies in appropriate manners. Good data governance practices are required for smooth and effective digital transformation. Universities are required to create an effective functional team for the data governance tasks, develop an internal audit of data governance, follow-up the regulatory compliance procedures, define the priorities of data governance activities, provide frequent data governance training for employees and faculty members, set enforcement and follow-up standards, and conduct frequent assessments of data governance plans and policies. Although the study is limited to the Saudi universities, results and implications of this study can be a useful reference to choose effective data governance practices that can be successfully adopted and implemented to effectively manage critical information and use it to transform their day-to-day operations.

Keywords—*COVID-19; data governance; digital transformation; higher education; Saudi Arabia*

I. INTRODUCTION

The recent years have witnessed unprecedentedly rapid changes that have their impact on the global economy today [1, 2]. In response to these changes, businesses, corporations, institutions, and even governments have realized the importance of adopting a change culture that enables them to be more willing to implement required changes and initiatives effectively and efficiently so that they can survive and remain competitive within this fast-changing business environment [3]. In this regard, different digital transformation initiatives have been adopted by organizations and institutions in both public and private sectors in different countries around the world. It is almost agreed that digital transformation is imperative for optimizing processes and making workflows faster, easier, and more efficient [4-6].

The Kingdom of Saudi Arabia has recently adopted some initiatives in relation to the digital transformation of the government sectors including education, healthcare, and telecommunications. These initiatives seek to improve operations and business value with the ultimate goal of achieving sustainable development and global effectiveness, increase the contribution of the digital economy to GDP, and improve the overall quality of life. The Digital Transformation national plan was designed to be carried out over three stages. The first stage lasted from 2005 to 2010; the second from 2012 through 2016; the third, currently in effect, began in 2019 and to be completed roughly in 2022; its main components are digital health, digital education, e-commerce, and smart cities.

In education, great efforts have been recently made in order to achieve the goals of the digital transformation clearly stated in the strategic objectives of the Saudi Vision 2030. In this regard, the Ministry of Education launched the Future Gate project for implementing digital transformation processes in all the educational institutions in the Kingdom. The project seeks to convert paper curricula into digital, and work to convert traditional classes into smart ones. The project was launched in collaboration with the TETCO Company for educational technologies in 2018.

The project is considered by many as one of the most promising projects in the field of education, especially with the rapid growth of its beneficiaries. It aims to create a good and enjoyable learning environment in order to reduce the Kingdom's traditional environment for education. With digital education, a positive interaction can be created between students on the one hand, and students and instructors on the

other by means of modifying the traditional style of education, and replacing it with digital education.

With the emergence of the Coronavirus crisis known as COVID-19, it was clear that the shift into digital transformation in education is imperative. Educators stressed the importance of developing a well-built infrastructure that can accommodate and facilitate the digital transformation processes in education. Taken the consequences of COVID-19 into consideration, the digital transformation of higher education has to be accelerated [7]. It is now important to accelerate digital transformation processes and these cannot be delayed in any way [8].

For the successful implementation of digital transformation processes, however, data governance policies and procedures need to be developed and enforced [9]. In spite of the importance of the issue, very little has been done on data governance in the Saudi institutions in general and higher education institutions and universities in particular. This study addresses this gap in the literature through proposing a model for data governance in the Higher Education institutions in Saudi Arabia. In order to achieve the study objectives, a questionnaire survey was designed and in-depth interviews and discussion focus-groups were conducted with the IT departments in 7 Saudi Higher Education institutions and universities. A mixed approach of quantitative and qualitative methods was adopted in order to evaluate the data governance policies and practices adopted by the selected institutions.

The remainder of this article is organized as follows. Section 2 is a brief survey on the emergence of digital transformation and the relationship between digital transformation on one side and data governance and integration on the other. Section 3 describes the methods and procedures. Section 4 is an analysis of the data. Section 5 summarizes the findings of the study and discusses their implications to higher education institutions.

II. LITERATURE REVIEW

Digital transformation refers simply to change processes whereby a shift is brought about so that technical and digital developments are replacing traditional and conventional methods for faster and better performance [10, 11]. Digital transformation provides huge potentials to build effective, competitive and sustainable societies, by achieving a fundamental change in the services of various parties, consumers, employees and beneficiaries, while improving their experiences and productivity through a series of commensurate processes, accompanied by a reformulation of the necessary procedures for activation and implementation [11, 12]. Digital transformation requires enabling a culture of creativity as well as a well-built infrastructure that accommodates any required changes in the work environment, from infrastructure, operating models, to marketing services and products [9, 13, 14].

According to Fenton, Fletcher, and Griffiths, digital transformation has the potentials of reshaping the way people live, work, think, interact and communicate with people, depending on available technologies, with continuous planning and a constant quest to reformulate practical experiences [15].

In this way, successful digital transformation processes can improve efficiency, reduce spending, and implement new services quickly and flexibly. They have the effect of achieving a fundamental change in the services provided to individuals in different fields including health, education, safety and security, and improving their experiences and productivity [13].

The literature suggests that digital transformation has been key in almost all innovation and change processes adopted by institutions and organizations due to its positive impacts on increasing workflow efficiency and reducing error, improving performance productivity, and quality, and thus enhancing customer satisfaction [14, 16].

In order for successful implementation of digital transformation, however, some factors need to be considered [17, 18]. These include data governance and integration. In fact, data is one of the most critical elements in all digitalization processes; therefore, it must be correct, complete, and accurate [19]. Practically every process within an organization relies on data to operate effectively, so the organization's strategy of data governance impacts the entire organization, not just the IT departments. Leignel, Ungaro, and Staar argue institutions should have effective, working and reliable data governance practices and policies in order to face the challenges of digital transformation. Through such practices and policies, organizations can effectively manage their investment in IT systems and make the most of their development [20].

Data governance can be defined as the concept of managing data and information to produce high-quality data, clean, accurate, and safe data [21, 22]. It is an umbrella term that encompasses processes of the creation, storage, use, evaluation, archiving and deletion of information [23, 24]. It includes processes, roles, policies and standards that ensure the effective and efficient use of information, and in that way it enables institutions to achieve their goals effectively and efficiently [21].

Data governance is based on many processes that are associated with the life cycle of information and data. These include data creation, data organization, data storage, data backup and recovery, data management and maintenance, and finally data retention and secure destruction. Organizations need to have a comprehensive and integrated data governance strategy that helps them maximize the value of their digital information assets while minimizing potential risks. In this sense, data governance is designed to guide records management and corporate content management, not replace it, while also provide a framework for making smarter decisions regarding valuable information and where it is best stored within the IT infrastructure [19].

Based on the critical role of data in digitalization processes, it is almost agreed that data governance is indispensable to digital transformation processes for all organizations including universities today [20]. Given the tremendous growth in the amount of data that universities today create, collect and store, the concept of data governance has increased in recent years. Universities have realized the need to establish laws governing this vast amount of data and information in addition to their

fear from the risk of security, privacy, compliance and legal violations.

Universities are also encouraged to adopt working data governance policies in order to make use of what has come to be known as Big Data - which can represent great opportunities for universities and educational institutions if they are properly employed. Through effective data governance practices, universities and educational institutions can move faster, make smarter decisions, and provide better insights into consumer behavior, which will contribute to increased efficiency. Before the emergence of what was known as big data, decisions regarding how to store data were mainly made by the employees and individuals who created and used the data. Many universities have realized that they need a more structured structure and central control that oversees how data is stored, managed, kept and protected from violation whether unintentionally or by penetration of sensitive information.

With the increasing importance of data in almost all university operations today, data governance remains a major issue. Nevertheless, universities still have difficulties in terms of data management which definitely has negative impacts on its processes. These difficulties are generally attributed to poor data governance. In light of this argument, this study seeks to propose an integrated model of data governance for successful data management and proper implementation of digital transformation processes in the Saudi Universities.

III. METHODS

This study is built on a case study design. Nine universities in Saudi Arabia were selected for the purpose of the study. These include public, community, and private universities that make up the Higher Education system in Saudi Arabia. The selected institutions represent the three categories. For data collection purposes, three methods were selected. These were the survey, focus group discussions and in-depth interviews. For data analysis, a sequential mixed method design using a combination of quantitative and qualitative approaches.

The employees, faculty members, executives and stakeholders in this study represent the population. The sampling was done in the form of survey participants and interviewees, which was an appropriate selection from the whole population. It is always difficult to include the whole population, hence a typical sample was only chosen for the purpose of generalizing the result to the population. For data representativeness reasons, the survey was publicized through all possible means in order to reach a maximum number of employees and stakeholders using the universities' services so that the findings and results should be interpreted and generalized as representing the views of employees and stakeholders. 477 responded to the survey. These represent different employment levels and stakeholders. In the qualitative phase of this study, six senior members (two of each institution), holding the positions of manager and above levels, were interviewed and 27 other employees were sampled for the informal focus groups as shown in Table I. All the participants in the interviews and discussion groups had also responded to the survey so it was easy to triangulate the findings of the survey with the observations during the interviews and focus groups. The data collected through a survey questionnaire was

triangulated through transcripts of the Focus groups, interviews, as well as institutional documentation and website material including its annual reports and shareholders' feedback. The data findings were underpinned with the established theories to determine whether the Saudi universities have the potentials to make the shift towards digital transformation smoothly and effectively. Interview and Focus Group Questions were both close-ended and open-ended in which very short text answers were expected.

TABLE I.   DETAILS ABOUT THE DATA SOURCES

Institutions	Survey	Focus Group	Interviews
General		-	-
Top management	60	3	3
IT staff	105	9	3
Employees & Faculty	240	9	-
Other stakeholders	72	6	-
Total	477	27	6

The questionnaire is in three main parts. In Part One, respondents were asked about their perceptions about data governance practices in their institution, how employees and stakeholders can be educated about data governance policies, and how the institutions can ensure the efficiency of the digital transformation processes in their institutions. In Section Two, the participants were also asked about data governance principles followed in their institutions, their frequency and employees' opinions about those principles. In Section Three, participants were asked whether good data governance practices had been in place to facilitate digital transformation processes.

The questions prepared for the interviews and focus groups were only guiding questions (about the subject of the study), but during the interview and discussion, a few probe questions were also added and supplemented to the main questions. This helped the researchers to cope with the respondents' stories and to get thick and detailed descriptions. All the participants had an opportunity to talk about their experiences in relation to data governance principles. Their narratives indicated their understanding of the data governance and its benefits which the researchers used in identifying the themes and subthemes for its content analysis.

IV. ANALYSIS AND DISCUSSIONS

The data collected from the questionnaire revealed interesting information about the application and implementation of data governance in the Saudi universities. The employees, faculty members, executives and stakeholders showed great familiarity with the principles of data governance. The data also revealed that data governance was deeply embedded in the universities' portfolios. This information was triangulated through the data collected from interviews and focus groups. The participants revealed almost a similar account of data governance being followed at their institutions and how these practices contributed to making the shift towards digital transformation smoothly and effectively. The researchers inferred from both the questionnaire of data

and the narratives of participants that they had familiarized themselves with both data governance principles and also the manner they were being implemented and monitored. The interview respondents admitted that during the learning and training programs, they involved in self-reflection and collaborative discussions with one another for a better understanding of the data governance principles. Respondents also realize the importance of data governance practices and policies for a smooth and effective shift towards digitalization. Nevertheless, it is also clear that universities still have difficulties in terms of dealing with big data and making use of these data in ways that have positive impacts on the performance and quality of the organizations. Some executives reported that only 11% of their data is appropriately used. This means that approximately 90 percent of the digital information is stored as "dark data", unorganized, undisclosed and untapped.

The findings of the study agree with the data governance literature in relation to the effective role data governance can play in the implementation of digital transformation processes. Successful organizations have to adopt working data governance policies that ensure the effective creation of data and dealing with these data in an appropriate manner [18, 25]. The claim here is that organizations which do not have a governance strategy face critical risks while those which work within a governance framework perform measurably better [26]. It is also true that data governance and management are closely related to the performance of institutions and organizations. Big data should be considered as opportunities for growth, performance and competitiveness [27, 28]. Ramingwong and Manopiniwes stress that data governance and data security management are critical success factors for the productivity and survival of industries and organizations today [29].

In light of these findings, it is suggested that universities need to develop a virtual culture to meet the challenges of the digital age we witness today. This, however, requires both imaginative and creative implementation, as well as open and innovative leadership. Universities should devise strategic plans for digital transformation and that includes the selection of appropriate technologies taking into account training and experience. The components of these strategic plans are shown in Fig. 1.
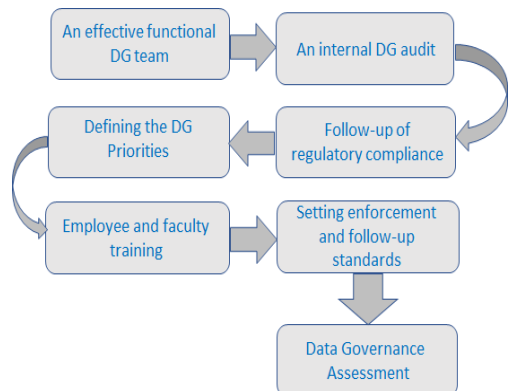

Fig. 1. A Proposed Model for Good Data Governance Practices in the Saudi Universities.

In the first place, each university should create an effective functional team for data governance tasks. No data governance plan will succeed if it does not reflect the needs and goals of all major stakeholders, including legal entities, compliance officers, risk management, human resources, information technology, and others [21, 30]. Each of these entities should be represented at the initial planning stage and have a say in determining success factors, key outcomes and potential risks.

Second, universities should have an internal audit for data governance that can be usefully used for developing a robust information management framework [23, 31, 32]. This should have the effect of defining and identifying the main data sources universities interact with including outdated systems, and data archives that are potentially not actively managed at all.

Universities also need to carefully evaluate and follow-up the legal and regulatory requirements of data governance practices. It is extremely important to understand all external retention requirements, decide the duration data should remain and how long it will be kept and review these requirements often to ensure that the governance plan will be kept up to date [31]. Equally important, universities should define the priorities of data governance activities [33, 34]. Information governance plans are not created to be implemented overnight, so it is important to address the most urgent issues first. Data evaluation should define the priorities of primary information governance activities. For example, if a review reveals a large number of decentralized data stored randomly across a set of shared drives, a good first step might be to create a data map that connects employees and faculty members to the data sources they interact with frequently. Likewise, if the initial evaluation reveals the accumulation of a large number of unnecessary backup files, the data governance plan should focus more on developing and implementing a better deletion policy.

It is important for universities also to provide training programs for employees and faculty members on the best data governance practices since the successful implementation of digitalization processes relies heavily on the workforce. Training should not be confined to IT departments and personnel only [35, 36]. Employees and faculty members should understand how data governance policies and procedures affect their daily activities. It is also important to explain for them the importance of data and data governance policies and their role in the digital transformation processes [36-39]. Training programs should also be supplemented with enforcement and follow-up standards that ensure the implementation of data governance programs [21, 40]. Finally, universities need to conduct periodic and frequent assessments of data governance plans, activities, and policies. Assessments should consider the way the effectiveness of data governance programs can be proved and whether the data governance goals and objectives are achieved [22, 30].

V. CONCLUSION

The outbreak of the pandemic COVID-19 has accelerated the digitalization developments in many fields including higher education institutions. Universities now are urgently required to speed up their digitalization initiatives and cope up with the

global digital developments in order to survive. For the successful implementation of digital transformation, however, data governance should be considered. This study addressed the nexus between data governance and digital transformation in higher education institutions. The findings of this study indicate that data governance is an effective tool in the implementation of digital transformation processes in higher education institutions and thus should be embedded into strategies of the universities for using digital technologies in appropriate manners. Good data governance practices are required for smooth and effective digital transformation.

Universities are required to create an effective functional team for the data governance tasks, develop an internal audit of data governance, follow-up the regulatory compliance procedures, define the priorities of data governance activities, provide frequent data governance training for employees and faculty members, set enforcement and follow-up standards, and conduct frequent assessments of data governance plans and policies. Although the study is limited to the Saudi universities, results and implications of this study can be a useful reference to choose effective data governance practices that can be successfully adopted and implemented to effectively manage critical information and use it to transform their day-to-day operations. Further research is required to provide higher education institutions with workable and reliable solutions for the proper management of big data and the assurance of data quality for improving operational efficiency.

REFERENCES

[1] S. Furusten, Institutional Theory and Organizational Change. Edward Elgar Publishing, Incorporated, 2013.

[2] P. Myers, S. Hulks, and L. Wiggins, Organizational Change: Perspectives on Theory and Practice. OUP Oxford, 2012.

[3] D. C. Dunphy, A. Griffiths, and S. Benn, Organizational Change for Corporate Sustainability: A Guide for Leaders and Change Agents of the Future. Routledge, 2003.

[4] G. Westerman and D. Bonnet, Leading Digital: Turning Technology into Business Transformation. Harvard Business Review Press, 2014.

[5] N. Palmer et al., Digital Transformation with Business Process Management. Future Strategies Inc., 2019.

[6] S. Shariffuddin and J. Razali, "Transformation of university colleges to full-pledged universities: A proposed conceptual framework for Malaysian higher learning institutions," International Journal of Advanced and Applied Sciences, vol. 4, no. 12, pp. 168-173, 2017.

[7] S. Martin-Barbero, "COVID-19 has accelerated the digital transformation of higher education," World Economic Forum21 Jul 2020 2020, Available: https://www.weforum.org/agenda/2020/07/covid-19-digital-transformation-higher-education/.

[8] M. D. P. Uys, One World One School: Education's Forthcoming Fundamental Transformation. Independently Published, 2020.

[9] A. Maheshwari, Digital Transformation: Building Intelligent Enterprises. Wiley, 2019.

[10] T. U. Daim, Digital Transformation: Evaluating Emerging Technologies. World Scientific Publishing Company Pte Limited, 2020.

[11] B. Daniotti, M. Gianinetto, and S. D. Torre, Digital Transformation of the Design, Construction and Management Processes of the Built Environment. Springer International Publishing, 2019.

[12] A. Gilchrist, Digital Success: A Holistic Approach to Digital Transformation for Enterprises and Manufacturers. alasdair gilchrist, 2018.

[13] V. Kale, Digital Transformation of Enterprise Architecture. CRC Press, 2019.

[14] D. R. A. Schallmo and C. A. Williams, Digital Transformation Now!: Guiding the Successful Digitalization of Your Business Model. Springer International Publishing, 2018.

[15] A. Fenton, G. Fletcher, and M. Griffiths, Strategic Digital Transformation: A Results-Driven Approach. Taylor & Francis, 2019.

[16] T. Saldanha, Why Digital Transformations Fail: The Surprising Disciplines of How to Take Off and Stay Ahead. Berrett-Koehler Publishers, 2019.

[17] B. Borden, C. Fudge, J. Nelson, J. Porell, and I. Redbooks, Accelerating Digital Transformation on Z Using Data Virtualization. IBM Redbooks, 2018.

[18] P. Jackson and C. Carruthers, Data Driven Business Transformation: How to Disrupt, Innovate and Stay Ahead of the Competition. Wiley, 2019.

[19] J. M. Barker, Data Governance: The Missing Approach to Improving Data Quality. University of Phoenix, 2016.

[20] J. L. Leignel, T. Ungaro, and A. Staar, Digital Transformation: Information System Governance. Wiley, 2016.

[21] J. Ladley, Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program. Elsevier Science, 2012.

[22] U. Gupta and S. Cannon, A Practitioner's Guide to Data Governance: A Case-Based Approach. Emerald Publishing Limited, 2020.

[23] H. Sen, Data Governance: Perspectives and Practices. Technics Publications, 2019.

[24] B. van Gils, Data Management: a gentle introduction: Balancing theory and practice. Van Haren Publishing, 2020.

[25] F. P. G. Márquez and B. Lev, Big Data Management. Springer International Publishing, 2016.

[26] Y. Y. Haimes and A. P. Sage, Risk Modeling, Assessment, and Management. Wiley, 2015.

[27] Y. Yuan and C. Ho, "Needs analysis of national parks for applying big data solutions in tourism management," International Journal of Advanced and Applied Sciences, vol. 4, no. 12, pp. 20-36, 2017.

[28] W. Lee, A. Chong, and T. Ramayah, "Organizational culture and performance of Malaysian manufacturing firms," International Journal of Advanced and Applied Sciences, vol. 5, no. 12, pp. 59-66, 2018.

[29] S. Ramingwong and W. Manopiniwes, "Supportmentfor organization and management competences of ASEAN community and European Union toward Industry 4.0," International Journal of Advanced and Applied Sciences, vol. 6, no. 3, pp. 96-101, 2019.

[30] D. Cervo, M. Allen, and J. Dyché, Master Data Management in Practice: Achieving True Customer MDM. Wiley, 2011.

[31] R. D. McDowall, Data Integrity and Data Governance: Practical Implementation in Regulated Laboratories. Royal Society of Chemistry, 2018.

[32] H. Elbardan and A. O. Kholeif, Enterprise Resource Planning, Corporate Governance and Internal Auditing: An Institutional Perspective. Springer International Publishing, 2017.

[33] D. B. A. M. Khosrow-Pour, Encyclopedia of Information Science and Technology, Fourth Edition. IGI Global, 2017.

[34] D. B. A. M. Khosrow-Pour, Advanced Methodologies and Technologies in Business Operations and Management. IGI Global, 2018.

[35] N. Bhansali, Data Governance: Creating Value from Information Assets. CRC Press, 2013.

[36] E. M. Power and R. L. Trope, Sailing in Dangerous Waters: A Director's Guide to Data Governance. American Bar Association, Section of Business Law, 2005.

[37] S. K. Strydom and M. Strydom, Big Data Governance and Perspectives in Knowledge Management. IGI Global, 2018.

[38] R. Mahanti, Data Quality: Dimensions, Measurement, Strategy, Management, and Governance. ASQ Quality Press, 2019.

[39] A. D. Giordano, Performing Information Governance: A Step-by-step Guide to Making Information Governance Work. Pearson Education, 2014.

[40] R. S. Seiner, Non-Invasive Data Governance: The Path of Least Resistance and Greatest Success. Technics Publications, 2014.

Prudently Secure Information Theoretic LSB Steganography for Digital Grayscale Images

Khan Farhan Rafat
Riphah Institute of Systems Engineering (RISE)
Islamabad, Pakistan

Abstract—The endangerment of online data breaches calls for exploring new and enhancing existing sneaky ways of clandestine communication to tailor those to match the present and futuristic technological and environmental needs, to which malicious intruders wouldn't have an answer. Cryptography and Steganography are the two distinct techniques that, for long, have remained priority choices for hiding vital information from the unauthorized. But the visibility of the encrypted contents makes these vulnerable to attack. Also, the recent legislative protection agreed to law enforcement authorities in Australia to sneak into pre-shared cryptographic secret keys (PSKs) shall have a devastating impact on the privacy of the people. Hence, the need of the hour is to veil in the encrypted data underneath the cover of Steganography, whose sole intent is to hide the very existence of information. This research endeavor enhances one of the most famous images Steganography technique called the Least Significant Bit (LSB) Steganography, from the security and information-theoretic standpoint by taking a known-cover and known-message attack scenario. The explicit proclamation of this research endeavor is that the security of LSB Steganography lies in inducing uncertainty at the time of bit embedding process. The test results rendered by the proposed methodology confers on the non-detectability and imperceptibility of the confidential information along with its strong resistance against LSB Steganalysis techniques.

Keywords—*Clandestine communication; covert channel; hiding data in plain sight; inveil communication; LSB steganography*

I. INTRODUCTION

The world is undergoing and witnessing significant information technology shift from a paper-based environment to a green, paperless digitized environment [1]. This drifting digitized world is called the Internet of Things (IoT) that constitutes a multitude of technologies, such as vehicles, locomotives, traffic lights, cameras, televisions, satellites, sensors, therapeutic apparatus, the drones, and smartphones, just to name some [2]. Today's Hi-Tec technological gadgetry is not only enchanting but also inebriating. Accustomed to the Hi-Tec tools, people now carry these everywhere in their journey across the world. Round the clock networking, coupled with allied Cloud computing, enables them to remain in contact with their loved ones besides handling appointments of significance and doing business on the fly, simultaneously [3], [4]. This fact is apparent from Fig. 1, which expounds on the dependency and reliance of the people on their smart digital gadgetry.

The situation above, though, speaks high of the technological cum digital revolution, indeed poses a higher risk of people falling prey to the oblivious enemy who continuously is keeping them under surveillance to invade their privacy [5]. It is but only the unawareness as regards breaches through user's unconscious sharing of personal data on social media that Gartner, in its data security breach report stated "cyber vulnerability" as one of the three significant areas of concern for the year 2020 [6].

A. Privacy

The word "privacy" refers to the degree or extent up to which people willingly share their private information with others [7]. However, the digitized world of today finds it cumbersome to preserve user's privacy from some intentional malevolent intrusion that has further amplified to many folds because of cutting edge technology and the skills with which the opponents have augmented themselves [8]. The realization of one's privacy dates back to the history of human civilization but is often considered not as their legitimate right. That disregarded facet stemmed from the fact that what felt private in one region contrasts in another [9]. This discrepancy, to the European Court of Human Rights and some prominent scholars upholding the notion, is due to devoid of consensus on one prescribed agreement on the legal definition of privacy [10]. This implication of privacy tied to its relationship with self-respect, and an individual's autonomy and freedom find itself confronted each day with the illicit masqueraders in this fast progressive Hi-Tec information-sharing cyberspace [11].

Amongst several techniques of information security, the two foremost runners in keeping information confined only to the authorized include Cryptography and Steganography [12].
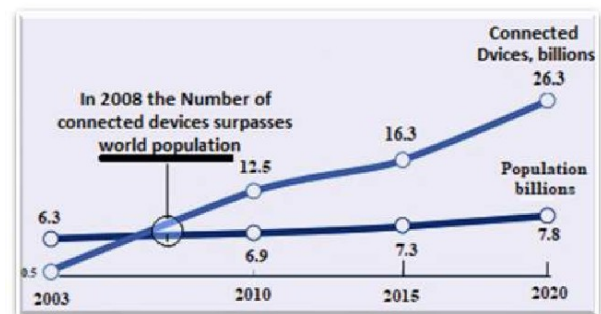

Fig. 1. Increasing Dependency and Reliance of the People on Smart Devices. Image Source [1].

B. Cryptography

The term Cryptography of Greek origin is a blend of two words: (i) crypto: meaning "hidden, secret", and (ii) graphy: refers to' writing' respectively. It is more towards making the information unintelligible as regards its protection from unauthorized disclosure. The first acknowledged indication of its usage via some rarely used hieroglyphic symbols dates back to 1900 BC in an inscription imprinted in the tomb of Khnumhotep II, a nobleman of Egypt. The anticipation was not to hide the contents but to change the mannerism in which it appeared [13].

From the simplest Ceaser cipher [14] to Elliptic Curve Cryptography (ECC) [15], cryptography has undergone sophisticated computational advances because the techniques used in World War I and World War II have now become a matter of seconds to break on a personal computer by the hobbyist [16]. Today, cryptography has found its way in almost every field of life, including e-commerce, credit cards, digital cryptocurrencies, and Government and Corporate Sectors alike by transforming itself into an integral part of any information security infrastructure [13].

The originator of an encrypted entity shares with the intended recipients in prior, the decryption technique, and the encryption keys to impede compromise. As a convention, literature often uses the names Alice for the originator (abbreviated as 'A'), Bob (shortened to 'B') for the intended addressee, and Eve (' E' - the eavesdropper or 'W' - Warden Wendy) for the potential adversary [17]. Fig. 2 explains the encryption and decryption process, along with the terminology used.
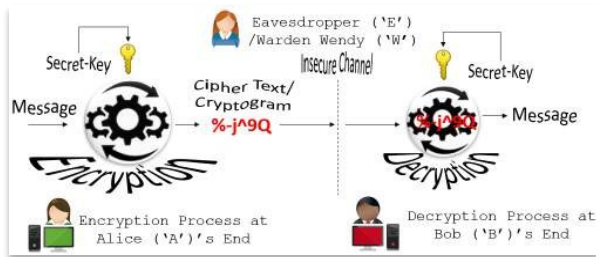

Fig. 2. Cryptographic System with allied Terminology.

1) Advantages: Some of the essential information security services that cryptography provides include:

- Confidentiality - Substituting message contents with arbitrary characters under the control of a shared secret called "secret/cryptographic key", fades out the true context, thereby making it gibberish for an impostor/intruder.

- Authentication and Non-Repudiation - Digital signatures, using the Public-Private key-pair, protects against and counterfeits any spoofing attack.

2) Limitations: Following are some of the constraints that cryptography fell short of addressing:

- It fails to ensure the availability of information as and when required.

- It does not guarantee the veracity of the information.

- It cannot defend against weaknesses, which are a direct consequence of design flaws, whether that of system, processes, or protocols.

- It requires resources both in terms of computational time and power and money (infrastructure support).

a) Attacks on Cryptography: A variety of attacks [17–20] discussed below exists, whose very intent is to extract the meaning out of the encrypted contents without the knowledge of the secretly shared cryptographic key.

- Known Plaintext Attack - Knowing the plaintext for some portions of the ciphertext, the attacker tries to decrypt the remainder of the ciphertext via formulating some relationship between the two akin to determining the shared key.

- Chosen Plaintext Attack - Here, the attacker gets the plaintext of his/her choice encrypted. Using this plain-ciphertext pair make things easier in deriving the key used in encrypting the plaintext.

- Ciphertext Only Attack - Without having the corresponding plain contents, the attacker has access to a set of ciphertexts. The attack is a success if the conforming plaintext gets determined from those ciphertexts. This type of attack facilitates in deriving the encryption key, and hence any cryptosystem must guard against it.

- Brute Force Attack [21] - Knowing the keyspace, the attacker tries every possible combination to find out the key used for encrypting the plaintext. For example, given a key-size of 8-bits, the attacker applies all the possible 256 8-bit patterns (that is, 2^8=256) to decrypt the ciphertext. The attack, however, is resource-intensive, where the attacker might succeed in the very first attempt or continues until the last bit pattern.

- Dictionary Attack [22] - The attacker constructs a dictionary of ciphertexts along with corresponding plaintexts. Later, whenever a ciphertext needs to be analyzed, the same is searched in the dictionary in an attempt to retrieve the plaintext with subsequent dictionary updates.

b) Steganography: The name Steganography is an amalgamation of the two Greek words (i) steganos, which means "covered", and (ii) graphy: which refers to "writing". The name itself comes from Latin - Steganographia. Unlike cryptography, Its practical usage dates back to 440 BC [23]. It intends to hide the information traces as if those do not exist at all. Over some time, this technique has also evolved - from old unconventional methods ranging from shaving of the head and writing a message on the scalp, hiding message by engraving it on a lamb's belly to more recent microdot, invisible ink, null cipher, drifting from spatial to the frequency domain and hiding of data in multimedia files [24]. Fig. 3 gives the terminology and the process of secret message embedding and extraction [25].
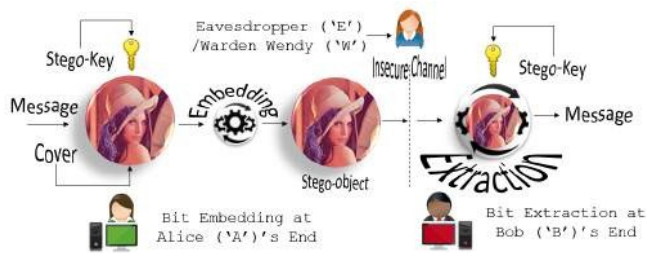
Fig. 3.   The Steganography System with allied Lexicon.

3) Advantages: Following are some of the pros of the technique:

- Hides the existence of information as if that does not exist at all.

- With effective embedding methodology, the similarity between the cover and the Stego image makes detection of the embedded bits trivial because it shows some resembles for Gaussian noise.

- protects the anonymity of its recipients to some extent.

4) Limitations: Some of the cons of the technique include:
- The original cover used as a data carrier, ought to remain secret.

- The embedding algorithm losses its effectiveness if used without a Stego key.

- Security of Steganography System is often found misleadingly attributed to encryption of secret message bits before its embedding inside the cover.

a) The Gauging Parameters: The criteria to judge the effectiveness of a Steganography System include the triad of security, capacity, and imperceptibility [26], as shown in Fig. 4.

- Security - The Steganography System can withstand the traceability of the hidden information.

- Capacity - The maximum amount of information that the Steganography System can safely hide within the cover.

- Imperceptibility akin to undetectability - The ability of the Steganography System to at least minimize if not avoid the cover's degradation as regards secret information inset.
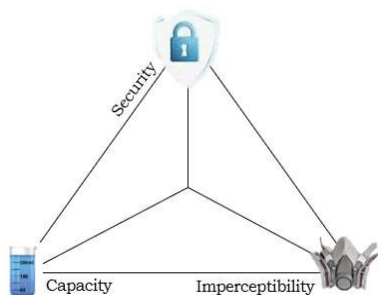

Fig. 4.   Criteria to Gauge the Appropriateness of any Steganographic System.

It is, however, opined that imperceptibility is one of the components that ensure the security of the Steganographic System.

b) Attacks on Steganographic Systems: [27] has explained five types of aggression against any Steganographic system, and hence, it is but imperative for data hiding schemes to guard against those.

- Chosen Message attack - By selecting messages of choice, the attacker tries to examine the effects on the Stego-object to establish some relationship between the two.

- Known-Message attack - With the original message in hand, an attacker contrasts it with the generated Stego-object to patronize the bits for possible signatures for futuristic usage.

- Stego-Only attack - With only the Stego-objects in the possession and with the known algorithm, the attacker studies them to demystify the embedding.

- Known-Cover attack - Having the original cover alongside its corresponding Stego-object, the task of an attacker reduces to finding the differences between these two and extracting the hidden information.

- Chosen Stego attack - With selected Stego-objects, the attacker, works in conjunction with the algorithm to detect the embedded information.

This work is an extended version of the preliminary research carried out under the title "Nondeterministic Secure LSB Steganography for Digital Images", registered for presentation and publication in the international conference on Cyber Warfare and Security (ICCWS 2020), Pakistan.

The structuring of this endeavor is as follows: Section II discusses the basics of a digital image, it's processing, related domains, and the concept of bit-plane slicing, which is followed by a review of some recent LSB based Steganography schemes and its variant research in Section III. Section IV explains the research gap to address, which is derived from the literature review. Modeling considerations for our proposed method are discussed in Section V whereas LSB embedding and subsequent extraction methodology are the topics of Section VI. The details on the State-of-the-art in digital Image analysis and Steganalysis are showcased in Section VII. Section VIII explicates on the test results and their contrast with those of antecedents. Discussion on the contemporary and context-based issues under current pretext appears in Section IX. Section X concludes the proceedings.

II.   DIGITAL IMAGE STEGANOGRAPHY

A. Digital Image

A digital image composed of picture elements alias pixels has a finite, discrete numeric representation. That representation corresponds to image intensity called gray level at a specific place, which is the direct outcome of its two-dimensional spatial coordinates denoted with 'x' on the x-axis, and 'y' on the y-axis, respectively. Based on its modes of derivation or acquisition, such as placement of bits in a 2D

format or scanning, a digital image may have the categorization of either a vector or raster type, respectively. The pixels reside as a raster image or raster map in a computer's memory, visualized as a two-dimensional array of integer values [28-29].

B. Digital Image Processing

Digital Image Processing (shortened as DIP) refers to pixels manipulating and transformation techniques that either enhances image quality such as noise reduction smoothing edges, or extracts information like features, with computerized algorithms.

C. Bit Plane Slicing

Set of bits that correspond to a specific bit position in each of the binary sequence representing the pixel value/intensity. For example, for 8-bit data representation (byte), there are 8-bit planes: the first-bit plane, while traversing from left to right, comprises the most significant bit (MSB), whereas the last that is, 8th position, contains the least significant bit (LSB). Fig. 5 gives the pictorial illustration of the concept.
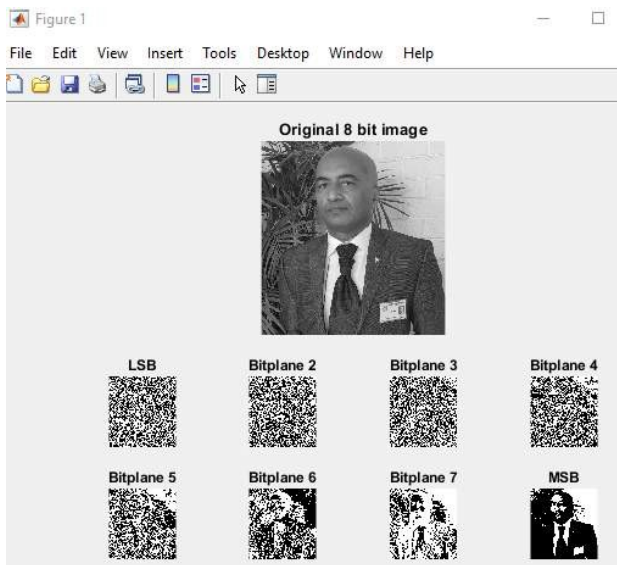

Fig. 5.   Bit-Plane Slicing Illustrated.

As apparent from the figure above, the LSB plane contains information that does not profoundly contribute to image composition, whereas the MSB plane significantly contributes towards image visualization. The said characteristic of bit-plane slicing serves the basis for LSB Steganography.

D. Domains of Image Processing: Most often discussed Image Processing Domains Include

- Spatial Domain - In a 2D image representation (that in a matrix form), each element represents corresponding pixel intensity. This intensity distribution state of 2D matrices is called Spatial Domain. Recommended for working on images of real objects, some of the Steganography techniques related to this domain are the Modulus function, Pixel Value Differencing (PVD) [30], Least Significant Bit (LSB) (the focus of our proposed research), and Combination of LSB and PVD.

- Frequency Domain - A space where an image value at point P represents the amount that the image intensity values vary over an exact distance concerning P; that is, it gives the rate of change of pixel values. Preferred for working on images of modeled contours, some of the Steganography techniques of this domain include transformation methods such as Discrete Cosine Transform (DCT) [31] and Discrete Fourier Transform (DFT) [32].

E. LSB Steganography

Let $p[i] = \{0, 1, . . ., 2^n-1\}$ represent an n (=8) bit integer sequence where $p[i]$ denotes the pixel intensity of the ith pixel in an 8-bit grayscale image. Hence, each $p[i]$ can then have a big-endian bit representation in terms of $b[i; 1]$; $b[i; 2]$; ... ; $b[i; n]$ for n number of bits using Eq. 1:

$$p[i] \leftarrow \sum_{k=1}^{n} b[i,k] * 2^{(n-k)} \qquad (1)$$

LSB Steganography, as the name implies, operates by replacing the LSBs of each $p[i]$ with each message bit denoted by $msg[i]$, thereby rendering the Stego image as $stegO[i]$. The insertion of message bits is sequential and consecutive for each pixel comprising the cover image. The embedding process for the LSB insertion takes the form, per n pixels for distinct k, where $k = 1 . . . n$, as shown in Eq. 2.

$$stegO\ (p[i]) = p[i, b_{lsb}] \leftarrow msg[w_{(j,l)}, b_k] \qquad (2)$$

The $w\ (j, l)$ is the corresponding l^{th} character of the word constituting the message, where $j = \{1,..., $ total words in the message$\}$, $l = \{1, . . ., $ number of characters in the jth word$\}$, and b_k represents the distinct character bits.

III.   REVIEW OF THE LITERATURE

One of the earliest accounts on digital Steganography goes to Kurak and McHugh [33]. The methodology proposed by them while examining image downgrading and contamination bear some resemblance to today's 4-bit LSBs (least significant bits) embedding. The author in [34] used a linear feedback shift register (LFSR) to generate pseudorandom numbers within a specified given range to locate the exact pixels for embedding bits of the secret data using the LSB technique. Arguably the assertion is that random pixel selection shall brace the security of LSB technique. The author in [35] first came up with an improved one-dimensional (1D) chaotic map by eliminating the inherent drawbacks of its narrower range of chaotic behaviors, and the uniform distribution of the key sequence. It followed the proposal of a color image LSB Steganography using their upgraded 1D chaotic map, which conferred on the exactness of the proposed bit embedding method. The author in [36] introduced an Optical Character Recognition (OCR) based Steganographic technique in which the feature form of the message got embedded in the target cover image. Extracted character-level features contained were embedded in the cover image to strengthening the data hiding because an impostor shall first know the hidden features, and even after that has to have a qualified OCR model to recover from the decoded contents (features). The results were found confirmatory to an English Printed Character dataset (Chars74K Dataset) for mixt LSBs. The author in [15] used the Elliptic Curve Cryptography (ECC) algorithm to encrypt the secret message and, after that,

interleaved it into the cover by employing the LSB Inversion procedure. The author in [37] encrypted secret message before embedding it within LSBs of random multiple color pixels of the cover image using the Stern-Brocot Sequence. The author in [38] proposed a new technique using integer wavelet transform (IWT) to conceal patients' information by applying modified least significant bit (m-LSB) method to embed that into the randomly selected transform coefficients of an ECG signal via chaotic maps. The author in [39] doubly- layered reversible information hiding (RIH) method using the least significant bit (LSB) matching technique to improve the bit embedding efficiency (EE) and to enhance its quality by restraining the falsification instigated on to the Stego-image. The author in [40] used the LSB Steganography technique to hide images (as secret information) within the image, followed by scrambling Arnold technique to ascent the cover image for added security. The author in [41] proposed a mixt of cryptography, Steganography, and digital watermarking, calling it "Next Gen". The encrypted patient's information got embedded in the cover image using the LSB Steganography technique, and for authentication digital watermarking was used. The author in [42] presented a novel way of modifying true-color image pixels to facilitate high message bit embedding with least distortion by modifying at least one bit per Red (R), Green (G), and Blue (B) pixel to a maximum of 7 bits/pixel. Any bit position starting from LSB to 7th bit of the byte can appear modifiable. Arbitrary pixels with random bits used for hiding purposes for added security. The author in [43] suggested replacing the familiar words of English literature constituting the message with the values (128–255) of ASCII codes to use 8-bits in place of 8n bits, and converting those into respective binaries. Respective corresponding ASCII values are used for translating words other than those specified into bits. After rotating the cover image by some angle derived from the MSB and LSB of the secret key, the secret bits embedded in a spiral form pixel by pixel that is, either following a clockwise direction or counterclockwise based on the angle of rotation within the RGB. The channel that is, R, G, or B having the highest value gets secret data bits embedded into it under control of a secret key for which its most significant bit (MSB) is XORed with the LSB of the channel having the highest value. The XORed results decide which channel to use for embedding purposes. The offering of randomness by the two-step embedding serves the basis for [44]. Here, in the first stage, the secret encrypted quantum image is embedded into another quantum watermark image. After that, the quantum watermarked modified image is embedded into a quantum cover image using the optimal LSB-based algorithm. in the recovery phase, a series of inverse operations applied to retrieve the secret quantum image can be reconstructed by; only the Stego image, and the key can extract the secret quantum image. The author in [45] used modular arithmetic to hide secret data by avoiding the delinquent of overlapping. In the first instance, the target data to embed undergoes an intermediary conversion via another numbering system. The digits thus converted are embedded through articulating the intensities of the cover image in a manner that facilitates its easy retrieval at the receiving end. In their effort to increase bit embedding capacity of LSB Steganography technique [46] took inverting cover pixels in place of replacing those. The

method involves finding the maximum and minimum values of the data that needs protection. This step is followed by subtracting all the secret contents from the maximum value. The results follow a division, and the new values thus obtained are inserted into the cover image to get the Stego-object/image. The author in [47] shared initial findings regarding modified LSB Steganography. By using r - indiscernibility relations, the authors embedded the secret data in a cover image in a semi-random manner. However, the same is reconstructable deterministically by employing a mask used during embedding. Each Byte of a cover pixel that contains a fixed combination of exact indiscernible bits with the mask serves as a placeholder for LSB replacement. The author in [48] linked the LSB matching method with the image enlargement technique to extend the extents of the Stego image. Doing so ensures an appropriate spread of secret information inside the Stego-image. The author in [49] proposed n-right most bit replacement technique for image Steganography for 1 ≤ n ≤ 4. The method iterates by converting the n-right most bits of every pixel and n-bits of the secret data to respective discrete values. By translating the difference of the two values, followed by the replacement of the cover pixels, produces Stego-image. The author in [50] proposed an LSB based Steganography method to hide secret data in digital images in a pseudorandom fashion via three chaotic noise generators based on the skew tent chaotic map. The author in [51] suggests utilizing a bit reversal method based on 2 schemes for improving the quality of Stego-image. The suggested schemes employee rearrangement of least significant bits of some of the pixels of the cover image if they are in proximity with a specific pattern of a few bits of the secret pixels. The author in [52] presented a high capacity image Steganography technique by blending pixel differencing and swapping mechanisms by dividing the image into 3×3 pixel non-overlapped chunks. LSB substitution is then applied on for every pixel within that chunk, followed by the application of quotient value differencing (QVD) on the leftover six bits. Because the proposed methodology is prone to fall off boundary condition, hence, that particular chunk stands undone from the said hybrid embedding, giving way to 4-bit LSB replacement. The author in [53] performs XOR-*ing* of the LSB of the red (R) in the RGB channel with the secret key bit that is, XOR (R_i, K_j), which determines the subsequent hiding of message bit in either Green (G) or the Blue (B) channels of RGB. If the result of the XOR operation is 1, the secret message bit is inserted as the LSB of G-channel else the B-channel LSB is replaced with a secret message bit.

IV. RESEARCH PROBLEM

As evident from the literature reviewed, the majority of the effort rests in increasing bit embedding capacity of the cover image or to increase the perceptibility of the Stego image akin to increasing the system's security from the perspective of visual attack. However, the exertion in detecting hidden bits under a known-cover and known-message attack scenario is significantly missing. Things get further worse when the above situation gets linked to Kerckhoff's principle [54], which states that the security of a system lies in its key when that system is in a public domain. Hence, with the known algorithm, the cover, and the Stego image, the task of Wendy (the aggressor)

equates only to detect and extract the embedded bits. The example that follows and as illustrated in Fig. 6, clarifies on the assertion.
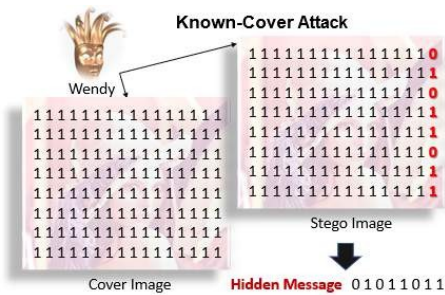

Fig. 6.   Wendy Detects and Extracts the Hidden Message.

- Example: Let matrix A of order ($m \times n$) represent the cover image and let B denote the Stego image of the same order. Then the warden Wendy, in possession of the cover image and the known algorithm, needs only to contrast the two images to retrieve the hidden secret.

Using pseudo-random number generator (PRNG) to scatter bit embedding is illusionary, as [55] called the use of PRNG "a sin." It ought to mention here that most Steganography Systems prohibits the reusability of the cover, which is another pitfall of existing schemes.

Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR) seems overemphasized in recent studies in the context of Steganography as compared to the barely touched mean and standard deviation (STD) of an image. It is pertinent to mention here that PSNR is a measure of signal strength that does not take into consideration the human visual system (HVS) as compared to M/SSIM, which links HVS to illustrate the quality of an image.

In image processing, the "mean" accounts for the contribution of individual pixel's intensity towards the entire image, whereas standard deviation accounts for the dispersion of a pixel from mean that helps categorizes image regions.

V.   Modeling the Proposed Steganography Algorithm

Simmons [56] proposed the first model for Steganography by taking a prisoner's problem whose elucidation is as follows:

- Alice and Bob were taken to prison, where they were locked up in separate cells. They were allowed to communicate with one another, but only through Warden Wendy, who on finding a clue of covert communication between the two, was authorized to send them in long imprisonment from where they could never return. Alice and Bob mutually shared some parameters for communication before they were taken into custody. The two must now agree on a scape plan using pre-agreed parameters.

Inspiration for our proposed model came from the fact that in addition to the limitations highlighted in Section IV, the pseudorandom number generation is also constraint by the fact that all the recipients engaged in communication must also have the same random number generator or that random

sequence of numbers at their respective ends for appropriate bit extraction/message retrieval to succeed.

A facet of LSB insertion is that the inserted LSBs have a direct impact on the output statistics within the purview of the pixels/composition of the selected cover image. In our proposed algorithm, for example, the quality of the random number generator contributes significantly to the stego image quality. It is so because a True Random Number Generator (TRNG) produces different results each time of its use. It is this feature of TRNG that is, pure randomness, which information theory also supports [57] that explicitly points at the strength associated with its usage to achieve a secure information-theoretic solution as in our proposed solution. Additionally, the use of a TRNG, in our case, also eliminates the need for having it at recipient end for extracting hidden information, which is a compulsory requirement of techniques employing pseudo-random number generators. Doing above is advantageous to attain:

- an Information-Theoretic Secure Steganography Solution, and.

- reusability of the cover as opposed to the existing techniques where the original image is to be kept secret on its usage.

To achieve our set goal, we experimented with a MATLAB (*R2020a*)' *rand*' function for lateral replacement with a TRNG.

To facilitate ease and to speed up of the bit extraction and information retrieval process at the receiving end, we further favored (from futuristic requirements considerations as well) for a message header to keep the original message length (first six bytes of the header). It is followed by the file name, along with its associated content type (subsequent twelve bytes), that is, text, image, audio, video, and such other file types, which is followed by the actual message. The notation '*m*H' shall be used in the lateral discussion regarding any such message.

VI.   Proposed Steganography Algorithm

The bit embedding and extraction processes are explicated as follows.

A.   Bit Embedding Process

    ❖ Inputs:
    i.   Secret Message/Contents
    ii. Cover Image
    iii. Stego-Key (at least 4096 bits)
    ❖ Output:
    i.   Stego Image

To send a secret message, the initiator shall take the following steps:

1. Select a Stego-key of length $\geq$ 4096 bits.
2. Select the secret message/contents for embedding:
   a. by appending the message length (6 bytes), followed by content's filename (8 bytes) and type (4 bytes), that is, its extension.
   b. translating the whole text (*m*H) into its equivalent bits.

c. translate Stego key into its equivalent bits.

d. Exclusive-Or (XoR) mH with the Stego key bits. Extend the Stego key by replicating it till it equals mH length if needed.

3. Select the cover image.

4. Iterate through the cover image by taking one pixel at a time till the end of file (EOF).

    a. Replace the pixel's LSB with a random bit, preferably generated via a TRNG.

    b. Check the following:

        i.   is the Stego key bit (moving from right to left) is ON?

        ii. is the pixel's MSB is OFF?

        iii. is the Stego key bit (moving from left to right) is ON?

        iv. are there still some secret message bits to process?

    c. If any of the answers above are FALSE then move to step (e).

    d. Replace the pixel's LSB with the secret message bit.

    e. If all the pixels are processed, then save the Stego image and move to step 5.

    f. Increment the message bit counter and decrement the Reverse counter.

    g. If the Reverse Counter reaches zero, reset it to the length of the Stego key.

5. Exit the bit embedding process.

The process above is illustrated in Fig. 7, and the corresponding source code is written in MATLAB as shown in Fig. 8.

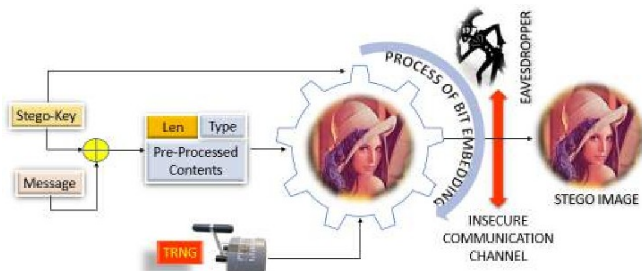Fig. 7.   LSB Substitution Illustrated.

```
m=1;                            % Forward Counter
ml=0;                           % Message bit Counter
l=length(SecM);                 % Length of Secret Message/Contents
rc=4096;                        % Reverse Counter
for ii=1:size(image,1)
    for jj=1:size(image,2)
        pixel=bitset(image(ii,jj),1,rand);        % Random LSB manipulation
        if Ky(rc)=='1' && bitget(pixel,8)==0 ...
                && ml<=1 && Ky(m)=='1'
            ml=ml+1;
            pixel=bitset(pixel,1,SecM(ml));  %Embedding Secret Message bit
        end
        stego_image(ii,jj)=pixel;           %Stego Image
        m=mod(m, 4096)+1;                    %Incrementing Forward Counter
        if rc-l==0
            rc=4096;
        else
            rc=rc-1;                         %Decrementing Reverse Counter
        end
    end
end
```

Fig. 8.   MATLAB Source Code for LSB Substitution in Grayscale Images.

B. Bit Extraction Process

❖ Inputs:

    i.   Stego Image

    ii. Stego-Key (Same, as used in embedding)

❖ Output:

    i.   Hidden/Extracted Message

Following are the steps to extract the hidden message from the Stego image:

1. Select the pre-agreed Stego key and translate it into equivalent bits.

2. Select the Stego Image and iterate through it by taking one pixel at a time, first up to 144 bits to extract the hidden message length, file name along with its extension, and then up to the message length (just pulled) as per following procedure:

    a. Check the following:

        i.   is the Stego key bit (moving from right to left) is OFF?

        ii. is the pixel's MSB is OFF?

        iii. is the Stego key bit (moving from left to right) is ON?

        iv. are 144 bits extracted? (or are there still some secret message bits to process?)

    b. If any of the answers above are FALSE, then move to step (e).

    c. Extract and store the pixel's LSB.

    d. If the extracted bits equal 144 in length or all the bits extracted as per hidden message length then:

        i.   Exclusive-Or (XoR) the extracted bits with the Stego key bits.

        ii. Translate the results into bytes (8-bit chunks).

        iii. Once gone through, save or discard the message as applicable.

        iv. Move to step 3.

    e. Increment the message bit counter and decrement the Reverse counter.

    f. If the Reverse Counter reaches zero, reset it to the length of the Stego key.

3. Exit the bit extraction process.

As evident in the procedure above, supplemented by the MATLAB code of Fig. 9 and the bit extraction process shown in Fig. 10; does not necessitate the same TRNG at the receiving end. Moreover, there is no need to have a TRNG at the receiver's end for unidirectional communication.

Example - The following exemplifies the bit embedding and extraction processes of our proposed Information-Theoretic Secure LSB Steganography solution for grayscale images.

```
m=1;
ml=1;%Message bit counter
hl=48; %Hidden Message Length
rc=4096; %Reverse stego key counter
H=0;
for ii=1:size(stego_image,1)
    for jj=1:size(stego_image,2)
        if  Ky(rc)=='1' && bitget(stego_image(ii,jj),8)==0 && Ky(m)=='1'
            SecM(ml)=bitget(stego_image(ii,jj),1);
            ml=ml+1;
        end
        m=mod(m,4096)+1;
        if rc-1==0
          rc=4096;
        else
          rc=rc-1;
        end
    end
end
```

Fig. 9.   MATLAB Bit Extraction Source Code.
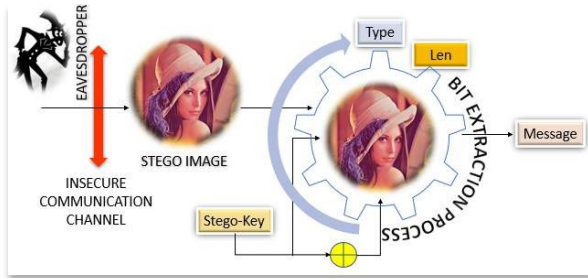

Fig. 10. Illustrating Bit Extraction Process.

Let:

i. S_k = 11001111 be the Stego Key bits

ii. mH = 00000110 denotes the secret message bits whose Exclusive-OR with the Stego key gives the result as mH ← 11001001 ← =XoR (11001111, 00000110)

iii. R_b = 1000010 be the randomly generated bits, and

iv. C_b = 00101000 01111010 represents the cross-section of pixels of the grayscale Cover Image.

1. Embedding Secret Message Bits - Step by step explanation is as follows:
    a. Pre-Processing of pixels
        i.   Take the 1st pixel of the Cover Image that is, 10101000
        ii. Take the 1st random bit, that is, '1' and substitute it as the LSB in the selected Pixel as follows: 00101001 ← 00101000 ← 1
    b. Secret Message Bit embedding
        i.   Because the MSB of the 1st cover pixel is 0, it is placed as the 1st pixel of the Stego image without secret message bit embedding.
        ii. The forward and reverse counters are incremented and decremented, respectively.
        iii. The 2nd cover pixel, that is, 01111010, is selected.

iv.   The pixel remains unaltered after pre-processing because the corresponding random and LSB bit is '0'.

v.   Because the 2nd most bits of the Stego key while traversing from left to right and right to left is ON (/1), and the MSB of the selected pixel is OFF (/0) as well, hence the 1st message bit that is, '1' is inserted as LSB of that pixel.

vi.   The processed pixel is then placed as the 2nd pixel of the Stego image.

c. The Steps (a - b) above are repeated for all the pixels   by continuous replacement of LSB bits with random bits meeting the aforesaid criteria, once all the secret message bits are processed. Thereafter, the Stego image is saved, and the bit embedding process terminates.

d. The Stego image takes the form as 00101001 01111011 and is transmitted to the receiving end.

2. Bit Extraction Process - Step by step explanation is as follows:
    ❖ Given:
        i.   Stego key S_k = 11001111
        ii. Pixels values (Stego Image) S_o = 00101001 01111011
    ❖ We need to find the hidden message = (mH) =?
    a. Extracting the Hidden Bits
        i.   Because the MSB of the 1st cover pixel is 0, it is excluded from the bit extraction process.
        ii. By taking the 2nd pixel of the Stego image, it is observed that MSB of the pixel is OFF (/0). Also, the 2nd most bits of the Stego key while traversing from left to right and right to left is ON (/1) respectively, hence the LSB of that pixel, which is ON (/1), is extracted, and it is the first hidden encrypted message bit.
        iii. Exclusive-Or (XoR) the extracted bit with corresponding Stego key bit gives the hidden message bit. That is mH←0←XoR (1, 1).
        iv. Likewise, the above process continues until the extraction of all the secret message bits.

VII. Steganalysis Techniques and Image Quality Assessment Tools for Digital Image Steganography

The following elucidates on the state-of-the-art in image analysis that served as the foundation towards gauging the output as rendered by our proposed secure Steganography solution:

A. Steganalysis

The art and science of Steganalysis [58] aim at detecting and possibly extracting potentially veiled artifacts known as the

601 | P a g e

payload (referred to as active Steganalysis [59]) from pragmatic data either with or without the prior knowledge of the underneath Steganography algorithm and allied parameters. This technique has gained significant prominence in forensic sciences and attained state-level recognition [60] in technically advanced countries [61] because detection or unveiling of the concealed information may help avoid and overcome catastrophic security situations. Recent interest in digital Steganalysis dates to the publication of the report regarding illegal usage of Steganography by the malevolent engaged in terrorist activities [62], which further got intensified after the 9/11 calamity [63-64].

1) LSB Steganalysis: With specific reference to the LSB Steganography technique, the LSB Steganalysis methods fall into three categories whose concise explanation follows subsequently and serves as a preferred choice in analyzing our proposed methodology:

a) Structural detectors [65] - Explicitly analyze the pairing structure of LSB substitution in pixel groups.

b) Weighted Stego-image (WS) [66-67] – Strives to estimate the embedded bits.

c) Statistical Detectors [68-69] – It is the application of analytical techniques to the embattled image.

The author in [70] presented a Least Significant Bits Steganalysis technique capable of detecting the existence of randomly scattered hidden data embedded in the LSBs of natural continuous-tone images. The method precisely measures the embedded message length, even for lengths that are relative to the target image size. It works by forming some subsets of pixels whose cardinalities vary with LSB embedding, and which can precisely be quantified.

The author in [67] enhances the Weighted Stego-Image (WS) Steganalysis method evolved for LSB replacement payload size estimation in digital images. In doing that, the study suggested for up-gradation of the three components, namely bias correction, the cover pixel prediction, and the least-squares weighting. Experimental results spread over more than two million attacks in total, which were based on images from numerous sources, and pre-processing antiquities showed significant improvement in the accuracy leaving behind the best of the structural detectors by avoiding their high rate complexity.

In contrast to the Least Significant Bit (LSB) Steganography, Steganalysis uses structural or combinatorial traits of the LSB embedding. The author in [65] suggested a general framework for detecting hidden messages along with giving an estimate of their length by including all the combinatorial structures covering those of the earlier research. Experimental evidence suggested a higher success rate of detection for the proposed method in contrast to that of its competitors.

The author in [71] suggested the detection of hidden messages in the Least Significant Bit (LSB) plane of an original image under the assumption that the mean level and the covariance matrix of that image are unknown. The adaptive statistical test was so designed that the anticipated distribution shall remain independent of the parameters of the referral

image, and yet ensuring the highest probable degree of detection of the hidden bits. The test replaces the estimates developed on a local linear regression model with those of the unknown parameters. It was shown that the probability of detection gets maximized with the increased image size, which served as an asymptotic upper bound for the detection of hidden informative bits.

2) Image Quality Assessment (IQA): With the abundance of varied digital multimedia contents like audio, video, and such other file formats for data concealment [72-73], this research focuses on 8-bit grayscale digital images for its usage as a cover in carrying secret information and hence, shall only discuss the said technique from that facet. The justification in selecting digital images for our proposed secure Steganography solution is that being the most preferred media type after textual communication, these easily pass unnoticed through information barriers. This trait is because of their success in exploiting the human visual system (HVS) [74–77], contrary to the text where a single bit change results in an erroneous character. Further, their layout provides several redundant and partisan areas such as edges that serve as regions of interest (ROI) [77–79] for embedding information, which most of the contributions on the subject [80–82] have exploited for increased payload.

However, the insertion of information bits within a digital cover is likely to affect the cover's quality [83], which tends to make it a subjective matter [84]. It is because the perception of quality varies from person to person, and hence, it is unlikely to have an unbiased agreement on that matter. The situation above calls for having an Image Quality Assessment (IQA) methods/metric to quantify the image's quality objectively for reference [85], and which shall remain globally acceptable.

Since the approaches adopted for digital image analysis considers either a change in features between the original image/cover and the modified image called a visual attack or rely on the statistics of the modified image contents, including its type, expected payload length and such other attributes, hence, in the purview of a reference image, the IQA methods fall into three categories as follows:

a) Full Reference [86]: The method assumes to have an undistorted original reference image for comparison with an altered image, and hence destined to maintain accuracy in terms of the results. In the context of Steganography, this method is analogous to the most lethal known cover attack, which is the prima face of our proposed research. Following serves as some of the performance measures:

1) Absolute Difference (AD) – It is an effective similarity measure that gives the absolute difference between the referenced and the filtered (altered) image by subtracting the corresponding elements of the two matrices, as shown in Eq. 3.

$$AD = \sum_{i=1}^{N} \sum_{j=1}^{M} |R(i,j) - T(i,j)| \qquad (3)$$

2) Maximum Difference (MD) – It quantifies an image's contrast level by using Eqn. 4 and ranges from 0 to any of the

positive values. However, the higher the value, the more inferior the image's quality.

$$MD = max \, |R \, (i,j) - T \, (i,j)| \tag{4}$$

3) Mean Absolute Error (MAE) – Suitable to measure the blurring effect of an image with the ideal value being zero. Calculated using Eqn. 5, a higher value indicates a degraded quality image.

$$MAE = \frac{1}{(M \times N)} \sum_{i=1}^{N} \sum_{j=1}^{M} |R \, (i,j) - T \, (i,j)| \tag{5}$$

, where M × N is the size of the images, R (i, j), and T (i, j) are the referenced and tarnished image sequentially at the ith and j^{th} location.

4) Mean Square Error (MSE) – As the name suggests, it computes the mean square difference between the referenced and distorted images by using the formula shown in Eq. 6. MSE is the image quality measure typically when used for noise detection or blur removal, and henceforth.

$$MSE = \frac{1}{(M \times N)} \sum_{i=1}^{N} \sum_{j=1}^{M} |[R \, (i,j) - T \, (i,j)]^2| \tag{6}$$

, where M × N is the size of the images that is, $i = 1, \, .., \, M$; $j = 1, \ldots, N$. The anticipated value for MSE ≥ 0, where zero is the ideal result.

5) Root Mean Square Error (RMSE) – It is the square root of mean squared error (MSE) computed via equation Eq. 7.

$$RMSE = \sqrt{\left(\frac{1}{(M \times N)} \sum_{i=1}^{N} \sum_{j=1}^{M} |[R \, (i,j) - T \, (i,j)]^2|\right)} \tag{7}$$

The range of RMSE is ≥ 0, where zero stands as the ideal value.

6) Peak Signal-to-Noise Ratio (PSNR) – It is a measure of the signal's strength and calculated using equation Eq. 8. For Steganography, it is being used as a quality measure though it is independent of the human visual system (HVS).

$$PSNR = 10 \, log_{10} \, \frac{P^2}{MSE} \tag{8}$$

The ideal value for PSNR is ∞. However, values > 0 are acceptable, where P is the highest gray level in the image, and MSE is computed using Eq. 6.

7) Laplacian Mean Squared Error (LMSE) – It is a measure of image degradation on account of factors such as edges, noise, and such other effects and calculated as shown in Eq. 9. The larger the value of LMSE, the poor shall be the quality of the target image.

$$LMSE = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} |[O(R \, (i,j)) - O(T \, (i,j))]^2|}{\sum_{i=1}^{N} \sum_{j=1}^{M} |O(R \, (i,j))|^2} \tag{9}$$

, where O (R (i, j)) = R (i + 1; j) + R (i − 1, j) + R (i, j + 1) + R (i, j − 1) − 4×R (i, j).

8) Normalized Cross-Correlation (NCC) ‒ It enhances the image brightness by the normalization process and is widely used for image restoration purposes. Computed by

using Eqn. 10, the range of NCC varies between ±1, where a −1 indicates a perfect correlation, and +1 a negative correlation.

$$NCC = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} R \, (i,j) \times T \, (i,j)}{\sum_{i=1}^{N} \sum_{j=1}^{M} R \, (i,j)^2} \tag{10}$$

9) Cosine Similarity (CS) ‒ It is computed by taking into consideration the cosine of the angle between two vectors in a multidimensional space. It is a measure of the similarity between the two vectors because the value of the measure increases with the decrease in the angle between them, and is calculated using equation Eq. 11.

$$CS = \frac{\sum_{i=1}^{N} R_i \times T_i}{\sum_{i=1}^{N} R_i^2 \times \sum_{i=1}^{N} T_i^2} \tag{11}$$

, where R_i and T_i are the components of vectors R and T respectively. The similarity equals 1 when both the R and the T are the same, or is -1 if the two are opposite. A value of zero for CM means no correlation.

10) Structural Content (SC) ‒ The structural content is concerned with the spatial arrangements of the pixels in an image. It is a similarity measure between the two images human eye cannot differentiate and is computed by using Eq. 12.

$$SC = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} T(i,j)^2}{\sum_{i=1}^{M} \sum_{j=1}^{N} R(i,j)^2} \tag{12}$$

The best value of SC is 1.

11) Structure Similarity Index (SSIM) – It gives a comparison between the luminance as given in Eq. 13, contrast, as in Eq. 14, and that for the structure, as shown in Eq. 15 of the referenced and target images where the SSIM is computed, as shown in Eq. 16 and Eq. 17.

$$L(uminance)_{AB} = \frac{2\mu_A\mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1} \tag{13}$$

$$C(ontrast)_{AB} = \frac{2\sigma_A\sigma_B + C_2}{\sigma_A^2 + \sigma_B^2 + C_2} \tag{14}$$

$$S(tructure)_{AB} = \frac{\sigma_{AB} + C_3}{\sigma_A^2 \sigma_B^2 + C_3} \tag{15}$$

$$SSIM_{AB} = L_{AB}{}^\alpha, C_{AB}{}^\beta, S_{AB}{}^\gamma \tag{16}$$

, with the weights α, β, and γ = 1, Eq. 16 takes the form, as shown in Eq. 17.

$$SSIM_{AB} = \frac{(2\mu_A\mu_B + C_1)(2\sigma_A\sigma_B + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)} \tag{17}$$

The values of α, β, and γ define the weight assigned to each model, μ_A, μ_B are the averages of signal A and B as in Eq. 18, $C_1 = (K_1P)^2$, P is the highest gray level value in Eq. 19, C2= $(K_2P)2$, $K_2 \leq 1$. $C_3 = C_2/2$, and σ_{AB} denotes the standard deviation between signals (A, B) as in Eq. 20.

$$\mu_A = \frac{1}{M} \sum_{i=1}^{M} A_i \tag{18}$$

$$\sigma_A = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (A_i - \mu_A)^2} \tag{19}$$

$$\sigma_{AB} = \frac{1}{M-1} \sum_{i=1}^{M} (A_i - \mu_A)(B_i - \mu_B) \qquad (20)$$

b) Reduced-Reference (RR) − Proposed by [87], and also known as the partial reference (PR) method, predicts the quality of the target image where only certain aspects of the original image are known. Here, the sender, apart from sending the image over a noisy/insecure/narrow bandwidth channel, transmit along an auxiliary channel some of the extracted features that contribute towards the quality assessment of the reference image, doing this aimed at facilitating the correct /simple retrieval of the original image by the recipient. Depending on the methodology, the receiver may adjust a specific aspect of the altered/received image to reconstruct the same and expound on the quality in the context of the referenced image. Fig. 11 is an illustration of the said concept. However, the availability of an auxiliary channel is a prerequisite for the above method to work.
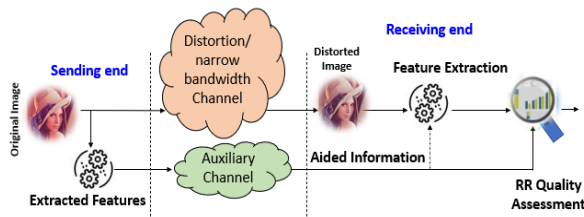

Fig. 11. Illustrating Reduced Reference Image Quality Assessment System.

c) No-Reference (NR) [88] − As the name suggests, the NR method akin to the blind image quality assessment (BIQA) method, does not require a reference image. Instead, the quality assessment is based on the features of the target image. This method parallels a real-time scenario where no reference image is available for cross-comparison and hence, must also be given due consideration.

1) Blind/Reference less Image Spatial Quality Evaluator (BRISQE) − [89] presented a blind/no-reference (NR) image quality assessment (IQA) model that works for the spatial domain. The model, blind image spatial quality evaluator (BRISQE), uses scene statistics of locally normalized luminance coefficients to enumerate probable losses of "genuineness" in the image due to the presence of misrepresentations, that leads to a holistic quality measure. The model does not need transformation to another domain, such as Discrete Cosine Transformation (DCT), Wavelet, and such other transformations. Statistical results show BRISQE superiority over full-reference peak signal-to-noise ratio (PSNR) and the structural similarity index (SIM).

2) Naturalness Image Quality Evaluator (NIQE) – It is an opinion-unaware, IQA method that uses Naturalness Image Quality Evaluator to calculate the no-reference image quality score for the input image. NIQE can ascertain the quality of a distorted image. Decreasing NIQE increases the perceptual quality of the image [90].

3) Perception-based Image Quality Evaluator (PIQE) − [91] proposed a novel no-reference IQ Evaluator real-world imagery. Unlike opinion-based supervised learning, an opinion-unaware methodology quantifies the distortion within an image without any training data but relies on pulling out local features for forecasting quality. Additionally, to mimic human behavior, we estimate quality only from perceptually significant spatial regions. Low computational complexity is another facet of the algorithm.

4) Singular Value Decomposition (SVD) − [92] presented a new grayscale image quality assessment measure to predict the distortion contributed by a variety of noise sources. Five test images, namely airplane, boat, Goldhill, Lena, and peppers with six types of alteration including JPEG, JPEG 2000, Gaussian blur, Gaussian noise, sharpening, and DC-shifting, each with a distortion level of five, were subjected to quality assessment via this method. The measure performed well when compared with PSNR and such other similar measures.

The SVD requires one input matrix and yields three matrices as output. Using a Matrix, A of size m × n the computation is performed as shown in Eq. (21),

$$S = SVD(X) \qquad (21a)$$

$$[U, S, V] = SVD(X) \qquad (21b)$$

$$[U, S, V] = SVD(X; 0) \qquad (21c)$$

where U and V are the orthogonal matrices while S is a diagonal matrix. SVD computes the norm of the diagonal matrix S, which gives the correlation between the pixels in a specific matrix. The best matching occurs when the difference between the two norms equals zero. The authors in [93], [94] also explicate on blind image analysis. To implement the above in MATLAB programming language, [95] serves a good reference.

VIII. EVIDENCE-BASED TEST RESULTS

A total of 30 grayscale images of dimension 512 × 512, as shown in Fig. 12, were selected from the dataset [96] and other freely available web resources for Steganalysis and tested against full, reduced, and no-reference models. All the cover images were in Tagged Image File Format (TIF/TIFF), where maximum sustainable randomly generated bits were used as a message.

A. Results for Steganalysis

The Steganalysis outcome and other statistics shown in Table I elucidates on non-detection of LSB insertion by [65–70] for the said images when processed through our proposed secure steganographic bit embedding algorithm, the theoretical aspects of which are covered in Sec. IX.

B. Full Reference Test Results

Full Reference validity conducted using the MSSIM, SC, MSE, LMSE, NAE, MAE, WPSNR, NCC, CS, and AD. Table II expounds on the output, which speaks high of our proposed secure bit embedding algorithm.

It is also pertinent to mention here that the cosine similarity (CS) for all the test images was found out to be one and hence, is not listed. The NCC of some of the test images is shown in Fig. 13 for visual illustration.

Aerial	Airplane	APC	Barbra	Boat
Breakfast	Cameraman	Cars and APC	clown	Couple
Darkhair	Fish	Frog	Goldhill2	House
Jetplane	Lake	Lena	Library	Lighthouse
Man	Mandril_gray	mountain	Palehair	Pentagon
Pepper	Rice	Walkbridge	washsat	zelda

Fig. 12. 512×512 Grayscale Test Images [96].

TABLE I. SECURITY

Test Images	Maximum		Entropy				Steganalysis				
	M. Bits	bpp	R. Ent	S. Ent	MI	J. Ent	RS	SP	Triples	WS	AUMP
APC	17127	0.0654	5.0534	4.9338	4.7897	5.1976	0.1972	0.1046	0.0145	0.0783	3.152
Aerial	6108	0.0234	6.994	6.048	5.9971	7.0449	0.1187	0.0161	0.0013	-0.027	0.3923
Airplane	2412	0.0093	4.0045	3.9954	3.9751	4.0248	0.0048	-0.0109	0.0013	0.0012	0.2105
Barbra	37856	0.1445	7.4664	6.7877	6.4676	7.7865	0.2975	0.2064	0.3684	0.1269	4.1021
Boat	18672	0.0713	7.1238	6.2889	6.1322	7.2805	0.1258	0.0207	0.0175	0.0222	1.9404
Breakfast	14920	0.0570	7.5423	6.668	6.5434	7.6668	0.0888	-0.0696	0.0152	0.0161	5.9851
Cameraman	25622	0.0978	7.048	6.2691	6.0536	7.2634	0.1126	0.0892	0.0332	0.0661	9.0315
Cars and APC	44140	0.1685	6.5632	6.3966	6.0239	6.9359	0.3961	0.2096	0.3684	0.1861	3.576
Couple	32220	0.1230	7.2952	6.6634	6.3925	7.5661	0.2544	0.1388	0.3684	0.1029	4.1144
Darkhair	40408	0.1542	7.2767	6.6226	6.2823	7.617	0.2079	0.1143	0.0739	0.1429	14.0275
Fish	39801	0.1519	7.3988	6.7599	6.4236	7.7351	0.1808	0.1955	0.0658	0.1266	19.0433
Frog	35655	0.1361	7.0366	6.3404	6.0397	7.3372	0.3442	0.3047	0.3684	0.1863	3.5217
Goldhill2	43237	0.1650	7.4778	6.8427	6.4787	7.8418	0.2892	0.2054	0.3684	0.1217	4.7349
House	35878	0.1369	5.7529	5.2777	4.9738	6.0567	0.1067	0.051	0.0221	0.0887	15.9681
Jetplane	11545	0.0441	6.7135	5.8139	5.717	6.8104	0.0726	0.0437	0.0061	0.0121	1.7099
Lake	32667	0.1247	7.4826	6.7572	6.4846	7.7553	0.1576	0.0352	0.0226	0.0558	1.5588
Lena	31227	0.1192	7.4456	6.7083	6.4469	7.707	0.1372	0.2942	0.0325	0.0761	3.8846
Library	28315	0.1081	6.8562	6.2446	6.0066	7.0941	0.1016	0.06	0.0053	0.0341	2.12
Lighthouse	40906	0.1561	7.4486	6.8356	6.4905	7.7936	0.283	0.281	0.0627	0.1212	5.5697
Man	35528	0.1356	7.2367	6.5587	6.259	7.5365	0.254	0.1616	0.0611	0.1328	7.9212
Mandril gray	30728	0.1173	7.2925	6.5508	6.2933	7.55	0.3758	0.1408	0.0413	0.0893	4.869
Palehair	22802	0.0871	6.9542	6.3037	6.1122	7.1457	0.1467	0.2247	0.0279	0.0708	2.7151
Pentagon	19534	0.0746	6.6548	5.8213	5.6566	6.8195	0.2653	0.1268	0.0272	0.1005	3.2404
Pepper	33117	0.1264	6.7624	6.5306	6.2504	7.0427	0.1522	0.2028	0.0631	0.1234	7.6063
Rice	45744	0.1746	7.0171	6.4051	6.0196	7.4026	0.2398	0.1807	0.0633	0.1307	6.2447
Walkbridge	39859	0.1521	7.683	7.0288	6.6927	8.0191	0.3136	0.1452	0.0489	0.1138	3.5319
clown	49431	0.1886	5.3684	5.7861	5.3684	5.7861	0.2089	0.1414	0.0993	0.1702	14.5023
mountain	29538	0.1128	7.7828	7.0395	6.7919	8.0304	0.269	0.1813	0.0253	0.047	2.7038
washsat	62489	0.2384	2.8676	3.3938	2.8676	3.3938	0.2502	0.2348	0.1159	0.2001	12.4087
zelda	48486	0.1850	7.2668	6.6769	6.2676	7.6761	0.2666	0.2094	0.3684	0.1479	6.5424

bpp = Bits Per Pixel; R. Enrt = Entropy of Reference Image; S. Entr = Entropy of Stego Image;

J. Entr = Joint Entropy; MI = Mutual Information

TABLE II. FULL REFERECE (FR) IQA

File Name	Maximum	Test Results								
	M. Bits	MSE	RMSE	LMSE	NAE	MAE	PSNR	WPSNR	NCC	AD
APC	17127	0.2994	0.5472	0.008	0.0024	0.2994	53.3686	60.6857	0.0684	-0.2696
Aerial	6108	0.5024	0.7088	0.0028	0.0028	0.5024	51.1205	63.7239	-0.0105	-0.4907
Airplane	2412	0.4334	0.6583	0.0146	0.0025	0.4334	51.7619	62.6107	0.0700	-0.4285
Barbra	37856	0.5001	0.7072	0.0024	0.0044	0.5001	51.1405	57.3499	0.0052	-0.4271
Boat	18672	0.4907	0.7005	0.0088	0.0036	0.4907	51.2224	59.7254	0.1088	-0.4544
Breakfast	14920	0.4982	0.7058	0.0198	0.0029	0.4982	51.157	59.9982	0.0277	-0.4696
Cameraman	25622	0.5002	0.7072	0.0293	0.0042	0.5002	51.1392	60.2802	0.0955	-0.4515

Cars and APC	44140	0.4765	0.6903	0.0056	0.0045	0.4765	51.3502	60.4525	0.0097	-0.3946
Couple	32220	0.4963	0.7045	0.0067	0.0041	0.4963	51.1738	60.5239	0.0436	-0.4344
Darkhair	40408	0.5013	0.7080	0.0713	0.0046	0.5013	51.1299	59.3311	-0.0283	-0.425
Fish	39801	0.488	0.6986	0.0612	0.0053	0.488	51.2468	63.9927	-0.0516	-0.4072
Frog	35655	0.4998	0.7070	0.0063	0.0041	0.4998	51.1425	67.009	0.0057	-0.4308
Goldhill2	43237	0.4989	0.7063	0.0096	0.0044	0.4989	51.1505	59.5063	0.0729	-0.416
House	35878	0.3683	0.6069	0.1732	0.0027	0.3683	52.4692	60.6775	0.1353	-0.2998
Jetplane	11545	0.4991	0.7065	0.0157	0.0028	0.4991	51.1493	55.0727	0.0361	-0.4773
Lake	32667	0.4997	0.7069	0.0066	0.004	0.4997	51.144	64.0182	-0.0360	-0.4373
Lena	31227	0.4998	0.7070	0.0154	0.004	0.4998	51.1428	59.1577	0.0209	-0.4394
Library	28315	0.5983	0.7735	0.0016	0.0041	0.5983	50.362	58.3347	-0.0422	-0.5459
Lighthouse	40906	0.5031	0.7093	0.0033	0.0044	0.5031	51.1143	60.5441	0.0661	-0.425
Man	35528	0.5012	0.7080	0.0078	0.0045	0.5012	51.1306	58.645	0.0046	-0.4329
Mandril gray	30728	0.5017	0.7083	0.0077	0.0039	0.5017	51.1267	64.7751	0.0009	-0.4416
Palehair	22802	0.5039	0.7099	0.0049	0.0037	0.5039	51.1071	61.7554	0.0450	-0.4614
Pentagon	19534	0.5046	0.7104	0.0042	0.0036	0.5046	51.1011	63.5935	0.0068	-0.4668
Pepper	33117	0.4913	0.7009	0.0068	0.0042	0.4913	51.2173	58.1202	-0.0195	-0.4238
Rice	45744	0.4996	0.7068	0.0513	0.0045	0.4996	51.1447	62.2896	0.0587	-0.412
Walkbridge	39859	0.4982	0.7058	0.0032	0.0044	0.4982	51.1569	63.1365	0.1072	-0.4204
clown	49431	0.4131	0.6427	0.01	0.0061	0.4131	51.97	54.6714	0.0584	-0.2924
mountain	29538	0.5029	0.7092	0.0013	0.0036	0.5029	51.1163	60.3184	0.0285	-0.447
washsat	62489	0.4745	0.6888	0.0233	0.0072	0.4745	51.3688	55.1983	0.0045	-0.3463
zelda	48486	0.5004	0.7074	0.0342	0.0055	0.5004	51.1381	58.2598	-0.0107	-0.4076

Mean of Normalized Cross-Correlation (NCC) is tabulated above

Fig. 13. Normalized Cross-Correlation Illustrated for some Popular Images.

In image processing (IP), the mean of an Image often termed as spatial filtering is used for noise reduction. Standard Deviation (STD), on the other hand, accounts for the variation or dispersion from the average mean, or anticipated value. A low STD means that the image pixels tend to be closer to the mean, whereas a higher value indicates that the pixels are spread over a broad range of values. For Stego images, the mean and the STD tell on the noise and change in the luminance/intensity of the pixel values. Table III gives a comparison of the mean and STD of the cover images used in testing our proposed methodology.

C. Reduced-Reference Test Results

Reduced-reference Image Quality Assessment was carried out in purview of [57] and conjunction with MATLAB GitHub repository. The results rendered for the test images are as shown in Table IV. It is pertinent to mention that lower the quality score (QS), the higher is the Stego-image quality.

D. No Reference/Blind IQA Test Results

It tends to build a computational model to predict the subjective quality quantitatively from the partisan image without having the aid of original or reference copy. Functions, namely BRIQE, NIQE, and PIQE that accept the distorted/Stego-image as input are used for the computational purpose. Table IV summarizes the output quality scores for the said IQA's. Note that for PIQE, high perceptual quality is associated with a low score value, whereas a high score means low perceptual quality. For NIQE, the case is just the opposite of PIQE.

E. Comparative Analysis

Undoubtedly, the sequential LSB replacement is a security threat, as pointed out by [66]. LSB Steganography is also undermined to have structural flaws [97] and in offering a weak association between consecutive bit planes [98]. However, LSB Steganography, until today, is a strong contender amongst its counterparts. The said fact is evident from Table V that demonstrates high structural similarity and less interference for our proposed secure bit embedding algorithm. The data for the said table is from the recently published [99], which has referred to the studies of [100–102].

Other recent work, including [103], [104] for blending pixel-value difference (PVD) and LSB techniques, and local binary pattern-based (LBPB) [105], published findings that are contrasted in Table VI, further expounds on higher PSNR values as rendered by our proposed bit embedding methodology.

TABLE III. REDUCED REFERENCE (RR) & NO/BLIND REFERENCE (NR) IQA

Test Images	Maximum		Structural Similarity			R. Ref.	No Reference - IQA		
	M. Bits	bpp	SSIM	MSSIM	SC	RRIQA	BRISQE	NIQE	PIQE
APC	17146	0.0654	0.9986	0.9995	0.9961	0.8093	27.233	5.3995	13.6299
Aerial Airplane	6127	0.0234	0.999	0.9998	0.9948	0.9883	12.2286	2.7113	22.4038
Barbra Boat	2431	0.0093	0.9972	0.9992	0.9952	2.3488	16.6887	5.7391	21.7968
Breakfast	37875	0.1445	0.9981	0.9996	0.9932	0.5127	31.8209	4.6049	33.0083
Cameraman	18691	0.0713	0.9979	0.9996	0.994	1.5277	6.9412	5.0183	18.2527
Cars and APC	14939	0.0570	0.9987	0.9998	0.995	0.9059	41.539	5.4934	51.7844
Couple Darkhair	25641	0.0978	0.997	0.9993	0.9937	1.6629	33.0189	5.1635	41.3617
Fish Frog Goldhill2	44159	0.1685	0.9989	0.9997	0.9933	0.3350	9.9519	4.1246	20.3461
House Jetplane	32239	0.1230	0.9985	0.9997	0.9934	1.1657	28.2067	3.362	25.824
Lake Lena Library	40427	0.1542	0.9966	0.9993	0.9936	1.4869	20.5537	4.0567	20.368
Lighthouse	39820	0.1519	0.9967	0.9989	0.9934	0.7906	41.5285	4.0947	65.3779
Man Mandril gray	35674	0.1361	0.9987	0.9997	0.9933	0.6727	19.1534	4.0332	24.3395
Palehair Pentagon	43256	0.1650	0.9983	0.9997	0.9935	1.1303	13.7579	3.5827	18.5686
Pepper Rice	35897	0.1369	0.9978	0.9995	0.9968	1.1906	52.6812	4.4929	46.2368
Walkbridge	11564	0.0441	0.9977	0.9995	0.9949	1.5764	26.5814	3.1438	21.7911
clown mountain	32686	0.1247	0.9983	0.9996	0.9941	0.8629	21.7387	4.36	19.3017
washsat	31246	0.1192	0.9976	0.9995	0.9936	1.4719	4.8896	4.2099	17.0984
zelda	28334	0.1081	0.9992	0.9998	0.9935	0.9255	42.1952	4.2689	48.093
	40925	0.1561	0.9985	0.9996	0.9937	1.2733	18.636	2.9754	38.8563
	35547	0.1356	0.9981	0.9996	0.9931	0.5882	23.1349	2.8212	21.3023
	30747	0.1173	0.999	0.9998	0.9935	0.6472	51.9592	7.9671	22.9831
	22821	0.0871	0.998	0.9996	0.9936	1.3137	18.5818	4.8185	22.3224
	19553	0.0746	0.9988	0.9998	0.9934	0.8922	15.1769	4.6506	16.8965
	33136	0.1264	0.9984	0.9995	0.9934	1.3861	35.5626	7.288	27.3557
	45763	0.1746	0.9977	0.9996	0.9933	1.0565	35.7323	7.1433	24.1532
	39878	0.1521	0.9991	0.9998	0.9937	0.7311	9.5149	2.5131	28.9098
	49450	0.1886	0.9979	0.9992	0.994	1.2194	21.7439	4.8926	43.1609
	29557	0.1128	0.9986	0.9995	0.9948	1.0588	15.9205	3.0943	38.3954
	62508	0.2384	0.9971	0.9992	0.9896	0.9676	9.8024	4.185	18.5335
	48505	0.1850	0.9972	0.9995	0.9923	1.2972	19.0354	5.6148	11.2756

TABLE IV. CONTRASTING FR-IRQ OF THE PROPOSED ALGO

Test Image	Reference	MSE	MAE	PSNR	NCC	SSIM
Airplane	[36]	0.47	N/A	51.36	N/A	0.9957
	[106]	1.3211	0.6107	41.5506	0.99999	0.9831
	Proposed	0.4334	0.4334	51.7619	0.07	0.9972
Baboon	[36]	0.51	N/A	51.01	N/A	0.9995
	[106]	0.4284	0.2244	48.8172	1	N/A
	Proposed	0.5017	0.5017	51.1267	0.0009	0.999
Barbara	[36]	0.52	N/A	50.94	N/A	0.999
	[106]	0.3442	0.1656	49.6504	0.99999	N/A

	Proposed	0.5001	0.5001	51.1405	0.0052	0.9981
Boat	[36]	0.46	N/A	51.41	N/A	0.9989
	Proposed	0.4907	0.4907	51.2224	0.1088	0.9979
Car man	[106]	0.2211	0.2221	51.6788	0.99999	N/A
	Proposed	0.5002	0.5002	51.1392	0.0955	0.997
Lena	[36]	0.52	N/A	50.95	N/A	0.999
	[106]	0.1014	0.1146	55.0332	0.99999	0.9684
	Proposed	0.4998	0.4998	51.1428	0.0209	0.9976
Peppers	[36]	0.9765	N/A	47.5	N/A	0.9579
	[106]	0.4585	0.2907	46.6995	0.99998	0.9706
	Proposed	0.4913	0.4913	51.2173	-0.0195	0.9984

TABLE V. A COMPARISON WITH SOME OTHER POPULAR STEGANOGRAPHY TECHNIQUES

Test Parameters Images	PVD [100]	TBPC [101]	ATBPC [102]	ATCEQES [99]	Proposed Method
PSNR	52.51	53.34	55.34	56.49	51.1428
Lena WPSNR	67.41	67.45	67.45	72.47	59.1577
SSIM	0.9982	0.9987	0.9987	0.9999	0.9976
PSNR	52.23	55.3	55.3	56.49	51.1267
Baboon WPSNR	86.58	79.55	79.55	84.82	64.7751
SSIM	0.9993	0.9995	0.9995	0.999	0.9990
PSNR	53.03	55.3	55.31	56.47	51.1392
Camer WPSNR	66.9	65.32	66.01	63.92	60.2802
aman SSIM	0.9978	0.9983	0.9984	0.9999	0.9997
PSNR	52.49	55.39	55.38	56.51	51.2173
Peppers WPSNR	67.63	68.49	68.84	70.65	58.1202
SSIM	0.9984	0.9988	0.9988	0.9999	0.9984

TABLE VI. A COMPARISON OF PSNR RENDERED BY PVD+ LSB AND FREQUENCY DOMAIN STEGANOGRAPHY ALGOS

Stego Image	LBPB [105]	PVD [104]	3LSB [104]	PVD+LSB [103]	Frequency Domain[103]	Proposed Method
Lena	56.82	39.56	37.92	36.32	45.05	51.1428
Baboon	53.57	37.38	37.93	35.4	41.82	51.1267
Peppers	58.64	39.11	37.93	35.91	45.65	51.2173
Jet	56.23	39.12	37.94	36.41	44.77	51.1493
Boat	57.95	N/A	N/A	35.72	45.7	51.2224
Lake	N/A	N/A	N/A	35.89	0	51.1445
Elaine	N/A	N/A	N/A	34	0	N/A
Couple	N/A	N/A	N/A	35.78	45.95	51.1738

IX. DISCUSSION

The comparison made in the preceding section is in the purview of the fact that the detection and extraction of hidden information are independent of the type of Steganography used and the misapprehension associated with its allied bit embedding capacity such as [106]. It is because the more the message bits embedded inside the cover, the greater shall be the threat of exposure of the underneath bit embedding algorithm. Moreover, changing cover bits in proportion to the message bits may also reveal the hidden content length. More importantly, the reusability of cover images may lead to a compromised situation. As regards images, what matters most at a glance is the perceptibility of the Stego image, which, however, is dependent on the perception of the onlooker/attacker. Practically, subjective image analysis is almost impossible because millions of images are uploaded each day on the web [107]. Hence, a combination of all the theoretical/subjective and objective image analysis techniques contributes to the acceptance or otherwise of any image-based Steganography technique. It is, therefore, imperative to quantify the results of each technique through some Universal rating into a quality score to assess and grade any Steganographic method. It shall help uphold the impartiality by removing the bias towards self-proclaimed efficiency and the effectiveness of one's proposed Steganography method.

For instance, take the example of an original image having an MSE = 0 and an SSIM = 1, as shown in Fig. 14 adapted from [108]. It is apparent from Fig. 15 that despite having the same MSE = 144, the structural distortion in the images is

quite visible/easily detectable. Moreover, different images can have the same PSNR, SSIM, and even entropy. For a thorough understanding, [109] is a good starting point.

REFERENCE IMAGE   MSE=0, SSIM=1

Fig. 14. Original. MSE = 0; SSIM = 1 [108].

MSE = 144, SSIM = 0.988   MSE = 144, SSIM = 0.840

MSE = 144, SSIM = 0.913   MSE = 144, SSIM = 0.694

Fig. 15. Deceptive Statistics for MSE = 144. Adapted from [108].

While [110] stating Watermarking techniques as belonging to a particular group of the Steganography field indicated a trade-off between the three gauging parameters of imperceptibility, capacity, and the robustness as regards watermark system. That is, increased payload shall likely distort the quality of the image. In contrast, the requirement of being secured against Steganalysis also depends on the depth of the embedded bits as LSB, 2-LSB, and such other choices. Likewise, to increase the quality of the Stego image, one needs to either lessen the amount of data to be hidden or the depth of the embedded bits, which seems the main reason behind recent studies for publishing results to 30 or 50% LSB bit embedding. However, as far as the strength of any proposed LSB Steganography bit embedding algorithm goes, it needs to be tested against its full capacity that is 100% bit embedding to get an unbiased view on its security, capacity, and imperceptibility aspects. Our test results are within the purview of [111] that explicitly stated as "excellent", a PSNR with 53 for 100% embedding (1bit per pixel), and calls for the test image 'Lena,' a PSNR value of ≥ 50.

Another distinct limitation noticed while reviewing the literature of state-of-of-the-art image Steganography

techniques is that the least emphasis is given to the Steganalysis that gauges the security/limitations of the bit embedding methodology without which the discussion of data obliviousness through Steganography remains incomplete. Table VII contrasting the Steganalysis results with the recent research [36], which is for images of size 128×128 as compared to our proposed that are carried out on 512×512 images, each with a depth of 8-bits.

TABLE VII. A Comparison of RS and SP Steganalysis Between [36] and the Proposed Method between [36] and the Proposed Method.

	[36]		Proposed		
	128 x 128		512	x	512
	RS	SP	RS	SP	
Airplane	0.0603	0.0710	0.0048	-0.0109	
Barbra	0.0576	0.0448	0.2975	0.2064	
Boat	0.1120	0.1365	0.1258	0.0207	
Lena	0.0170	0.0569	0.1372	0.2942	
Mandril	0.4919	0.4653	0.3758	0.1408	

A. Theoretical Facet

Kerckhoff (1883) stated the principle of security for a cryptographic system, which could easily be extended to Steganography as well. According to Kerckhoff, the security of a public domain cryptographic system resides in its secret key. Likewise, for digital Steganography to work effectively, there must exist some surreptitious bit manipulation mechanism because otherwise contrasting the cover and Stego object shall expose the hidden secret.

- Example: Let s denotes the Stego object computed by some function γs by taking some cover c and the payload m, then the bit embedding γs and extraction ξe processes can be expressed as shown in Eq. 22:

$$
\begin{aligned}
&\% \text{Stego Object} \\
&s = \gamma s(c, m) \\
&\% \text{ Extracting Message Bits from Cover Image} \\
&c' = \xi e(s) \\
&\quad = \xi e(\gamma s(c, m)) \\
&m' = (c, c')?
\end{aligned}
\tag{22}
$$

Interestingly, the above also holds (partially or in full) with the introduction of Stego key k with known cover and the Stego object, shown in Eq. 23. Additionally, with the same cover, it may also unconsciously lead to a Stego key compromise as well.

$$
\begin{aligned}
&s = \gamma s(c, k, m) \\
&c' = \xi e(s) \\
&c' = \xi e(\gamma s(c, k, m)) \\
&m' = (c, c')_k ? \\
&\quad = m \text{ (message segment(s) or full message)}
\end{aligned}
\tag{23}
$$

It follows from Eq. 23 that the mutual information $I_m(c; s)$ explains on the information about c if s is known, as shown in Eq. 24, with H(c), and H(c/s) being the respective entropies:

$$I(c; s) = H(c) - H(c/s) \qquad (24)$$

Eq. 24 equally holds when the secret information (say) E gets embedded inside the cover c, as shown in Eq. 25.

$$I(E, (c, s)) = H(E) - H(E|(c, s)) \qquad (25)$$

Eq. (25) implies that security objectives can not be met when both cover and Stego object is known or unless Eqn. (26) equates to zero:

$$I(E, (c, s)) = 0 \qquad (26)$$

It follows from above that a technically viable solution would then need a dynamic indeterministic bit embedding mechanism capable of giving different results on each of its occurrences, as shown in Eq. 27.

$$
\left.
\begin{array}{l}
\%\text{Stego Object} \\
s = \gamma s(c, r, k, m/r \leftarrow condition\ for\ substitution) \\
\%\ \text{Extracting Message Bits from Cover Image} \\
m = \xi e(s) \\
= \xi e(\gamma s(c, k, m/r \leftarrow conditonal\ extraction)) \\
= [r\ or\ m] \ // \ \text{uncertainty in predicting } r \text{ and } m
\end{array}
\right\} \qquad (27)
$$

, where the Stego key k is different for every new message.

It is evident from Eq. (27) that security lies in disconnecting the one-to-one/linear mapping of cover bits with those of the message bits through some unpredictable phenomena during the bit embedding process.

In practice, TRNGs are known to provide unpredictable results each time these get executed, and the information theory also supports the concept of randomness.

Based on the above analogy, our proposed bit embedding methodology remains novel amongst the recently suggested Image-based secure Steganography solutions.

X. CONCLUSION

In today's world, it is difficult to remain isolated from the digitalization progression that is doubling the data every two years. Hence, regardless of the willful or unconscious data generation, its protection from unauthorized exposure and illegal usage remains the primary concern of information security forefront. In this regard, several information hiding techniques have poured in. Still, the most significant and widely adopted amongst those is that of digital image Steganography because of its aptness in deceiving the human visual system and the redundancy of picture elements to serve as place holders for secret bit embedding. However, recent studies seem short in considering and addressing to its full, the security aspects of Steganalysis, detectability of embedded bits through known cover and known message attack, using Stego key while proposing such furtive schemes or aligning those to information-theoretic perspective. Misconception regarding mean square error (MSE) and peak signal to noise ratio (PSNR) are visible in the tests conducted, which are preferred over the structural similarity index (SSIM). Much of the effort rests in increasing bit hiding capacity in place of squeezing the critical contents to avoid compromise of underneath bit embedding methodology or facilitating in the reusability of the same cover for multiple communication. This research

endeavor attempts to subdue the above-cited limitations by suggesting an information-theoretic secure secret bit embedding Steganography solution through the use of a TRNG. An explicit take away of the proposed study is that the security of LSB Steganography lies in disconnecting the consecutive replacement of secret message bits by inducing uncertainty in the bit embedding process. An implicit assertion of the proposed research is to highlight the need and significance of standardization of the quality scores to remove bias in gauging newly evolved or enhanced Image-based Steganography solution for Universal acceptance.

REFERENCES

[1] A. A. Kharlamov and G. Parry," The impact of servitization and digitization on productivity and profitability of the firm: a systematic approach," Production Planning & Control, pp. 1-13, 2020.

[2] S. S. Arslan, R. Jurdak, J. Jelitto, and B. Krishnamachari," Advancements in distributed ledger technology for Internet of Things," ed:Elsevier, 2020.

[3] M. Wairiya, A. Shah, and G. Sahu," Mobile Learning Adoption: An Empirical Study," in 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 757-761.

[4] R. Ernst and J. Haar," Competitiveness," in Globalization, Competitiveness, and Governability, ed: Springer, 2019, pp. 47-67.

[5] M. Jozani, E. Ayaburi, M. Ko, and K.-K. R. Choo," Privacy concerns and benefits of engagement with social media-enabled apps: A privacy calculus perspective," Computers in Human Behavior, vol. 107, p.106260, 2020.

[6] Gartner," Gartner Says Data and Cyber-Related Risks Remain Top Worries for Audit Executives," 2019.

[7] P. Shukla, H. Kazemian, F. FIET, and C. Eng," Privacy in The First Line of the First Code," Science Magazine, vol. 17, p. 04, 2020.

[8] K. Sanjeev, B. Janet, and R. Eswari," Automated Cyber Threat Intelligence Generation from Honeypot Data," in Inventive Communication and Computational Technologies, ed: Springer, 2020, pp. 591-598.

[9] A. Stetsenko," Transformative-Activist and Social Justice Approaches to the History of Psychology," in Oxford Research Encyclopedia of Psychology, ed, 2020.

[10] V. Kvashis and Y. Sluchevskaya," Limits of Acceptable State Interference in Privacy," in XVII International Research-to-Practice Conference dedicated to the memory of MI Kovalyov (ICK 2020), 2020, pp. 255-258.

[11] Y. Lu and S. Li," From data flows to privacy issues: a user-centric semantic model for representing and discovering privacy issues," in Proceedings of 53rd Hawaii International Conference on System Sciences, 2020.

[12] O. A. Al-Harbi, W. E. Alahmadi, and A. O. Aljahdali," Security analysis of DNA based Steganography techniques," SN Applied Sciences, vol. 2, pp. 1-10, 2020.

[13] A. Singh," Cryptography: A Never Ending Technology," CYBERNOMICS, vol. 2, pp. 45-47, 2020.

[14] M. G. Vigliotti and H. Jones," Cryptography for Busy People," in The Executive Guide to Blockchain, ed: Springer, 2020, pp. 23-40.

[15] R. Shanthakumari and S. Malliga," Dual layer security of data using LSB inversion image Steganography with elliptic curve cryptography encryption algorithm," Multimedia Tools and Applications, vol. 79, pp. 3975-3991, 2020.

[16] D. John F," Review of The Third Reich is Listening by Christian Jennings," Cryptologia, vol. 44, pp. 91-95, 2020.

[17] O. Rachael, S. Misra, R. Ahuja, A. Adewumi, F. Ayeni, and R.Mmaskeliunas," Image Steganography and Steganalysis Based on Least Significant Bit (LSB)," in Proceedings of ICETIT 2019, ed:Springer, 2020, pp. 1100-1111.

[18] N. T. Courtois, M.-B. Oprisanu, and K. Schmeh," Linear cryptanalysis and block cipher design in East Germany in the 1970s," Cryptologia, vol. 43, pp. 2-22, 2019.

[19] N. T. Courtois and M. Georgiou," Constructive non-linear polynomial cryptanalysis of a historical block cipher," arXiv preprint arXiv:1902.02748, 2019.

[20] G. Wu, F. Zhang, L. Shen, F. Guo, and W. Susilo," Certificateless aggregate signature scheme secure against fully chosen-key attacks," Information Sciences, vol. 514, pp. 288-301, 2020.

[21] M. Bouam, C. Bouillaguet, and C. Delaplace," Brute-Force Cryptanalysis with Aging Hardware: Controlling Half the Output of SHA-256," 2019.

[22] S. Khatoon and B. Singh Thakur," Cryptanalysis and improvement of authentication scheme for roaming service in ubiquitous network," Cryptologia, pp. 1-26, 2020.

[23] C. B. Smith," The Comparison of Steganography and Cryptography:Concealing Information," Utica College, 2019.

[24] N. F. Johnson and S. Jajodia," Exploring Steganography: Seeing the unseen," Computer, vol. 31, pp. 26-34, 1998.

[25] R. J. Anderson and F. A. Petitcolas," On the limits of Steganography," IEEE Journal on selected areas in communications, vol. 16, pp. 474-481, 1998.

[26] S. Arunkumar, V. Subramaniyaswamy, V. Vijayakumar, N. Chilamkurti, and R. Logesh," SVD-based robust image steganographic scheme using RIWT and DCT for secure transmission of medical images," Measurement, vol. 139, pp. 426-437, 2019.

[27] "Definitions.", LIA - Laboratory of Advanced Research on Computer Science, 2020.

[28] R. Gonzalez and R. Woods," Digital Image Processing 3rd edn Pearson Prentice Hall," 2008.

[29] S. Boutnaru," Steganography obsfucation," ed: Google Patents, 2020.

[30] H.-C. Wu, N.-I. Wu, C.-S. Tsai, and M.-S. Hwang," Image steganographic scheme based on pixel-value differencing and LSB replacement methods," IEE Proceedings-Vision, Image and Signal Processing, vol. 152, pp. 611-615, 2005.

[31] R. Kavitha, U. Eranna, and M. Giriprasad," DCT-DWT Based Digital Watermarking and Extraction using Neural Networks," in 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), 2020, pp. 1-5.

[32] A. Jalali and H. Farsi," A new Steganography algorithm based on video sparse representation," Multimedia Tools and Applications, vol.79, pp. 1821-1846, 2020.

[33] C. W. Kurak Jr and J. McHugh," A cautionary note on image downgrading," in ACSAC, 1992, pp. 153-159.

[34] D. Ghosh, A. K. Chattopadhyay, K. Chanda, and A. Nag," A Secure Steganography Scheme Using LFSR," in Emerging Technology in Modelling and Graphics, ed: Springer, 2020, pp. 713-720.

[35] C. Pak, J. Kim, K. An, C. Kim, K. Kim, and C. Pak," A novel color image LSB Steganography using improved 1D chaotic map," Multimedia Tools and Applications, vol. 79, pp. 1409-1425, 2020.

[36] A. Chatterjee, S. K. Ghosal, and R. Sarkar," LSB based Steganography with OCR: an intelligent amalgamation," Multimedia Tools and Applications, pp. 1-19, 2020.

[37] M. A. Al Mamun, S. M. Alam, M. S. Hossain, and M. Samiruzzaman," A Novel Image Steganography Using Multiple LSB Substitution and Pixel Randomization Using Stern-Brocot Sequence," in Future of Information and Communication Conference, 2020, pp. 756-773.

[38] N. Soni, I. Saini, and B. Singh," Integer Wavelet Transform-Based ECG Steganography for Hiding Patients' Confidential Information in e-Healthcare Systems," in Soft Computing: Theories and Applications, ed: Springer, 2020, pp. 513-525.

[39] A. K. Sahu and G. Swain," Reversible Image Steganography Using Dual-Layer LSB Matching," Sensing and Imaging, vol. 21, p. 1, 2020.

[40] A. Jain," A Secured Steganography Technique for Hiding Multiple Images in an Image Using Least Significant Bit Algorithm and Arnold Transformation," in International Conference on Intelligent Data Communication Technologies and Internet of Things, 2019, pp. 373-380.

[41] B. Praveen, D. Samanta, G. Prasad, C. R. Kumar, and M. Prasad," Protecting Medical Research Data Using Next Gen Steganography Approach," in International Conference on Information, Communication and Computing Technology, 2019, pp. 340-348.

[42] A. Saikia and T. Tuithung," A Novel True Colour Image Bit Modification Technique for Image Steganography," in International Conference on Soft Computing and Signal Processing, 2019, pp. 317-327.

[43] N. S. R. Chandra, V. Sneha, and P. V. Paul," A Novel Image Steganography Model Using LSB with Extended ASCII Codes," in Smart Intelligent Computing and Applications, ed: Springer, 2020, pp.107-116.

[44] G. Luo, R.-G. Zhou, and Y. Mao," Two-level information hiding for quantum images using optimal LSB," Quantum Information Processing, vol. 18, p. 297, 2019.

[45] B. Datta, S. Roy, S. Roy, and S. K. Bandyopadhyay," Multi-bit robust image Steganography based on modular arithmetic," Multimedia Tools and Applications, vol. 78, pp. 1511-1546, 2019.

[46] D. Nashat and L. Mamdouh," An efficient steganographic technique for hiding data," Journal of the Egyptian Mathematical Society, vol. 27, pp. 1-14, 2019.

[47] P. Artiemjew and A. Kislak-Malinowska," Using r-indiscernibility Relations to Hide the Presence of Information for the Least Significant Bit Steganography Technique," in International Conference on Information and Software Technologies, 2019, pp. 209-220.

[48] N. M. Al-Aidroos and H. A. Bahamish," Image Steganography Based on LSB Matching and Image Enlargement," in 2019 First International Conference of Intelligent Computing and Engineering (ICOICE), 2019, pp. 1-6.

[49] A. K. Sahu and G. Swain," A novel n-rightmost bit replacement image Steganography technique," 3D Research, vol. 10, p. 2, 2019.

[50] Pichardo-M'endez, JL and Palacios-Luengas, L and Mart' ınez-Gonz'alez, RF and Jim' enez-Ram 'ırez, O and V'azquez-Medina, R," LSB Pseudorandom Algorithm for Image Steganography Using Skew Tent Map," Arabian Journal for Science and Engineering, pp. 1-20, 2019.

[51] I. Maurya and S. Gupta," Inverted LSB Image Steganography," in Soft Computing: Theories and Applications, ed: Springer, 2020, pp. 19-29.

[52] G. Swain," Very high capacity image Steganography technique using quotient value differencing and LSB substitution," Arabian journal for science and engineering, vol. 44, pp. 2995-3004, 2019.

[53] P. Agarwal, D. Moudgil, and S. Priya," Encrypted Transfer of Confidential Information Using Steganography and Identity Verification Using Face Data," in Artificial Intelligence and Evolutionary Computations in Engineering Systems, ed: Springer, 2020, pp. 155-166.

[54] A. Kerckhoffs," La cryptographie militaire. 9: 5–38," ed: January, 1883.

[55] J. Von Neumann," 13. various techniques used in connection with random digits," Appl. Math Ser, vol. 12, p. 5, 1951.

[56] G. J. Simmons," The prisoners' problem and the subliminal channel," in Advances in Cryptology, 1984, pp. 51-67.

[57] Z. Wang and E. P. Simoncelli," Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in Human Vision and Electronic Imaging X, 2005, pp. 149-159.

[58] J. Z¨ollner, H. Federrath, H. Klimant, A. Pfitzmann, R. Piotraschke, A. Westfeld, et al.,"Modeling the security of steganographic systems," in International Workshop on Information Hiding, 1998, pp. 344-354.

[59] J. Wen, X. Zhou, P. Zhong, and Y. Xue," Convolutional neural network based text steganalysis," IEEE Signal Processing Letters, vol. 26, pp.460-464, 2019.

[60] H. Lee and H.-W. Lee," New Approach on Steganalysis: Reverse-Engineering based Steganography SW Analysis," in Proceedings of the 2020 9th International Conference on Software and Computer Applications, 2020, pp. 212-216.

[61] F. Nabi," A Survey on Image Steganography." academia.edu

[62] S. Trivedi and R. Chandramouli," Active steganalysis of sequential Steganography," in Security andWatermarking of Multimedia Contents V, 2003, pp. 123-130.

[63] J. Kelley," Terror groups hide behind Web encryption," USA today, vol. 5, p. 2001, 2001.

[64] J. Cosic and M. Baˇca," Steganography and steganalysis-does local web sites contain "Stego" contents?," in Proceedings ELMAR-2010, ed, 2010.

[65] G. Rajput and R. Agrawal," Evaluation of feature selection measures for Steganalysis," in International Conference on Pattern Recognition and Machine Intelligence, 2009, pp. 432-439.

[66] A. D. Ker," A general framework for structural Steganalysis of LSB replacement," in International Workshop on Information Hiding, 2005, pp. 296-311.

[67] J. Fridrich and M. Goljan," On estimation of secret message length in LSB Steganography in spatial domain," in Security, Steganography, and watermarking of multimedia contents VI, 2004, pp. 23-34.

[68] A. D. Ker and R. B¨ohme," Revisiting weighted Stego-image Steganalysis," in Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, 2008, p. 681905.

[69] A. Westfeld and A. Pfitzmann," Attacks on steganographic systems," in International workshop on information hiding, 1999, pp. 61-76.

[70] O. Dabeer, K. Sullivan, U. Madhow, S. Chandrasekaran, and B. Manjunath," Detection of hiding in the least significant bit," IEEE Transactions on Signal Processing, vol. 52, pp. 3046-3058, 2004.

[71] S. Dumitrescu, X. Wu, and N. Memon," On Steganalysis of random LSB embedding in continuous-tone images," in Proceedings. International Conference on Image Processing, 2002, pp. 641-644.

[72] L. Fillatre," Adaptive Steganalysis of least significant bit replacement in grayscale natural images," IEEE Transactions on Signal Processing, vol. 60, pp. 556-569, 2011.

[73] M. A. Alsmirat, R. A. Al-Hussien, W. a. T. Al-Sarayrah, Y. Jararweh, and M. Etier," Digital video forensics: a comprehensive survey," International Journal of Advanced Intelligence Paradigms, vol. 15, pp.437-456, 2020.

[74] L. A. Sandoval-Bravo, V. I. Ponomaryov, R. Reyes-Reyes, and C.Cruz-Ramos," Coverless image Steganography framework using distance local binary pattern and convolutional neural network," in Real-Time Image Processing and Deep Learning 2020, 2020, p. 114010D.

[75] A. Gutub and F. Al-Shaarani," Efficient Implementation of Multiimage Secret Hiding Based on LSB and DWT Steganography Comparisons," Arabian Journal for Science and Engineering, pp. 1-14, 2020.

[76] T. Sudhakar, S. S. V. Sriraman, and N. Venkateswaran," Synthesis and Evaluation of Improved Reference Matrix Models for High Capacity Image Steganography," in 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), 2020, pp. 1-6.

[77] T. Rabie, M. Baziyad, and I. Kamel," High Payload Steganography:Surface-Fitting The Transform Domain," in 2019 International Conference on Communications, Signal Processing, and their Applications (ICCSPA), 2019, pp. 1-6.

[78] U. Pilania and P. Gupta," A Proposed Optimized Steganography Technique using ROI, IWT and SVD," International Journal of Information Systems & Management Science, Forthcoming, 2019.

[79] S. J. Gladwin and P. L. Gowthami," Combined Cryptography and Steganography for Enhanced Security in Suboptimal Images," in 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), 2020, pp. 1-5.

[80] H. S. Radeaf, B. M. Mahmmod, S. H. Abdulhussain, and D. Al-Jumaeily," A Steganography based on orthogonal moments," in Proceedings of the International Conference on Information and Communication Technology, 2019, pp. 147-153.

[81] A. R. Idrais, S. Harb, M. O. Ahmad, and M. Swamy," A Novel High Capacity Data Hiding Algorithm using Salt and Pepper Noise," in 2020 11th International Conference on Information and Communication Systems (ICICS), 2020, pp. 131-135.

[82] X. Duan, D. Guo, N. Liu, B. Li, M. Gou, and C. Qin," A New High Capacity Image Steganography Method Combined With Image Elliptic Curve Cryptography and Deep Neural Network," IEEE Access, vol. 8, pp. 25777-25788, 2020.

[83] A. Seif and W. Alexan," A High Capacity Gray Code Based Security Scheme for Non-Redundant Data Embedding," in 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE), 2020, pp. 130-136.

[84] W. Liu, X. Yin, W. Lu, J. Zhang, J. Zeng, S. Shi, et al.," Secure halftone image Steganography with minimizing the distortion on pair swapping," Signal Processing, vol. 167, p. 107287, 2020.

[85] J. Greffier, J. Frandon, A. Larbi, J. Beregi, and F. Pereira," CT iterative reconstruction algorithms: a task-based image quality assessment," European radiology, vol. 30, pp. 487-500, 2020.

[86] A. K. Moorthy and A. C. Bovik," Visual quality assessment algorithms:what does the future hold?," Multimedia Tools and Applications, vol. 51, pp. 675-696, 2011.

[87] L. Zhang, L. Zhang, X. Mou, and D. Zhang," A comprehensive evaluation of full reference image quality assessment algorithms," in 2012 19th IEEE International Conference on Image Processing, 2012, pp. 1477-1480.

[88] B. Lakshmi Sirisha," Image Steganography based on SVD and DWT techniques," Journal of Discrete Mathematical Sciences and Cryptography, pp. 1-8, 2020.

[89] S. T. Abdulrazzaq, M. M. Siddeq, and M. A. Rodrigues," A novel Steganography approach for audio files," SN Computer Science, vol. 1, pp. 1-13, 2020.

[90] W. Lyu, W. Lu, and M. Ma," No-Reference Quality Metric for Contrast-Distorted Image Based on Gradient Domain and HSV Space," Journal of Visual Communication and Image Representation, p.102797, 2020.

[91] A. Mittal, A. K. Moorthy, and A. C. Bovik," No-reference image quality assessment in the spatial domain," IEEE Transactions on image processing, vol. 21, pp. 4695-4708, 2012.

[92] A. Mittal, R. Soundararajan, and A. C. Bovik," Making a "completely blind" image quality analyzer," IEEE Signal Processing Letters, vol.20, pp. 209-212, 2012.

[93] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani," Blind image quality evaluation using perception based features," in 2015 Twenty First National Conference on Communications (NCC), 2015, pp. 1-6.

[94] A. Shnayderman, A. Gusev, and A. M. Eskicioglu," An SVD-based grayscale image quality measure for local and global assessment," IEEE transactions on Image Processing, vol. 15, pp. 422-429, 2006.

[95] C.-W. Kok and W.-S. Tam, Digital Image Interpolation in Matlab: John Wiley & Sons, 2019.

[96] U. Virtebi," The USC-SIPI Image Database."

[97] Ker, Andrew D and Pevn'y, Tom' aˇs and Kodovsk`y, Jan and Fridrich, Jessica," The square root law of steganographic capacity," in Proceedings of the 10th ACM workshop on Multimedia and security, 2008, pp. 107-116.

[98] T. Zhang and X. Ping," A new approach to reliable detection of LSB Steganography in natural images," Signal processing, vol. 83, pp. 2085-2093, 2003.

[99] A. Saeed, M. J. Khan, H. Shahid, S. I. Naqvi, M. A. Riaz, M. S. Khan, et al.," An Accurate Texture Complexity Estimation for Quality-Enhanced and Secure Image Steganography," IEEE Access, vol. 8, pp.21613-21630, 2020.

[100] S. S. Agrawal and R. M. Samant," Data hiding in grayscale images using pixel value differencing," in Technology Systems and Management, ed: Springer, 2011, pp. 27-33.

[101] R. Y. Li, O. C. Au, K. K. Lai, C. K. Yuk, and S.-Y. Lam," Data hiding with tree based parity check," in 2007 IEEE International Conference on Multimedia and Expo, 2007, pp. 635-638.

[102] H. Al-Dmour, N. Ali, and A. Al-Ani," An efficient hybrid Steganography method based on edge adaptive and tree based parity check," in International Conference on Multimedia Modeling, 2015, pp. 1-12.

[103] M. A. Hameed, M. Hassaballah, S. Aly, and A. I. Awad," An Adaptive Image Steganography Method Based on Histogram of Oriented Gradient

and PVD-LSB Techniques," IEEE Access, vol. 7, pp. 185189-185204, 2019.

[104] S. Prasad and A. K. Pal," Logistic map-based image Steganography scheme using combined LSB and PVD for security enhancement," in Emerging Technologies in Data Mining and Information Security, ed:Springer, 2019, pp. 203-214.

[105] S. Chakraborty and A. S. Jalal," A novel local binary pattern based blind feature image Steganography," Multimedia Tools and Applications, pp. 1-14, 2020.

[106] M. Nazari and I. D. Ahmadi," A novel chaotic Steganography method with three approaches for color and grayscale images based on FIS and DCT with flexible capacity," Multimedia Tools and Applications, pp. 1-32, 2020.

[107] N. K. Pandey and M. Diwakar," A Review on Cloud based Image Processing Services," in 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), 2020, pp.108-112.

[108] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli," Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, pp. 600-612, 2004. https://www.cns.nyu.edu/~lcv/ssim/

[109] Hor' e, A., and D. Ziou." Image quality metrics: PSNR versus SSIM." In Proceedings of the 2010 IEEE 20th International Conference on Pattern Recognition, vol. 1. 2010.

[110] A. Hanjalic, G. Langelaar, P. Van Roosmalen, J. Biemond, and R. Lagendijk, Image and video databases: restoration, watermarking and retrieval: Elsevier, 2000.

[111] Nag, Amitava. (2015). Re: What is the best PSNR value for the steganography method to hide the text in image?. Retrieved from: https://www.researchgate.net/post/What˙is˙the˙best˙PSNR˙value˙for˙the ˙steganography˙method˙to˙hide˙the˙text˙in˙image/54e5a9d5d11b8b330b 8b4581/citation/do- wnload.

AUTHOR'S PROFILE

Khan Farhan Rafat is a self-motivated individual having practical experience in the evolution, designing, and implementation of information security solutions together with formulation and drafting of allied security policies and procedures. He also possesses progressive experience in managing and securing IT operations within complex working environments. An enthusiastic, innovative individual who multitasks and has an excellent sense of counter strike to get results by instilling commitment, trust, fairness, and loyalty. Strengths include a strong sense of leadership, proficient communication and problem-solving skills and acts as a change catalyst. The first known individual among his compatriots to have furnished a Ph.D. (Computer Science) dissertation in ASCII Text Steganography, a research area regarded as the most difficult to comprehend by the gurus of the particular trait. Holding Master Degrees in Information Security, Project Management, and Telecommunication, he has contributed the research arena with international peer-reviewed journals and conference publications besides being the winner of the best conference paper award in Dubai's ICPINE, January 30-31, 2017.

Towards Securing Cloud Computing from DDOS Attacks

Ishtiaq Ahmed[1], Sheeraz Ahmed[2]
Department of Computer Science
Iqra National University
Peshawar, Pakistan

Asif Nawaz[3]
Faculty of Engineering
Higher College of Technology
Dubai, UAE

Sadeeq Jan[4]
Department of CS and IT
National Center for Cyber Security
UET Peshawar, Pakistan

Zeeshan Najam[5]
Department of Electrical Engineering
MNS Uni of Engg and Technology
Multan, Pakistan

Muneeb Saadat[6], Rehan Ali Khan[7]
Department of Electrical Engineering
University of Science and Technology
Bannu, Pakistan

Khalid Zaman[8]
Department of Computer Science
Near East University North Cyprus
Mersin 10, Turkey

Abstract—Cloud computing (CC) is an advanced technology that provides data sharing and access to computing resources. The cloud deployment model represents the exact type of cloud environment based on ownership, size, and accessibility rights, and also describes the purpose and nature of the cloud. Since all processes today are computerized, consumers need a lot amount of data and cache size. The security of the cloud is ensured in many levels, but the scope of intrusions makes it necessary to understand the factors that affect cloud security. CC-certified users rely on third parties for their other important security issues in third-party computing clouds. A DDoS attack is an attack-type in which it is not necessary to send a large number of packets to the server, which makes it impossible for legitimate users to access them. In this research work, a DDoS attack was launched and a tool for launching a DDoS attack was discussed. In this research, DDoS attacks were rejected using three different SNORT rules. In this research, rules predefined for detecting DDoS attacks on SNORT profiles detect and prevent DDoS attacks, but because they block certain legitimate requests and generate false alarms, this should be the subject of future research.

Keywords—*Cloud computing; denial of service; SNORT rules; network; energy consumption*

I. INTRODUCTION

Cloud computing, a popular topic in the past few years involves various technologies and provides scalable IT related services over the Internet. Cloud computing is the use of various services, such as software development platforms, servers, storage and software, over the internet, often referred to as the "cloud." Emergence of cloud computing technologies has changed the way we store, retrieve, and archive our data. With the promise of unlimited, reliable and always-available storage, a lot of private and confidential data are now stored on different cloud platforms.

Cloud computing is a concept used primarily in computer science, but in the last decade, the term "cloud computing" has always been used in the field of library and information science, as well as in other areas like Business, Industry, Medical Science and Corporate sector, etc. is also being done the application of cloud computing.

On the Internet, cloud computing (CC) is an advanced technology that provides data sharing and access to computing resources. CC is an Internet-based environment that provides services such as storage, applications, and servers [1,2]. CC is a very easy and fun method for today's consumers to use the Internet and do professional with CC. Since it is about providing remote resource resources to alleviate consumer problems, they need to use only those resources. They do not have to pay for local services such as infrastructure or storage. This atmosphere can be seen as a novel sculptural archetypal that offers greater flexibility and lower cost availability as shown in Fig. 1. Data and resources are available anytime, anywhere and accessible on the Internet. CC permits you to launch your own applications, software, and hosts on a virtual server, which can be restored when required. For example, Google App Engine, societal networks, Google Docs, AWS, etc.

CC provides a number of steps, including utilities and grid computing. Grid computing is a large-scale, decentralized calculation that provides a direct way to access a variety of useful resources [3,4]. To manage the resources of its users, service providers have been enabled through utility computing [5]. In 1960, ARPA (Advanced Research Institute) in the United States began to realize the connectivity of integrated devices [6]. They are concerned about integrated devices

because ARPA agents have different branch sizes and participate in different functions of branch search. ARPA funds its employees so they can find new ideas. Therefore, the agency must connect with the distribution agencies and share their personal efforts with everyone to achieve the best results. To this end, the IRPA's IRAPNET is proposed and four different branches have been established: Stanford Research Institute (SR), the University of California, Santa Barbara, and the University of Utah. These devices do not affect each other. The ARPA [7,8,9] proposed the current IMP protocol (the current message processor) for this purpose. IMP is designed to make communiqué conceivable and act corresponding a gateway. The ARPA connects its four twigs together, with branches from apiece host, and apiece branch connects with another branch of the IMP.

Since cloud technology offers a lot of benefits to consumers, these benefits must be classified with respect to the user's needs. The cloud deployment model represents the exact type of cloud environment based on ownership, size, and accessibility rights, and also describes the purpose and nature of the cloud [10].

A. Security in CC

Data safety transmitted to the Internet in the Cloud Configuration field is actually clouded computing security. Since all processes today are computerized, consumers need a lot amount of data and cache size. In CC, users must request that their data be stored and retrieved [11,12]. In this environment more storage capacity and services are accessed in many places. For cloud breadwinners, safety is an ever-present tricky, and security is a huge challenge regardless of Internet access. Because as soon as a real user receives their information from one place, the interloper can entree the information from alternative site, which may prevent the original user from approaching.

Fig. 1. Internet Depicted as Cloud in a Network.

B. DDoS Attacks in CC

DDoS attacks can disable cloud services at any time. The amount of DDoS ravages increased by 51% in 2013, but companies using DDoS security services can prevent these attacks. Denial of service attack usually creates a working server, system, or network that can advertise the target asset and prevent real clients from using the resource. The target of DOS Attack is the same as the target of DoS Attack. However, these attacks are a distributed form of DoS attacks, meaning they come from many locations and target a single victim [13,14,15].

The benefit of denying the distribution of service attacks is when attacks become accessible, expandable, flexible, and accessible from somewhere, because denial of service attacks is due to dissimilar sites [16]. Several hosts are involved in the ravage. In fact, they are named managers and representatives. An attacker prepares the first two or more managers and, with the help of these managers, manages the agent so that the service refuses to split the attack.

DDoS attacks are increasing speedily, causing great harm to large businesses and economic loss to global businesses and websites [17]. Although the Department of Defense condemns funding for targeted attacks, the economic withdrawal is often not the cause of such attacks. In each case, the attacker claims to destroy the company or sole attacker; in other cases, the attacker only tries to hit the target, causing the most damage, or hitting too many targets. There is a loss. When identifying a partner in denial of service attack, the target of the attack may be displayed. DDOS attacks architecture is explained diagrammatically in Fig. 2.

Many DoS attacks are actually split, and attack traffic comes from different systems. Although DOS attacks generated by a single source are more likely to be downplayed, as defense networks can intercept traffic through malicious sources, attacks from different systems are difficult to identify and defend [18]. The attack is because it's difficult to distinguish malicious packets and actual traffic.

Fig. 2. DDoS Attacks Architecture.

II. METHODOLOGY

DDoS Spell is a massively synchronized ravage on the search network's resources or the availability of the service. By managing a large number of hostages outside the network, a large number of data packets are sent to the target node to attack the DDoS. Victims use high traffic bandwidth and do not permit it to reach the victim of another large package [19]. There are many security vulnerabilities due to CC availability and scalability. The existence of DDoS over the network greatly increases the loss of packets, and causes security issues on the network. To identify such kind of attack various techniques are used in the previous researches like entropy-based anomaly recognition, this technique.

In the proposed work we test two approaches for the recognition of DDoS ravage, SNORT and Mutual Egress Filtering Approach following are the topology and configurations for both approaches. In this research work, we test two approaches for the recognition of DDoS ravage, SNORT and Mutual Egress Filtering approach following are the topology and configurations for both approaches.

A. Network Topology

Network topology is in what way we create networks to simulate DDoS ravage and sense it in the cloud. Fig. 3 temporarily describes the topology.

The topology has been implemented in Gns3 and consists of four networks. For the invaders, two different networks started for the ravage. The attack bump with IP address 1.1.1.2 has been fitted as Virtual Machine in VMware [20]. Another attacker with our IP address of 10.0.0.2 is our hosting system. The test server and SNORT node are also installed in VMware. The test node checks if the server responds after starting a DDoS attack. Install the SNORT to sense DDoS ravage and keep between these servers and the revenue generating network.

Windows Server 2012 is fitted in VMware and includes Internet Information Services (IIS). IIS is contained within in the server window and consist of the default Web site. The default Windows Server 2012 site is easy to manage.

Same topology is used for Mutual Egress Filtering Approach but instead of SNORT an Access Control List is configured on egress router.

1) Configuration of SNORT: First, the router is organized with four edges and assigns IP address to each interface:

a) Fast Ethernet (0/0) and the IP address of the line is 10.0.0.1

b) Fast Ethernet (0/1) and the IP address of the line is 1.1.1.1.

c) Fast Ethernet (4/0) and the IP address of the line is 192.168.2.2.

d) Fast Ethernet (4/1) and the IP address of the line is 192.168.0.2.

Then, configure SNORT in VMware. SNORT performs according to the rules, and the rule no 1 has already loaded SNORT. The loaded instructions hold a file because these files are similarly in the established SNORT. We need to use rules

to copy some of the downloaded folders and then go to into the installed directory of SNORT software and paste them there. In these files, there is a local named folder. Local can be edited by rule users. Notepad, Notepad++ or any folder editor in the file editor can modify rules or create your own rules according to your needs. We also changed the rules for detecting DDoS ravage in CC.

dropTCP any any> any 80 (\msg:"Reset outside window"; \ detection_filter:trackby_dst, count 30, seconds 1; \ new_action drop; timeout 50; sid:1000001;)".

This is our instruction for detecting DDoS ravage in CC and has been further to the local SNORT folder. Nodes that install SNORT also have two dissimilar boundaries: one that connects straight to the external network and one that connects to the internal server, and that is victim of the attacker. The IP addresses of the two edges are 192.168.0.3 and 2.0.0.1 and are linked over a network adapter. Install Windows Server 2012 and configure IIS (information on the Internet) on the server. Installing Windows Server 2012 is as meek as installing somewhat Windows OS. The IP address of the server is 2.0.0.2, and the hacker points to the address. The attacker's node has been configured as if it contained the LOIC letter, which is explained in Fig. 4.

Fig. 3. Network Topology.

Fig. 4. VMware Platform.

B. Simulator Setup

To test the impact of DDoS ravage and detect them in the cloud using SNORT, the following tools were configured.

1) Graphical Network Simulator (GNS3): GNS3 is used to create network topologies and simulate various kinds of ravage. Fig. 3, variant devices are displayed, consists of routers, switches, firewalls, and computers. These all devices are used in an actual network location. GNS3 uses these devices virtually, providing us with an atmosphere to test and model various kinds of topologies. In order to use GNS3IOS, the actual CISCO device images are used, which are really the OS (operating systems) of the router [21]. Users can initial fetch these pictures to the router configuration in the GNS3 simulator. Lacking an IOS image, the router will not be able to perform its operations.

2) VMware: A database which permits employers to develop many virtual machines by using a solo workstation is VMware. On the VMware stage, you can improve multiple virtual machineries to one physical apparatus, and the virtual machine can run as per another. It provides excellent routine when you need to practice a different operating system or other server in your user's system. For example, you can install a server and a Windows operating system in VMware to test the network or launch extra applications. It is informal to connect in any OS (operating system.

3) LOIC (Low Orbit Ion Canon): LOIC is an uncluttered basis net anxiety test device used to refuse service attacks. LOIC (for Microsoft Windows and Mac OS X) is a flood-making instrument for creating a large number of network streams for use with network resources or applications. This amount of traffic can be a result of service failure or lack of visibility due to server or application downtime. Users can use LOIC to refuse service attacks and reload the server using native TCP or UDP TCP packets.

Fig. 5 shows three parameters that trigger the attack: the target, the preparation, and the attack parameters.

- When selecting a destination, there will be a URL and an IP address. Click the URL to specify the target service URL. The IP Tap is secondhand to enter the IP address of the target package [22].

- Underneath the "Attack" choice, there are port challenges, technologies and risks. The number of ports that connect to the port for the application are identified. This technique is castoff for TCP, UDP, and HTTP, and risk associations are castoff to determine the amount of fears that arise.

- Once all the necessities for the attack have been determined, the ready state is reached. In the "Complete" section, there is a button that you can click to launch an attack.

Fig. 5.   LOIC A DDoS Attacking Tool.

C. Proposed Solution for DoS and DDoS Attacks

Various methods have been previously used to sense DoS and DDoS attacks in CC. In our effort, SNORT is used to sense DoS and DDoS. SNORT is software containing a list of instructions. It is also described in section, using these SNORT instructions to sense various kinds of attacks. Users can adjust these rules as needed. In a previous study [23], DoS and DDoS attacks were detected using SNORT according to the following rules:

> "dropTCP any any> any 80 (\msg:"Reset outside window"; \ detection_filter:trackby_src, count 30, seconds 1; \ new_action drop; timeout 50; sid:1000001;)".

This instruction triggers the SNORT device to sense a TCP appeal from any basis on port 80 (if more than 30 requests/sec) and cannot establish a connection with the IP address for 50 seconds.

In our effort, we made some variations to the rules, because the above rules only sense DoS attacks, which mean that this instruction can't sense DDoS attacks, since in the rules by_src, which means from any source An IP address, is detected every second for any IP address. The changes we made to detect DDoS attacks,

> "dropTCP any any> any 80 (\msg:"Reset outside window"; \ detection_filter: trackby_dst, count 30, seconds 1; \ new_action drop; timeout 50; sid:1000001;)".

In this instruction, we practice by_dst to sense the IP address close to its target and have found multiple IP addresses. Another solution to detect DoS and DDoS attack is combine both the rules and create another rule which will sense DoS and DDoS attacks from both sides either from source side or also from destination side. The rule is written as:

if (by_src,count==30, seconds 1;\)
```
    {
"droptcp                                              any
any>80(\detection_filter:new_actiondrop;timeout
50;sid1000001;)".})
    }
else
{
```
"dropTCP any any> any 80 (\msg:"Reset outside
window"; \ detection_filter:trackby_src, count 30, seconds 1; \
new_action drop; timeout 50; sid:1000001;)".
```
    }
```

This law detects DoS and DDoS attacks from both ends, according to this rule SNORT monitors traffic which is coming from one source, and from multiple sources as well.

1) Design procedure: After identifying different approaches for sensing DDoS ravage in CC, the suitable approach is used to detect DDoS ravage. In the proposed effort we implement two different approaches that is SNORT as IDS and Mutual Egress Filtering Approach. After testing both the approaches we make a comparative analysis that which approach is better for preventing DDoS attack in Cloud Environment. Fig. 6 describes the design procedure of our proposed work.

Fig. 6. Block Diagram for SNORT.

In the proposed work first we test IDS SNORT, for the detection of DDoS attack. Following are the major steps of our proposed solution.

- Start communication.

- SNORT implementation.

- Observation of data.

- Detection of DDoS attack.

- End of communication.

After creating the topology and setting up network configurations, communication is started. SNORT AN ID is implemented with predefined rules to observe the flow of data. In the proposed work we detect DDoS attack, for that a rule is written in the base file of SNORT, which is used for the recognition of DDoS attack. If the rule is triggered, then the flow of data is harmful. SNORT will aware the network administrator about the invasion and drop connection with the corresponding communicating party. If the rule is not triggered, that means that the flow of data is normal.

III. SIMULATION AND RESULTS

To use SNORT [18], the network topology and its format are configured. The DDoS ravage has been launched. It spoke fleetingly about its impact on cloud services; fixed between SNORT server and received network. The latest LOIC tool (Canon Ion Orbit) is also under discussion. This chapter contains various SNORT rules, simulation analysis, and implementation and comparison results.

To integrate shared DDoS features, a virtual background has been created. The virtual machine consists of a client and a cloud server. It is placed in different areas of the earth. SNORT is built in the center of external networks and servers. Locks can be used to test DOS attacks. This requires either the source IP address (board service) or the source URL, the port address (port numeral of the board bid), and the nature of the ravage under the target service. What kind of attack is TCP, UDP or HTTP form should apply. Once all these parameters have been defined, when you click this button, the Ready button will appear and the saturation of the selected service will begin. SNORT is used to detect attacks. SNORT contains config files. Here, the home network is defined as the net address of the server, and the exterior network is defined as something that isn't a homebased network. We can protect our home network from DDoS attacks. An external network is a network from outside or from anywhere.

We employed SNORT with two dissimilar directions. The last is the earlier search rule, which detects DDoS attacks, but not both. It can sense only one ravager at a time. Another principle is that we have to make variations to sense many ravagers at the same time and we have to give better results than the previous search rules. After testing SNORT, we implemented a mutual sewerage filtering method that completely prevented the attacker from flooding the server.

A. Monitoring Server Performance when DDoS Attack has not been Launched

Fig. 7 displays reserve monitoring previously any contact with the server: The chart in figure overhead shows six CPU procedure at the beginning of the performance check. The processor uses three percent, and its load reaches six percent after a period of time, and through continuous monitoring, it causes the processor to use three or four percent. According to some permissible necessities, it reaches 8 percent and 9 percent but no further than 10 percent.

B. Monitoring Server Performance with DDoS Attack Launched

Afterward starting the DDoS ravage, if SNORT is not implemented and the TCP SYN ravage is launched from two dissimilar sites, then the server act is checked. Fig. 8 shows how DDoS attacks use server properties.

Initial, there is no server request and CPU load are low. When an attacker knockouts server, server responds to the ravagers request. As a TCPSYN ravage, its server is on relevant IP address (attacker's IP address) to establish a TCP connection and three-way connection between client and server during TCP linking. First, the user directs a SYN package, which means that user wishes to launch a TCP linking to the server and responds with the SYN response and a salutation package. The client sends an identity packet. After this process, "Establishes the original connection", but does not wait for a response from the attacker server, and continuously sends the SYN packet. As a consequence, the server becomes busy, so sincere users cannot respond, as shown in Fig. 8. CPU ravage reaches the top end of CPU but reaches time, time, hour, hour, hour.

C. When SNORT is Implemented using existing Rule

Monitor server routine when launching SNORT, launch attack and SNORT the same solution - detect and prevent attacks. Fig. 9 use the previous experiment [14], to illustrate how SNORT can prevent DDoS attacks.

Fig. 7. Resource Monitoring before the Attack.

Fig. 8. Resources Monitoring after the Attack.

Fig. 9. Resource Monitoring when SNORT is Implemented with New Rule.

Fig. 9 shows that our proposed rules totally avoid DDoS attacks. In the above data (interval 1, interval 2, interval 3, interval 4, interval 5, and interval 6), the period interval for resource monitoring is different. As shown in the figure, when the ravage reaches the server, CPU tradition is constantly monitored at six different intervals. At interval 1, CPU usage is 3%. After a while, it will reach 15 reach and will gradually increase over the second and third intervals. To detect this ravage, SNORT is installed among the server and incoming network. When an attacker is hit by a server, resource consumption increases, but when the SNORT rule is started, it stops attacker. Resource feasting is reduced at intervals 4, 5, and 6, and CPU usage returns to its normal state.

D. Monitoring Server Performance when SNORT is Implemented with a Combined rule which is the Combination of First and Second Rule

After SNORT is executed by the blend rule, server act is constantly observed. Fig. 10 illustrates how to reduce server resource consumption from this point of view.

Fig. 10. Resource Monitoring when SNORT is Implemented with Combined Rule.

In the above image, we can look at different time intervals for monitoring resource consumption. At interval 1, the server resource is 2% on normal when the attack has not yet begun. When a flooded packet arrives at the server, the DDoS attack begins, increasing resource consumption between slots 2 and 3. After triggering the combined SNORT rule, resource consumption decreases when the condition is met and when the attack is immediately rejected. At intervals 4, 5 and 6, we can see that the server is normal.

E. Comparison of SNORT Rules

After testing SNORT Approaches, both are rules are giving better results in the detection of DDoS attack, SNORT works on rules to identify and prevent the attack. Fig. 11 shows the results and comparative analysis of SNORT rules.

Fig. 11. Comparative Analysis of SNORT Rules.

In the above Fig. 11, first rule is our proposed rule and Second rule is the combination of both previous (existing) rule and proposed rule. When SNORT is implemented using the first principle and SNORT is implemented using the second principle. The blue chart shows the server performance when SNORT is implemented using the first rule, and when the coral map shows the server performance when SNORT is implemented using the second rule. These results are discussed in Fig. 10 and 11, but on the contrary, when we look at the ranges of 4 and 5 in the two figures, the first SNORT

principle of SNORT will gradually reduce the impact of the DDoS attack. Yes, the second principle is to minimize the effect of denial of service and an immediate attack on DDoS. Both methods prevent predetermined rule-based attacks. SNORT is the best choice for the global network. For CC environments, SNORT works well as cloud users are spreading all over the world.

IV. CONCLUSION

The determination of the planned effort is to define the CC, its service model, safety problems, the address of the CC security issue, and the details of the DDoS attack and its influence on the cloud server. Computer security is a serious problem in modern clouds because it provides data storage, data exchange and other resources that are available anywhere in the world. The Internet is a public network that allows an attacker to access a cloud service without any access. CC-certified users rely on third parties for their other important security issues in third-party computing clouds. A DDoS attack is an attack-type in which it is not necessary to send a large number of packets to the server, which makes it impossible for legitimate users to access them. In this research work, a DDoS attack was launched and a tool for launching a DDoS attack was discussed. In this research, DDoS attacks were rejected using three different SNORT rules. Since SNORT uses rules after specifying rules in the SNORT base file, attacks are identified and blocked each time. In the proposed research, written rules for detecting DDoS attacks on SNORT profiles detect and prevent DDoS attacks, but because they block certain legitimate requests and generate false alarms, this should be the subject of future research.

REFERENCES

[1] Rajpurohit, Akshat, Akshat Jain, and Manish Sharma. "A Review on Cloud Computing and its Security Issues." (2018).

[2] Rakotondravony, Noëlle, Benjamin Taubmann, Waseem Mandarawi, Eva Weishäupl, Peng Xu, Bojan Kolosnjaji, Mykolai Protsenko, Hermann De Meer, and Hans P. Reiser. "Classifying malware attacks in IaaS cloud environments." Journal of Cloud Computing 6, no. 1 (2017): 26.

[3] Sridhar, S., and S. Smys. "A Survey on Cloud Security Issues and Challenges with Possible Measures." In International Conference on Inventive Research in Engineering and Technology, vol. 4. 2016.

[4] Mondal, Ranjan Kumar, and Debabrata Sarddar. "Utility Computing." International Journal of Grid and Distributed Computing 8, no. 4 (2015): 115-122.

[5] Nandi, Enakshmi, Ranjan Kumar Mondal, Payel Ray, Biswajit Biswas, Manas Kumar Sanyal, and Debabrata Sarddar. "Improved Cost-Effective Technique for Resource Allocation in Mobile Cloud Computing." In Progress in Computing, Analytics and Networking, pp. 551-558. Springer, Singapore, 2018.

[6] Leiner, Barry M., Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff. "A brief history of the Internet." ACM SIGCOMM Computer Communication Review 39, no. 5 (2009): 22-31.

[7] Leiner, Barry M., Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff. "A brief history of the Internet." ACM SIGCOMM Computer Communication Review 39, no. 5 (2009): 22-31.

[8] Jadeja, Yashpalsinh, and Kirit Modi. "Cloud computing-concepts, architecture and challenges." In 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET), pp. 877-880. IEEE, 2012.

[9] Goran Novkovic "Cloud computing's characteristics and benefits include on-demand self-service, broad network access, and being very elastic and scalable" nov 2017.

[10] Kumar, Sanjay, Syed Akbar Abbas Jafri, NishitArun Nigam, Nakshatra Gupta, Gagan Gupta, and S. K. Singh. "A New User Identity Based Authentication, Using Security and Distributed for Cloud Computing." In IOP Conference Series: Materials Science and Engineering, vol. 748, no. 1, p. 012026. IOP Publishing Ltd., 2020.

[11] Widyastuti, Dwininta, and Irwansyah Irwansyah. "Benefits and challenges of cloud computing technology adoption in small and medium enterprises (SMEs)."Bandung Creative Movement (BCM) Journal 4, no. 1 (2018).

[12] Odun-Ayo, Isaac, Olasupo Ajayi, and Sanjay Misra. "Cloud computing security: Issues and developments." (2018): 175-181.

[13] MKumar, P. Ravi, P. Herbert Raj, and P. Jelciana. "Exploring data security issues and solutions in cloud computing." Procedia Computer Science 125 (2018): 691-697.

[14] Maes, Stephane H., Rajeev Bharadhwaj, Travis S. Tripp, Kevin Lee Wilson, Petr Fiedler, and John M. Green. "Cloud application deployment." U.S. Patent 9,923,952, issued March 20, 2018.

[15] Agrawal, Gunjan. "A Survey on the "Visiosn of Cloud Compuiting–Its Referential Architecture Characteristics and Applications"." Journal Current Science 20, no. 1 (2019).

[16] Vaishnnave, M. P., K. Suganya Devi, and P. Srinivasan. "A Survey on Cloud Computing and Hybrid Cloud." International Journal of Applied Engineering Research 14, no. 2 (2019): 429-434.

[17] Khurana, Sumit, and Anmol Gaurav Verma. "Comparison of cloud computing service models: SaaS, PaaS, IaaS." International Journal of Electronics & Communication Technology IJECT 4 (2013).

[18] Liao, Yongxin, Fernando Deschamps, Eduardo de Freitas Rocha Loures, and Luiz Felipe Pierin Ramos. "Past, present and future of Industry 4.0-a systematic literature review and research agenda proposal." International journal of production research 55, no. 12 (2017): 3609-3629.

[19] Imran Ashrafa a Department of Information Technology, University of The Punjab, Gujranwala Campus, Pakistan "An Overview of Service Models of Cloud Computing" Accepted 15 Aug 2014, Available online 27 Aug 2014, Vol.2 (July/Aug 2014 issue)

[20] Siddiqui, Shams Tabrez, Shadab Alam, Zaki Ahmad Khan, and Ashok Gupta. "Cloud-Based E-Learning: Using Cloud Computing Platform for an Effective E-Learning." In Smart Innovations in Communication and Computational Sciences, pp. 335-346. Springer, Singapore, 2019.

[21] Molnar, D. and Schechter, S. "Self hosting vs. cloud hosting: Accounting for the security imp act of ho sting in the cloud", In Proceedings of the Ninth Workshop on the Economics of Information Security (WEIS), 2010.

[22] Akhawe, D., Barth, A., Lam, P. E., Mitchell, J.C. and Song, D. "Towards a formal foundation of web security", CSF, pp. 290-304,2010.

[23] Youngmin, J. and Mokdong, C. "Adaptive security management model in cloud computing environment", In the 12th International Conference on Advanced Communication Technology (ICACT), pp 1664-1669,2020.

An Innovative Approach of Verification Mechanism for both Electronic and Printed Documents

Md. Majharul Haque[1], Abu Sadat Mohammad Yasin[5],
Muhammad Shakil Pervez[6]

Bangladesh Bank, Dhaka, Bangladesh

Mohammod Akbar Kabir[3]

Department of Economics, University of Dhaka,
Dhaka, Bangladesh

Md. Nasim Adnan[2]

Department of Computer Science and Engineering,
Jashore University of Science and Technology, Jashore-
7408, Bangladesh

Mohammad Rifat Ahmmad Rashid[4]

University of Liberal Arts Bangladesh

Abstract—**Documents are inevitably relevant in our day-to-day life. Forgery of document could have severe repercussions including financial losses, misjudgments, damage of respect, goodwill, etc. Hence, documents need to be secured from threats such as counterfeiting, falsification, tempering etc., and there should be an easy way of verification about the originality of documents. There are several existing methods for ensuring authenticity and integrity with modern technologies like the block chain, Digital Signature, etc. Most of the methods are not appropriate for public usage instantly due to their intricacy, excessive costing, and implementation problem for which the easy approach of verification is yet not available for mass people. In this situation, a method of document verification has been proposed in this paper which intends to provide (i) authenticity, (ii) integrity, (iii) availability, and (iv) non-repudiation. The proposed method will serve the purpose of mass people as it has no licensing fee, easily implementable and effortlessly useable for both electronic and printed documents. It is worth to mention that the proposed method will provide a mechanism to confirm the originality of the document using only a Smartphone in no time.**

Keywords—*Document verification; integrity; non-repudiation; blockchain; printed documents*

I. Introduction

The electronic document is much more flexible to use and communicate; however, till date we cannot omit the necessity of printed document. A wide variety of legal documents that includes national identity cards, social security cards, birth certificates, driving licenses, passports, insurance documents, educational documents, wills, power of attorney, and land titles are still needed to be in the printed form for day to day life [1]. Modern technologies opened a window for mass people to exchange information in a simpler and faster way without any hassle. The extensive growth of Information and Communication Technology (ICT) has facilitated the generation, distribution, and reproduction of huge amount of electronic or printed documents effortlessly. With the help of cheap hardware and easy-to-use digital imaging software, any person can counterfeit a legal document at his/her convenience. In this way, the benefits of ICT for the electronic and printed documents come together with problems and threats associated with ensuring digital copyright protection, preventing digital counterfeiting, proof of authenticity, and content-originality verification [2].

There are a lot of organizations around the world providing several types of approval letters, certificates or other kinds of legal documents. In most of the cases, those documents are needed to be verified by various types of agency. For example, students are applying for new degree, scholarship or job to any native or foreign institute. After the application, certificates, transcripts, etc. are ubiquitously needed to be verified. Current verification processes, either automated or manual, comes with tremendous hassle, and spending of lots of money and precious time. Even, some of the verification processes are not always reliable. An important point to be mentioned that electronic documents are generated after approval of corresponding authority in a software system in a production server. Usually, a production server is kept highly secured; whereas development and test servers may not be as secured. In development or test servers, there can have provision to generate the same document which will be a simple bypass of the legal process. Printed documents can more easily be counterfeited with a scanning machine and using Photoshop or other popular software. We are aware that some doctors and teachers have been caught red handed with fake certificate after decades of their service. Hence, in this tough world, accurate verification of any electronic or printed document is extremely required as an instantaneous service for the people of all classes.

Till date, a number of techniques have been proposed for seamless document verification [1], [2]. Among them, digital watermarking techniques intend determine the ownership and integrity for digital content [3]. Watermarking techniques intend to detect noise introduced in a document while printing.

Amidror [4] presented an intrinsic security model which integrates print-based authentication and a security strategy for the protection of valuable documents. This technique is based on more intensity profiles and mathematical transformations of the microstructure of a document to generate an attractive, dynamic visual effect [3]. The author claimed that the counterfeiting of authenticity is difficult but the way of

verification from the level of mass people is almost impossible [4]. Garain & Halder (2008) proposed a framework for automatic authenticity verification. However, only a particular area of a document can be protected by using the method which can be usable for some specialized documents [5], [6]. Zhang et al. [5] proposed a method for protecting only pictures in documents by generating photo signatures based on the cosine transformation generated from the documents to validate their authenticity.

Eldefrawy et al. [7] introduced an approach to detect the differences between various document features produced using different printing technologies. Another approach was published by Mikkilineniet al. [8] to trace documents by comparing the printing characteristics of different printing devices. These methods fail to detect any forgery if a similar printer is used to reproduce the document [7], [8]. The most interesting aspect of the approach presented by Mikkilineniet al. [8] is that the security features introduced into a document to verify their authenticities are not themselves secure [3].

According to Nia et al. [9] and Gopal & Prakash [10], digital signature and blockchain can be a solution to ensure the security of electronic documents. More recently, two techniques [11], [12] have been proposed based on Quick Response (QR) codes. However, both of the techniques use digital signature to ensure the security of electronic documents. However, a soft copy of the document is needed at the user end for digital signature or blockchain verification in addition to a third-party involvement. Hence, verification through digital signature or blockchain is still at large for the mass people [13], [14].

A number of techniques on document protection can be found in literature based on digital watermarking, steganography, digital signature and blockchain, etc. It is important to mention that most verification procedures of those techniques are tedious and time-consuming as they often involve many legal entities across the globe [15]. Hence, it is evident that very little attention has been given on how to validate the authenticity and integrity of confidential documents effortlessly for the mass people.

Under these circumstances, in this paper we propose a simple yet sophisticated method to provide a verification mechanism for both electronic and printed documents. It is expected that the proposed method will be easily accessible for the mass people and will eliminate many hindrances of the document verification process.

The rest of the paper is organized as follows: Section II illustrates the proposed method in detail. Security potentials and benefits of the proposed method are discussed in Sections III and IV respectively. Finally, we present our concluding remarks in Section V.

II. METHODOLOGY

The proposed method intends to cover the authenticity, integrity, non-repudiation, and availability of a digital and printed document. The architecture is built as a client-server model where a client connects to the server and request for document generation. The entire method is described in the following five steps.

A. Requesting for Document Generation

Present your perspective on the issues, controversies, problems, etc. as they relate to theme and arguments supporting your position. Compare and contrast with what has been or is currently being done as it relates to the article's specific topic and the main theme of the journal.

1) There is a central server that has some necessary templates (with unique template' id) based on which document will be created. In this server, a client can be registered by an authorized agreement for a definite period through the Internet. After registration, the client will get a unique Id and password.

2) In the client-side database, there is a request queue table with the following fields:

a) Request Id (Auto-incremental unique Id).

b) Template Id (for the template to be selected in the server).

c) Application Id. (The specific application for which the client needs the approval letter).

d) Application Type Id. (There can be multiple types of applications).

e) JSON Data (Necessary information and template of the application that will be placed in the server).

f) Response from the server.

g) No_of_Try (The default value of this field is 0, which means it is yet to try for generating the document. The value is increased by 1 based on the number of trying. The value is set to -1 after 3 times trying and then no more trying).

h) Status (The value can be set as "success", "fail" or "rejected").

i) Response from the server (The entire response from the server is saved in JSON format)

3) The registered client requests for connection (step 1 of Fig. 1) to the server via an API (Application Interface) request with the Client Id and password. If the credential is found valid, the server responses with a unique token number (step 2B of Fig. 1). Then, a scheduler in client fetch one by one request from the queue where the "No_of_Try" value is not equal to "-1" and the "status" is not equal to "success" (success means the document is already generated successfully) and send to the server with a token number (step 3 of Fig. 1).
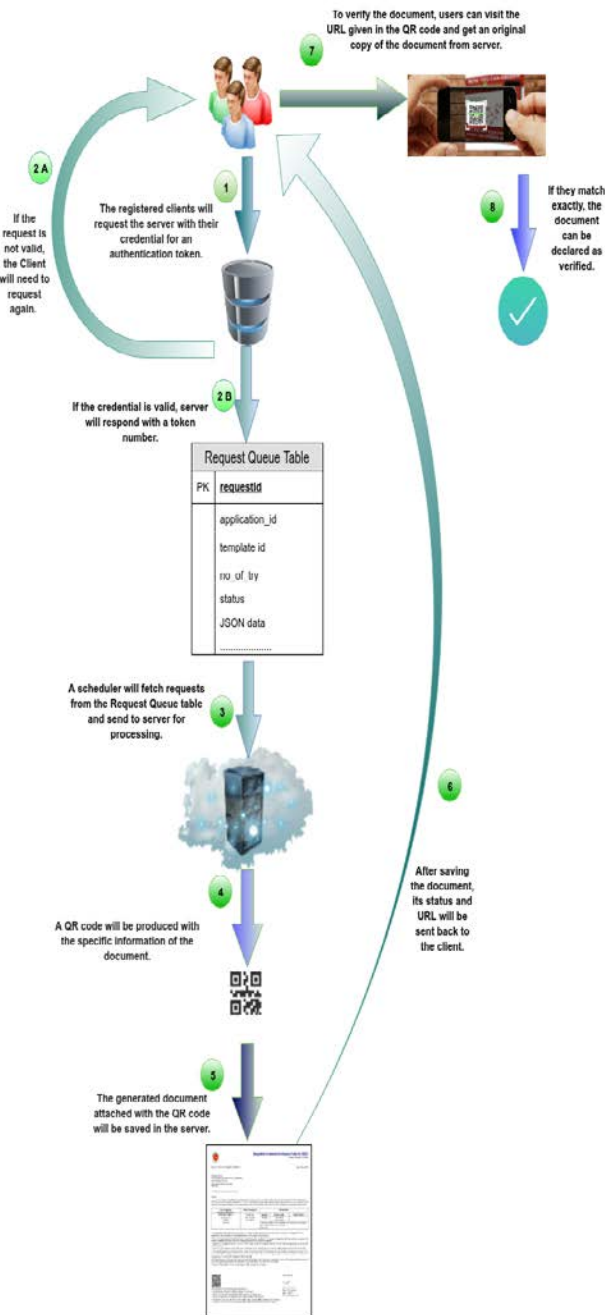
Fig. 1. Process Flow of the Proposed Method.

B. Creating Document

Present your perspective on the issues, controversies, problems, etc. as they relate to theme and arguments supporting your position. Compare and contrast with what has been or is currently being done as it relates to the article's specific topic and the main theme of the journal.

1) If the request is found valid from the Client Side, the corresponding template is selected in the server according to the Template Id. Then, the JSON formatted information (received by the server as an API request) is parsed and put in different predefined places of the selected template. Generally,

there can be three types of server for any Document Management system as follows: (i) Development, (ii) User Acceptance Test (UAT), and (iii) Live. In the Development and UAT server, the generated document will have a watermark as "Development" and "UAT", respectively. In the Live server, the generated document will have no such kind of watermark.

2) The generated document is saved in the central server with a unique file name where the file name is set using a defined naming convention as lientID_TemplateID_DD_MM_YYYY_N.pdf. Here DD means two (02) digits of a day, MM means two (02) digits of a month and YYYY means four (04) digits of a year, N means an incremental number (Starting from 1 for each day), ".pdf" means the document is generated in Portable Document Format.

3) In the pdf file (generated document), a QR Code is attached which contains the full URL of the document. This URL is the same as the path of the file where this file will be saved in the server (step 4 and 5 of Fig. 1). If anybody hits the aforementioned URL, he/she will get the corresponding pdf file.

4) After generating the document, it is saved in a specific folder as APPROVAL_LETTER/CLIENT_ID/YYYY/MM where the APPROVAL_LETTER is a folder which contains several sub-folders named CLIENT_ID, then Year and then Month.

5) If we want to archive some files, we can easily decide that the files of which Client, Year and Month needs to be archived.

C. Providing Response from the Server:

After generating the document and saving in the central server, a response is given to the corresponding client with the following information in JSON format (step 6 of Fig. 1): (i) status (it can be either "success" or "fail". The status can also be rejected, if there is any mismatch in the token number), (ii) the URL of the document (it is given if the status is success otherwise it is left blank), (iii) the duration between a request received and the response provided.

D. Receiving Response by the Client from the Server:

1) The client parses the JSON formatted response from the server and saves the value of "status" field, URL of the document and the entire response in request queue table (mentioned in the second point of sub-section A) in the database (client-end).

2) Client updates the value of "No_of_Try" field in request queue table as "-1" if the status is "fail" or "rejected" after trying 3 times so that will not be tried again.

E. Verifying Document:

After generating the document and saving in the central server, a response is given to the corresponding client with the following information in JSON format (step 6 of Fig. 1): (i) status (it can be either "success" or "fail". The status can also be rejected, if there is any mismatch in the token number),

(ii) the URL of the document (it is given if the status is success otherwise it is left blank), (iii) the duration between a request received and the response provided.

1) For the verification purpose, one needs to collect the document in either printed copy or soft copy (pdf). If it is printed copy, it is needed to confirm that the QR code has been printed well.

2) Almost all modern people have a smartphone in this age. By using those devices, the QR code can be scanned from the document where a URL will be found (step 7 of Fig. 1). If we hit this URL, a document will be downloaded from the server if the URL is valid otherwise it will give "not found" error. Now, the verification is very simple as follows:

a) if there is any document downloaded by visiting the URL, we need to match the downloaded document and the document we have in hand. If both are 100% same, we can declare that the document is valid (step 8 of Fig. 1)

b) if there is any mismatch or there is "not found" error, the document can be regarded as a fake one.

III. Ensuring Security

Though we are emphasizing on document verification, some security features have also been covered with the proposed method. Four significant points have been covered here as (i) authenticity, (ii) integrity, (iii) non-repudiation, and (iv) availability. Authenticity is the assurance that a message, transaction, or other exchange of information is from the source it claims to be from [18]. Integrity refers to the document has not been distorted by anyone. Non-repudiation confirms that the authority cannot disown the document. Last of all, availability means the document can be retrievable at any time. By using the proposed method, the user can claim that the approval letter has been given from a specific authority/organization. Now, if the same document has been found in the given URL in the QR code (explained in the third point of sub-section B under section III) at any time, it ensures the four points for security issue. (i) Authenticity is confirmed in such a way that the source of document is verified by downloading the document from the URL of specific organization, (ii) If the document has been distorted by anybody then the newly downloaded document will not be same as the old document and hence the loss of integrity can be detected, (iii) The authority cannot disown the document because this can be downloaded from the URL of the authority's system that is ensuring non-repudiation, (iv) Since the document is obtainable at any time from the URL, availability is also ensured.

IV. Comparison with Latest Existing Methods

In Table I, we present a brief comparison between the proposed method and some existing methods.

TABLE I. Comparison between the Proposed Method and Some Existing Methods

Criteria	Proposed Method	AL-Gawda et al. [3]	Warasart & Kuacharoen [15]
Security	Very High	Very High	Very High
Counterfeit Detection Accuracy	Very High	Very High	Very High
Applicability	Both Digital and Printed Documents	Both Digital and Printed Documents	Both Digital and Printed Documents
Use of Digital Signature	No	Yes	Yes
External Dependency	Service Provider Server (expected to be online for 24×7	Digital Signature Certifying Authority	Digital Signature Certifying Authority
Licensing Fees	No	Yes	Yes
Cost	No Cost (No Overhead for the Mass People)	High Cost (Maintenance of Digital Signature for all Concerned Actors)	High Cost (Maintenance of Digital Signature for all Concerned Actors)
Verification Process	Easy (Low Computation Overhead)	Complicated (High Computation Overhead)	Complicated (High Computation Overhead)
Applicability for Mass People	Yes	No	No

V. Benefit of the Proposed Method

We believe that the proposed system has the following benefits by comparing with other systems like a digital signature or any central authority for document verification:

1) The proposed system can be developed by any organization where license from any other third party is not required.

2) User can easily get another copy of the document if a downloaded copy is lost or distorted.

3) If a user wants to give a copy of the document to another person anywhere, he/she just has to give the URL of the document.

4) Since the central server is API-based, it can be connected to multiple clients without any hassle.

5) The process of verification has almost no cost, with no process-oriented harassment and requires a very short time assist only requires visiting a URL.

VI. Conclusion

It is well-known that counterfeiting of any necessary official or unofficial document is a criminal activity which adversely affects the country, society, and even personal reputation. For eradicating this problem, a method of document creation and verification has been proposed here to ensure authenticity, integrity, non-repudiation, and availability. The entire mechanism of providing security and performing verification has been discussed here in detail including the benefits of the proposed method compared to others. The key strengths of this method are to design an efficient, secure, low-cost, and easily accessible solution for the people of almost all levels without compromising quality. We believe that this method will prevent the falsification of documents. In the arena of software development, it will also be tremendously useful to issue any certificate or approval letter. Detection of falsified documents can also be augmented using several data mining algorithms [16], [17], [18], [19].

References

[1] Van Renesse, R. L. (1997). Paper based document security-a review.In European Conf. on Security and Detection-ECOS97.

[2] Tayan, O., Kabir, M. N., & Alginahi, Y. M. (2014). A Hybrid Digital-Signature and Zero-Watermarking Approach for Authentication and Protection of Sensitive Electronic Documents. The Scientific World Journal, 2014, 1-14.

[3] AL-Gawda, M., Beiji, Z. & Nurudeen, M. (2015). Printed Document Authentication Using Two-Dimensional (2D) Barcodes and Image Processing Techniques.International Journal of Security and Its Applications, 9(8), 347-366.

[4] Amidror, I. (2002). New print-based security strategy for the protection of valuable documents and products using moiré intensity profiles. Electronic Imaging 2002 International Society for Optics and Photonics, 3677, 89-100.

[5] Garain, U.,&Halder, B. (2008). On automatic authenticity verification of printed security documents., 2008. ICVGIP'08. In Proceedings of the 2008Sixth Indian Conference on Computer Vision, Graphics & Image Processing(pp. 706-713).Washington DC, United States: IEEE Computer Society.

[6] Li, J., Zhang,X.,Liu,S.,& Ren,X. (2009). Adaptive watermarking scheme using a gray-level computer generated hologram.Applied optics, 48(26),4858-4865. generated hologram.Applied optics, 48(26),4858-4865.

[7] Eldefrawy, M. H., Alghathbar,K.& Khan,M. K. (2012). Hardcopy Document Authentication Based on PublicKey Encryption and 2D Barcodes.In Proceedings of the 2012 International Symposium on Biometrics and Security Technologies (pp. 77-81). Washington DC, United States: IEEE Computer Society.

[8] Mikkilineni, A. K., Chiang, P.J.,Ali, G. N.,Chiu,G. T.,Allebach, J. P.,& Delp, E. J., (2005). Printer identification based on graylevel co-occurrence features for security and forensic applications. In Proceedeings ofthe conference on Security, Steganography, and Watermarking of Multimedia Contents VII (Vol.5681, pp. 430-440). California, USA: International Society for Optical Engineering.

[9] Nia, M.A., Sajedi, A., & Jamshidpey, A. (2011). An Introduction to Digital Signature Schemes.In Proceedeings of National Cpnference on Information Retrieval. Retrieved from https://arxiv.org/abs/1404.2820.

[10] Gopal, N., & Prakash, V.V. (2018). Survey on Blockchain Based Digital Certificate System. International Research Journal of Engineering and Technology,5(11), 1244-1248.

[11] Schultz, Michelle Kelly. "A case study on the appropriateness of using quick response (QR) codes in libraries and museums." Library & Information Science Research 35, no. 3 (2013): 207-215.

[12] André, P.S. and Ferreira, R.A.S., 2014. Colour multiplexing of quick-response (QR) codes. Electronics Letters, 50(24), pp.1828-1830.

[13] Avram, M. G.: Advantages and Challenges of Adopting Cloud Computing from Enterprise Perspective. Procedia Technology, Vol. 12(1), pp. 529–534 (2014).

[14] Li, J., Zhang,X.,Liu,S.,& Ren,X. (2009). Adaptive watermarking scheme using a gray-level computer generated hologram.Applied optics, 48(26),4858-4865. generated hologram.Applied optics, 48(26),4858-4865.

[15] Warasart, M., Kuacharoen, P. (2012). Paper-Based Document Authentication Using Digital Signature and QR Code.In Proceedings of the 4th International Conference on Computer Engineering and Technology(pp. 94–98). Phuket, Thiland: ASME Press.

[16] Adnan M. N.: On Dynamic Selection of Subspace for Random Forest. The 10th International Conference on Advanced Data Mining and Applications (ADMA 2014), pp. 19 – 21 (2014).

[17] Adnan M. N., Islam M. Z.: ComboSplit: Combining Various Splitting Criteria for Building a Single Decision Tree. The International Conference on Artificial Intelligence and Pattern Recognition (AIPR), pp. 17 – 19 (2014).

[18] Adnan, M. N., Islam, M. Z.: ForEx++: A new framework for knowledge discovery from decision forests. Australasian Journal of Information Systems, Vol. 21, pp. 1–20 (2017).

[19] Adnan, M. N., Islam, M. Z.: Forest PA: Constructing a decision forest by penalizing attributes used in previous trees. Expert Systems with Applications, Vol. 89, pp. 389–403 (2017).

Feature Expansion using Lexical Ontology for Opinion Type Detection in Tourism Reviews Domain

Lim Jie Chen[1], Gan Keng Hoon[2]

School of Computer Sciences, Universiti Sains Malaysia

Penang, Malaysia

Abstract—Tourism reviews platform such as Trip Advisor become a major source for tourists to share their experiences and get some ideas for decision making. Since there are millions of reviews generated daily in the travel websites, tourist is often overwhelmed with huge information. This is where opinion type detection is important as it makes it easy for a tourist to obtain useful reviews for their understanding and planning processes based on the reviews' opinion type. The opinion type of texts in travel mostly involves different aspects of opinion related to the travel process, such as transportation, accommodation, price, food, entertainment, and so on. The challenge of this research is to improve this detection by proposing the lexical ontology approach to address the issue of out-of-vocabulary (OOV) keywords during a supervised detection of opinion type. Besides, there are also issues where the training data for detection has poor coverage or limited in a certain domain. In this paper, we propose a review opinion type detection approach by integrating the word (feature) expansion approach in machine learning. The suggested approach consists of two stages namely feature expansion and classification. For feature expansion, Lexical Ontology (LO) is used to expand the feature-related word to the domains such as synonyms. For classification, the expanded feature is corporate to the Machine Learning approach to detect the opinion type.

Keywords—*Tourism domain; online review; opinion type detection; text classification; lexical ontology*

I. INTRODUCTION

Nowadays, tourists often rely on the online review when planning for their vacations such as Trip Advisor. TripAdvisor is the largest social travel website with about 500 million reviews of hotels, restaurants, attractions, and other travel-related businesses. Customer reviews provide reliable and valuable opinions about a tourist attraction such as its services, destination, and recommendations which helps tourists to understand more about a tourist attraction. However, due to the increase in the number of reviews recently, during decision making, it becomes difficult for the tourist to read all the reviews. Tourists are often overwhelmed and face difficulty in filtering relevant information from large number of reviews. Hence, it would be helpful if the opinion can be provided based on a certain type which is useful for decision making. For example, the tourism domain has Attractions, Concerts and Shows, Food & Drink, Transportation etc., job seekers domain contains Culture & Values, Work/Life Balance, Senior Management, Compensation and Benefits, and Career Opportunities. This is very important as online users trust customer reviews 12 times more than the product details provided by businesses [1]. Hence, automated identification of

reviews is important to help people to identify the opinion type of online reviews. This is the main reason of opinion type detection important. By applying opinion type detection, social network websites can structure user reviews in a fast and cost-effective way.

With this motivation, this paper focuses on opinion type detection where the problem of detection is formulated as the problem of text classification [2]. In the area of opinion analysis, classification is commonly used for topic classification or sentiment classification. In this research, our focus is on the topics or types in the context of tourism domain. For example, transportation, accommodation, food, entertainment, price and so on. Although the scope of contents of reviews is refined to tourism domain, nevertheless it is common to have various topics in the discussions. Hence, text classification in the tourism domain is an essential task to identify the topics mentioned in the text for further stage of analysis or application [3]. The outcome of this research is considered important in the decision making for both customers and the operators of tourism domain.

II. RESEARCH PROBLEM

There are some issues in opinion type detection/classification that lead to the needs of feature expansion. It is common to encounter contents variations and word variations when a classification model is applied in a new target data, even from the same domain. A classifier trained to detect the opinion type, e.g. food, from one point of interest may not guarantee similar performance when applied in data from another point of interest. Two main reasons behind this are out-of-vocabulary (OOV) keywords and limited labeled data when training is performed.

A. Out-Of-Vocabulary (OOV) Keywords

The first problem is caused by out-of-vocabulary (OOV) keywords during the detection involving new or unseen reviews. Once matching keywords are not captured in the model trained, this will create issue in the correct opinion type detection. For example, R1: the soup is very hot. Assumed that the dataset that we trained does not contain any keywords in R1, hence it is hard for a machine to determine its category. However, assume that there is keyword "spicy" in our training source. By using WordNet, the keyword "hot" in R1 shows the same meaning with the keyword "spicy" and they are in the same category. Expanding "hot" to "spicy" will improve the chances for machine to assign R1 to the correct opinion type. Fig. 1 shows the word "hot" has the same meaning as the word "spicy".
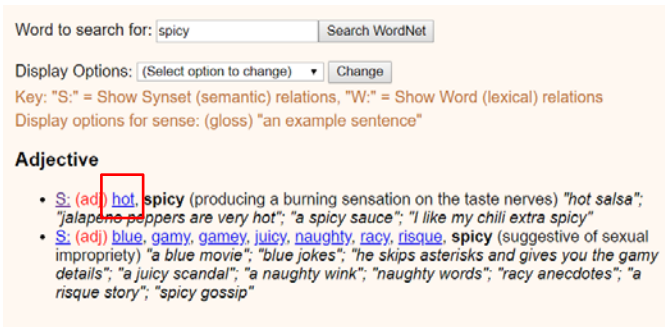
Fig. 1.   A Snapshot of the Word "hot" has the same meaning as the Word "Spicy" in Princeton's WordNet Web.

B. Limited Labeled Data that Represents the Concept of an Opinion Type

For example, R2: soup is very spicy and R3: soup is very tasty. Since R2 and R3 belong to the same concept of taste, they should be able to be detected as the same type. However, the machine might not determine they are in the same category. Assumed R3 is the upcoming reviews and the word "spicy" is in our training, by using WordNet, the keyword "spicy" in R2 and "tasty" in R3 shows they are similar to each other and they are in the same category. Thus, this improve the chances for machine to assign R2 and R3 to the correct opinion type. Fig. 2 shows the word "tasty" is similar to the word "spicy".

Fig. 2.   A Snapshot of the Word "Tasty" is Similar to the Word "Spicy" in Princeton's WordNet Web.

III. LITERATURE REVIEW

This section covers literatures related to different approaches of opinion type detection, using bag of words (BOW) approach. Works related to feature expansions are also discussed.

In a BOW, the words in a matrix do not represent sentences with structure and grammar, and the semantic relationship between these words are ignored in the construction of BOW representation. Another limitation of BOW is on its semantic meaning, basic BOW approach does not consider the meaning of the word in the document. It ignores the context in which it is used. The same word can be used in multiple places based on the context of nearby words [4].

According to E. Rudkowsky *et al.* [5], in the domain of social science, the use of word embeddings introduces a new approach to the field of sentiment analysis in the social sciences that offers potential to improve on current bag-of-words approaches. The main advantage of using word embeddings is its potential to detect and classify unseen or out-of-context words that are not included in the training data. Vector representations of text that allocate similar words closer to each other, such approach can supplement training data, which is promising in improving the results of machine learning tasks.

According to Sneha [6], the very first step of sentiment classification is to extract the phrases containing adverbs and adjectives in the review because they are good indicators of subjectivity. However, single-word adjectives and adverbs may have different meanings in different contexts where they modify the meaning of other words quickly. It is not sufficient to reply on single adjectives and adverbs as potential opinion word, noun and verb may represent aspects or its attributes in the review. Therefore, rather than selecting single word adjective or adverb, bigrams which contain noun phases with adjective and adverb are better choice.

From an e-commerce perspective, M. Hu *et al.* [7] and K. Vivekanandan *et al.* [8] have proposed a frequency-based method for aspect extraction. In this approach most frequent words in reviews usually, nouns and pronouns are considered to be candidate of aspects. However, S. Abeysinghe *et al.* [9], then improves the method by applying part-of-speech patterns to filter the terms added to the frequency terms as well. Another approach is using syntactic relations in words to determine the aspects.

For feature expansion techniques, several investigations have attempted to improve the out of vocabulary keywords problem. S. M. Rezaeinia et al. [10] conducts research in word embedding method, and they found that their Improved Word Vectors (IWV) which is based on the combination of natural language processing techniques, lexicon-based approaches and Word2Vec/GloVe methods which increased the accuracy of pre-trained vectors in sentiment analysis. They proposed a method that gets a sentence and returns improved word vectors of the sentence. They used Word2Vec which is based on continuous   Bag-of-Words   (CBOW)   and   Skip-gram architectures which can provide high quality word embedding vectors.

In the tourism domain, Muhammad Afzaal *et al.* [11] presented an aspect-based sentiment classification framework using tree-based aspects extraction method that classifies opinions/reviews of aspects into positive or negative. The opinion-less and irrelevant sentences are first removed by employing Stanford Basic Dependency on each sentence and the features are extracted from the remaining sentences with N-Grams and POS Tags to train the classifiers. Therefore, the limitation is that some opinion-less/irrelevant texts might be an important source/text for opinion type in the classification process. Removing them may result in OOV issue.

K. Soo-Min *et al.* [12] develops an automatic algorithm to produce opinion-bearing words by hybridizing two methods. First method is a small set of human-annotated data that shows that productive synonyms and antonyms of an opinion-bearing word can be found through automatic expansion in WordNet and use them as feature sets of a classifier. They also use all

synonyms of a given word as well as the word itself to determine a word's closeness to opinion-bearing or non-opinion-bearing synonym set.

TABLE I.   SUMMARY ON GAP ANALYSIS ON OPINION TYPE DETECTION

	Paper	Method	Problem Domain	Dataset
Bag-of-Word	[6]	Words representation of text content, ignoring grammar and order of appearance are ignored. The limitation of **Bags of Words (BOW)** is on its semantic & contextual meaning.	Automatic topic identification of health-related messages in online health community	Health-related messages (4041) in three message boards: treatment board, emotional support board and survivorship board.
Bag-of-Word	[7]	Word embedding method outperform **Bags of Words (BOW)** improving detection and classification of unseen or out-of-context words that are not included in the training data.	Sentiment analysis in social sciences	20580 sentences from party press releases, transcripts of parliamentary speeches, and media reports.
Natural Language Processing	[8]	**Single-word adjectives and adverbs** may have different meanings. Rather than selecting single word adjective or adverb, **bigrams** which contain **noun phases with adjective** and **adverb** are used.	Sentiment classification	Multi-Domain Sentiment Dataset (667 reviews) - Book, DVDs, Kitchen Appliances, and Electronics Appliances
Natural Language Processing	[9], [10]	**Frequency-based method** for aspect extraction. Most frequent words in reviews comprising of nouns and pronouns are considered to be aspects.	Mining opinion features in customer reviews	Customer reviews of five electronics products - from Amazon.com and C|net.com.
Lexical Ontology	[14]	Automatic **expansion using WordNet** and use it as feature sets of a classifier. Used all **synonyms** of a given word.	Sentence-level opinion detection system	Collections of opinion-bearing (2683) and non-opinion-bearing words (2548) manually from Columbia University
Lexical Ontology	[13]	Extracts **frequent nouns and noun phrases** from reviews text. Groups similar nouns using **WordNet**. **Decision tree** is employed on reviews where review words are used as internal nodes and extracted nouns as the leaf of a tree.	Aspect-based sentiment classification for tourist reviews	Restaurant (2000 reviews) and hotel (4000 reviews) domains dataset-collected from popular social media websites using crawler and APIs.

To wrap up the literatures for this research, we have summarized some of the relevant works in three aspects, i.e. Bag-of-Word, Natural Language Processing and Lexical Ontology in Table I.

IV. RESEARCH METHODOLOGY

In this section, we present our proposed work on the Opinion Type Detection framework for training and detecting review sentence's opinion type as shown in Fig. 1. In this framework, there are three main steps which are text pre-processing, feature expansion, and classification. The input to this framework is reviews sentence with its corresponding opinion type and the output is the accuracy of the model.

Fig. 3 shows the pipeline for opinion type detection using a supervised learning algorithm. In text pre-processing task, the datasets are collected, categorized, cleaned, and sorted based on some filtering task. In feature expansion, the expansion can be applied on features obtained from methods such as bags-of-words (BOW) and Natural Language Processing (NLP). In this study, we propose a Lexical Ontology (LO) approach to improve the opinion type detection in the tourism review. After feature expansion, machine learning approach such as Naïve Bayes (NB) classifier, Support Vector Machine (SVM), and Decision Tree (DT) is applied.

A. Tourism Review

Tourism review is a review made by a consumer who has experienced gain from travelling (see Fig. 4). Customer reviews provide true and valuable opinions about a tourist attraction which helps tourist to understand more about tourist attraction when making a decision. In tourism domain, there are some important opinion types in which a review can be categorized such as "Attraction", "Fee", "Time", "Weather", "Transport", "Service" and "Food".

Fig. 3. The Design Framework of Opinion type Detection.

Fig. 4.   Example of a Tourism review at Trip Advisor Website.

B. Text Pre-Processing

Text pre-processing is the process the cleans and prepare the text prior to classification process. Since real world data often contain noise and formatting errors, in pre-processing step, these unnecessary data will be removed to improve the quality of the input data. In our case, input that will be pre-processed are sentences of review text. Each sentence will go through the following steps:

Removing punctuation and convert text to lowercase: Each review sentence is converted to lowercase and has its punctuation removed.

Example: "Genting Theme Park is a full value for young once."

After this step, the sentence will become "genting theme park is a full value for young once".

Tokenization: A review sentence is treated as a string and split into a list of tokens.

Example: "genting theme park is a full value for young once"

After tokenization step, the sentence is divided into tokens, "genting", "theme", "park", "is", "a", "full", "value", "for", "young", "once".

Removing stop words: Stop words such as "the", "a", "and" etc. occur frequently, but do have significant role in the semantic/context of the text. Removing stopwords can potentially help improve the performance as there are fewer and only meaningful words retained. Thus, it could increase classification accuracy.

Example: "genting theme park is a full value for young once"

After removing stop words the tokens will be "genting", "theme", "park", "full", "value", "young".

Lemmatization: Lemmatization reduces a token in inflected form to the root form, called Lemma. A lemma is the canonical or dictionary form of a word.

Example: "runs", "running", "ran" will be transformed to its root form "run".

C. Feature Expansion

Bag-of-words (BoW)

Bag-of-words method is a simple representation of features (e.g. in word token form) obtained from the text documents. The model consists of bag, i.e. multiset, of words, where grammar rules are disregards. Word counts are represented in this model [13]. This method is often used in

1.   Natural Language Analysis.

2.   Document Classification.

3.   Information Retrieval.

An illustration of how word vectors are generated in bag-of-words model is shown below. Given two sentences,

1.   Dad likes to watch movies. Mum likes movies too.

2.   Dad also likes to watch indoor games.

These two sentences can be represented as follows as a collection of words.

1. ["Dad", "likes", "to", "watch", "movies", "Mum", "likes", "movies", 'too']

2. ["Dad", "also", "likes", "to", "watch", "indoor", "games"]

Remove the duplicate words and use the word count for the representation.

1.{"Dad":1,"likes":2,"to":1,"watch":1,"movies":2,"Mum":1,"too":1}

2.{"Dad":1,"also":1,"likes":1,"to":1,"watch":1,"indoor":1,"games":1}

Combine both sentences with their word frequency.

{"Dad":2,"likes":3,"to":2,"watch":2,"movies":2,"Mum":1,"too":1,"also":1, "indoor":1,"games":1}

Hence, we obtain a vocabulary for this small collection and by using this vocabulary we can create vectors for our sentences.

The length of the vector must be equal to the vocabulary size. Here the length of the vector is ten. Then, by comparing the sentences with the vocabulary and we get the vectors as follows.

Dad likes to watch movies. Mum likes movies too.

[1, 2, 1, 1, 2, 1, 1, 0, 0, 0]

Dad also likes to watch indoor games.

[1, 1, 1, 1, 0, 0, 0, 1, 1, 1]

The size of vector is proportionate to the size of vocabulary. Hence, for document with long texts, the size of the vocabulary is high. This also cause the vector to contain a greater number of zeros. It is called sparse matrix and the sparse matrix require more memory and high computational power [13].

Natural Language Processing (NLP)

Useful features can be identified using natural language analysis process like Part-of-Speech (POS) tagging. By tagging each word in terms of its POS, such as noun, pronoun, adverb, adjective, verb, etc. the syntactical meaning of the word can be used as reference to select relevant features. Table II shows the POS tags and their related meanings while Table III shows the example of the generation of lexical elements for one of the review sentences.

Lexical Ontology (LO)

To address the OOV words and limited labelled data issues, we propose to include Lexical Ontology in the opinion type detection task by expanding the features for each review sentences. Given a review, features are extracted from the reviews. Then, the features/keywords will be expanded by synonyms using WordNet, a well-known Lexical Ontology. We make assumption that these additional features can improve the accuracy of opinion type detection. Based on the two basic methods of features identification (BoW and NLP), we perform the expansion on four variants of feature sets: feature set F_{BOW}, F_{NLP}, F_{BOW+LO} and F_{NLP+LO}. Feature set F_{BOW} and F_{NLP} are used as the baseline feature set to assess the performance of the proposed expanded feature sets (F_{BOW+LO} and F_{NLP+LO}).

TABLE II. POS TAGS AND MEANING

POS TAGGING	MEANING
JJ	ADJECTIVE
RB, RBR, or RBS	ADVERB
NN or NNS	NOUN
VB, VBD, VBN, VBZ, VBG	VERB

TABLE III. THE GENERATION OF LEXICAL ELEMENTS FOR A REVIEW SENTENCE

STEPS	EXAMPLE
A review sentence	Genting Theme Park is a full value for young once.
POS Tagging and Stemming	Genting/NNP Theme/NNP Park/NNP is/VBZ a/DT full/JJ value/NN for/IN young/JJ once/RB ./.
Nouns	Genting, Theme, Park, value

Fig. 5. General Process of Feature Expansion using Lexical Ontology.

Base Features Extraction

Fig. 5 shows the process of feature expansion. A review sentence will first go through the pre-processing step, followed by Base Feature Extraction. The outcome for BOW method will be the feature of each review, $F_{BOW} = \{f_{b1}, f_{b2}, …, f_{bn}\}$.

After extraction, a list of features for each review $F_{BOW} = \{f_{b1}, f_{b2}, …, f_{bn}\}$ will be stored for feature expansion.

Similar feature set will be extracted using the NLP method, resulting in $F_{NLP} = \{f_{b1}, f_{b2}, …, f_{bn}\}$.

Feature Expansion

```
Input: Features for each review, F_W = {f_w1, f_w2, …, f_wn} with its
opinion type, where F_W can be F_BOW or F_NLP
Output: F_W+LO ← f_w1, f_w2, …, f_wn, f_w1+LO(1), f_w1+LO(2), …, f_w1+LO(n),
              f_w2+LO(1), f_w2+LO(2), …, f_w2+LO(n), f_wn+LO(1),
              f_wn+LO(2), …, f_wn+LO(n)
for all F_W do
        F_LO ← expand F_BOW with WordNet (Synonyms)
        F_LO ← f_w1+LO(1), f_w1+LO(2), …, f_w1+LO(n), f_w2+LO(1), f_w2+LO(2), …,
              f_w2+LO(n), f_wn+LO(1), fw_n+LO(2), …, f_wn+LO(n) }
        F_W+LO = F_W + F_LO
        F_W+LO ← f_w1, f_w2, …, f_wn, f_w1+LO(1), f_w1+LO(2), …, f_w1+LO(n),
              f_w2+LO(1), f_w2+LO(2), …, f_w2+LO(n), f_wn+LO(1),
              f_wn+LO(2), …, f_wn+LO(n) }
        F_W+LO ← Remove Duplicate feature
end for
return F_W+LO
```

Fig. 6. Pseudocode for Process of BOW+LO/NLP+LO Approach.

Fig. 6 shows the pseudocode for applying LO approach to BOW method and NLP method. By expanding the feature in F_W, the output will be stored in F_{LO} where F_{LO} is the feature obtained by expanding F_W with WordNet (Synonyms). Then, both F_W and F_{LO} are combined, F_{W+LO}. Using a sample review "*There is an area for arcade games, again, maybe more suitable for children.*" from "Genting Highlands Theme Park" POI, the features/keywords selected for four variant approaches is shown in Table IV. The steps of how the features are expanded are described in the subsequent paragraphs.

For F_{BOW+LO}, the review sentences go through pre-processing, which includes remove punctuation, stemming, tokenization, and remove stop words. This resulted the base features from the review, i.e. ['area', 'arcade', 'game', 'maybe', 'suitable', 'child'] as F_{BOW}.

TABLE IV. FEATURES/KEYWORD SELECTED FROM ONE OF THE SAMPLE REVIEWS

Symbol	Feature/Keywords	Approach
F_{BOW}	['area', 'arcade', 'game', 'maybe', 'suitable', 'child']	BOW
F_{NLP}	['area', 'game', 'child', 'maybe', 'suitable']	NLP
F_{BOW+LO}	['area', 'arcade', 'game', 'maybe', 'suitable', 'child', 'unfit', 'brave']	BOW+LO
F_{NLP+LO}	['area', 'game', 'child', 'unfit', 'brave', 'suitable', 'maybe']	NLP+LO

Similarly, for F_{NLP+LO}, all review sentences will be pre-processed as well, followed by POS tagging. Feature which is noun and other main tag types are selected as F_{NLP+LO} features, i.e. ['area', 'game', 'child', 'maybe', 'suitable'].

Then, we expand the base features with synonyms from LO, that results in F_{BOW+LO} ['area', 'arcade', 'game', 'maybe', 'suitable', 'child', 'unfit', 'brave'] and F_{NLP+LO} approach ['area', 'game', 'child', 'unfit', 'brave', 'suitable', 'maybe'].

Fig. 7 shows the process of expanding the keywords from the baseline methods (NLP and BOW). First, the features/keywords of BOW method reviews (training source) are generated, $(f_{b1}, f_{b2}, …, f_{bn})$. Then the expansion goes through each feature, e.g. f_{b1} means that the first feature is selected from T_1 training review (and the rest can be done in the same manner). Then, the expanded features are obtained using LO's synonym, $(f_{b1+LO(1)}, f_{b1+LO(2)}, …, f_{b1+LO(n)}, f_{b2+LO(1)}, f_{b2+LO(2)}, …, f_{b2+LO(n)}, f_{bn+LO(1)}, f_{bn+LO(2)}, …, f_{bn+LO(n)})$. $f_{b1+LO(1)}$ refers to the new feature that is expanded from the first feature. Finally, the BOW+LO features set can be generated by combining the features from the baseline and new features, which is called "BOW+LO approach". The same process applies to NLP method.

D. Classification

An important step in the opinion type detection pipeline is choosing a good classifier. This can be done by adopting different type of classifiers and measure their performances to serve as a guideline in the selection. Supervised machine learning models such as Naïve Bayes (NB) classifier, Support Vector Machine (SVM), and Decision Tree (DT) are chosen due to their popularities. Review data set for each opinion type is also split into training and test datasets to train and test the model.

1) Naïve Bayes (NB): Naïve Bayes classifier has been widely used for document categorization tasks [14]. It is theoretically based on Bayes theorem, which was developed by Thomas Bayes [15]. Recent studies show that NB is commonly used in information retrieval [16]. Naïve Bayes classifier is a generative model, which is a traditional method of text categorization. This classifier is chosen as the since it is the common base for classification task.

If the number of documents (n) fits into k categories where $k \in \{c_1, c_2, ⋯, c_k\}$, the predicted class as output is $c \in C$. The Naïve Bayes algorithm can be described as follows [17], [18]:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

$$P(d|X) = P(x_1|d) \times P(x_2|d) \times … \times P(x_n|d) \times P(d)$$

where d is document and c indicate classes.

where:

P($c | d$) = posterior probability. The probability of hypothesis h being true, given the data x, where P(c | d)= P(d_1| c) P(d_2| c)….P(d_n| c) P(x)

P($d | c$) = Likelihood. The probability of data d given that the hypothesis h was true.

P(c) = Class prior probability. The probability of hypothesis h is true

P(d) = Predictor prior probability. The probability of the data (irrespective of the hypothesis)

$$C_{MAP} = arg \max_{c \in C} P(d|c) \; P(c)$$

$$= arg \max_{c \in C} P(x_1, x_2, …, x_n|c) \, p(c)$$

2) Support Vector Machine (SVM): The original version of the Support Vector Machine (SVM) was formulated by Vapnik and Chervonenkis and adapted this version into a nonlinear formulation. SVM is a powerful statistical machine-learning technique [20]. Due to the ability to handle millions of inputs and good performance, SVM was widely used in text classification studies. SVM was originally designed for binary classification tasks. However, many researchers work on multi-class problems using this technique [19].

Since SVMs are traditionally used for the binary classification, a Multiple-SVM (MSVM) for multi-class problems is proposed by [20]. One-vs-One is a technique for multi-class SVM that builds $N(N – 1)$ classifiers as follows [21]:

$$f(x) = arg \max_i \left(\sum_j f_{ij}(x) \right)$$

3) Decision Tree (DT): A decision tree is a tree whose internal nodes are tests and leaf nodes are categories. Each internal node test one attribute and each branch from a node selects one value for the attribute. The attribute used to make the decision is not defined. So, attribute which gives maximum information can be used and the leaf node predicts a category or class. The decision trees are not limited to Boolean functions, but they can be extended for general categorically values functions [22].

$$H\left(\frac{p}{n+p}, \frac{n}{n+p} \right) = -\frac{p}{n+p} \log_2 \frac{p}{n+p} - \frac{n}{n+p} \log_2 \frac{n}{n+p}$$

Fig. 7.   Process of Feature Expansion by Lexical Ontology Approach.

V. EXPERIMENTS

A. Dataset

The data collection for this research is tourism review data about a tourism place, i.e. point of interest. The dataset is collected from reviews written by users regarding a point of interest, such as Penang Hill. The data is collected from Trip Advisor website. In this evaluation, a total of five point of interest are identified, and their reviews are collected. These five POIs are Genting Highlands Theme Park, Cameron Highlands Boh's Tea Centre, Club Med Cherating Beach, Escape Penang, and Penang Hill. For each POI, we have collected 50 reviews within a certain date range. If a point of interest has more reviews, the date range will be shorter and vice versa. Table V lists the Point of Interests and Reviews Data Range.

In this dataset, each review will be stored at the sentence level. Fig. 8 shows the output of the sentence segmentation of reviews. Table VI shows a listing of sentences for one example POI.

B. Data Benchmarking

To prepare the golden standard data collection, the reviews sentences are annotated based on the opinion types, i.e. "Attraction", "Fee", "Time", "Weather", "Transport", "Service" or "Food". A total of three annotators are recruited to perform the annotation on the sentences.

TABLE V. LIST OF POINT OF INTEREST AND REVIEWS DATA RANGE

Point of interest (POI)	Date		Number of Reviews Collected
	From	To	
Genting Highlands Theme Park	19/2/2019	17/8/2019	50
Cameron Highlands Boh's Tea Centre	12/5/2019	8/11/2019	50
Club Med Cherating Beach	2/12/2019	21/1/2020	50
Escape Penang	26/7/2019	16/8/2019	50
Penang Hill	26/11/2019	30/12/2019	50

Fig. 8. Output of Sentence Segmentation of Reviews.

TABLE VI. LISTING OF SENTENCES FOR ONE POI

Review Sentence
lots of indoor fun activities for kids and adults.
Pocket friendly as well.
We had a great time and 5 nights flew by so quickly.
Totally rested despite the rain.
went there by cable car.
Self Service check-in was a breeze.
There must be hundreds of food and drink options around.

Before an annotator starts his task, he is given the guideline and a set of data for the annotation task. Using the guideline of the definition of opinion type, the annotator is required to manually annotate each review as one of the seven opinion types, for example, "Attraction", "Fee" based on his judgement on the opinion type that best match the contesnt of the sentence. If the review sentence does not fit any of the seven opinion types, the review will be annotated with "N/A".

C. Data Statistics

From the 250 reviews data, 1691 sentences were obtained from the sentence segmentation. From a total of 1691 sentences, a total of 1576 review sentences are annotated with the opinion type. If there is more than an opinion type for a sentence, the best type will be chosen. Table VII shows the number of review sentences collected for each point of interest. Since 50 reviews will be collected is based on POI, and each review will have a different number of sentences depending on the length of the review, hence, it is natural to have difference numbers of review sentences for different POI.

As the data collection was made based on POI, there are also distinct differences in the number of review sentences related to each opinion type. Some opinions tend to be mentioned more in the reviews, e.g. Attraction, compared to Weather in this data collection. The number of review sentences for each opinion type is shown in Table VIII and Table IX (with POI).

TABLE VII. NUMBER OF REVIEW SENTENCES FOR EACH POI

Point of Interest (POI)	Number of Reviews Sentences
Genting Highlands Theme Park	275
Cameron Highlands Boh's Tea Centre	410
Club Med Cherating Beach	408
Escape Penang	300
Penang Hill	298
Total	1691

TABLE VIII. NUMBER OF REVIEW SENTENCES FOR EACH OPINION TYPE

Opinion Type	Number of Reviews Sentences
Attraction	631
Fee	86
Time	183
Weather	33
Transport	74
Service	229
Food	154

TABLE IX.   NUMBER OF REVIEW SENTENCES COLLECTED FOR EACH POI FOR ITS CORRESPONDING OPINION TYPE

POI / Opinion Type	Genting Highlands Theme Park	Cameron Highlands Boh's Tea Centre	Club Med Cherating Beach	Escape Penang	Penang Hill
Attraction	134	120	123	143	111
Fee	16	21	4	5	40
Food	23	44	64	12	11
Service	18	3	119	82	7
Time	24	41	29	26	63
Transport	23	31	0	1	19
Weather	5	3	24	2	9
Total	275	410	408	300	298

D. Evaluation Setting

In the experiment, each POI is selected in turns to be used as training data, with the remaining POIs used as testing. The training and testing approach used in this experiment is similar to cross-domain learning, where a source domain (i.e. source POI) is used as training and target domain (i.e. target POIs) are used as testing. As for the baseline comparison, in our experiment, we compare opinion type detection using the proposed approach with its baselines. Two experiment settings conducted are listed as below.

1) Classifier Selection for Opinion Identification. This experiment is carried out to select a classifier that will be used in our experiment. The experiment compares three classifiers, i.e. SVM, NB, and DT in their classification accuracy in opinion type detection. In this experiment, features are extracted using NLP and BOW approaches.

2) Feature Expansion for Opinion Identification. This experiment is carried out to compare the proposed approach with its baselines, i.e. LO+NLP vs NLP and LO+BOW vs BOW. The experiment compares opinion type identification under two settings, i. using Source Target with low number of training, i.e. $SOURCE_{LOW}$ and ii. using Source Target with high number of training, i.e. $SOURCE_{HIGH}$.

All evaluation will be performed based on the seven opinion types, and five POI of review sentences which are Genting Highlands Theme Park, Cameron Highlands Boh's Tea Centre, Club Med Cherating Beach, Escape Penang, and Penang Hill.

E. Experiment Results

1) Classifier selection for opinion identification: In this experiment, Genting POI is used as source training POI, while other POIs are used for testing. This experiment is carried out to select a classifier that will be used in the experiment for feature expansion evaluation.

Natural Language Processing (NLP)

By analyzing the results in Table X, the overall accuracy for each POI is higher when using SVM classifier. For

"Cameron", the overall accuracy by using SVM is higher at 61.59%, the percentage is higher by 14.82% compared with NB classifier (46.77%), and 24.33% compared to DT classifier (37.26). The results also can be seen for "Penang Hill", the overall accuracy by using SVM is higher at 51.54%, the percentage is higher by 8.46% compared with NB classifier (43.08%) and 16.92% compared to DT classifier (34.62). For "Escape", the overall accuracy by using SVM is higher at 63.47%, the percentage is higher by 9.6% compared with NB classifier (53.87%), and 7.38% compared to DT classifier (56.09%). For "Cherating", the overall accuracy by using SVM is higher at 52.69%, the percentage is higher by 14.73% compared with NB classifier (37.96%), and 1.12% compared to DT classifier (51.84%). From the results, it is observed that SVM is higher than NB and DT among all the four POI, hence, SVM will be chosen for further discussion for NLP approach.

Bag-of-Words (BOW)

By analyzing the results in Table XI, the overall accuracy for each POI is higher when using SVM and DT classifier. For "Cameron", the overall accuracy by using SVM is 57.41%, the percentage is higher by 11.78% compared with NB classifier (45.63%), and 1.14% compared to DT classifier (56.27). The results also can be seen for "Escape", the overall accuracy by using SVM is higher at 62.36%, the percentage is higher by 9.59% compared with NB classifier (52.77%) and 10.7% compared to DT classifier (51.66%).

For "Penang Hill", the overall accuracy by using DT is higher at 60.38%, the percentage is higher by 4.23% compared with SVM classifier (56.15%), and 17.69% compared to NB classifier (42.69%). For "Cherating", the overall accuracy by using DT is higher at 53.26%, the percentage is higher by 1.42% compared with SVM classifier (51.84%), and 17.57% compared to NB classifier (35.69%). Since only SVM perform better in NLP approach, hence for consistency purpose, only SVM classifier will be chosen for next evaluation.

TABLE X.   OVERALL ACCURACY OF EACH POI BY EACH CLASSIFIER IN NLP APPROACH (NOUN FEATURE)

Classifier / POI	SVM	NB	DT
Cameron	61.59	46.77	37.26
Penang Hill	51.54	43.08	34.62
Escape	63.47	53.87	56.09
Cherating	52.69	37.96	51.84

TABLE XI.   OVERALL ACCURACY OF EACH POI BY EACH CLASSIFIER IN BOW APPROACH

Classifier / POI	SVM	NB	DT
Cameron	57.41	45.63	56.27
Penang Hill	56.15	42.69	60.38
Escape	62.36	52.77	51.66
Cherating	51.84	35.69	53.26

2) Feature expansion for opinion type identification: This experiment is carried out to compare the proposed approach with its baselines, i.e. LO+NLP vs NLP and LO+BOW vs BOW under two settings, one is using Source Target with low number of training set, SOURCE$_{LOW}$ (Genting as training data, remaining POIs as testing data), second is using Source Target with high number of training set, SOURCE$_{HIGH}$ (Cameron as training data, remaining POIs as testing data).

SOURCE$_{LOW}$ (Genting as training data, remaining POIs as testing data)

Table XII illustrates the results for the overall accuracy for "Genting" as a training dataset. By studying the table, we can see that there is no difference in the overall accuracy as all the approach resulted in 0.57.

SOURCE$_{HIGH}$ (Cameron as training data, remaining POIs as testing data)

Table XIII shows the results for the overall accuracy for "Cameron" as a training dataset. BOW+LO. From the results, there is an accuracy of 0.60 by using LO+BOW which is 3% higher compared to BOW approach (0.57), with 6% higher compare to NLP (Noun) approach (0.54) and NLP+LO approach (0.54). In addition, NLP+LO approach does not perform well in NLP approach where the same results are presented for NLP and NLP+LO which is 0.54.

Table XIV shows that BOW+LO perform the best for SOURCE$_{HIGH}$ over SOURCE$_{LOW}$ in overall accuracy and over other approaches (BOW, NLP, NLP+LO). This proves that the number of training data has a significant impact on the classification accuracy. From the results, the highest accuracy of 0.60 is achieved by using BOW+LO. Therefore, we can conclude that LO can be potentially used with BOW (BOW+LO) in achieving a better overall accuracy.

TABLE XII. OVERALL ACCURACY FOR BOW VS NLP VS LO (GENTING AS TRAINING DATASET)

Approach / POI	BOW	NLP	BOW+LO	NLP+LO
Cameron	0.57	0.62	0.57	0.61
Penang Hill	0.56	0.52	0.55	0.52
Genting	0.62	0.63	0.62	0.62
Cherating	0.52	0.53	0.53	0.53
Overall Accuracy	0.57	0.57	0.57	0.57

TABLE XIII. OVERALL ACCURACY FOR BOW VS NLP VS LO (CAMERON AS TRAINING DATASET)

Approach / POI	BOW	NLP	BOW+LO	NLP+LO
Escape	0.54	0.53	0.54	0.53
Penang Hill	0.62	0.57	0.58	0.57
Genting	0.62	0.59	0.68	0.59
Cherating	0.50	0.48	0.59	0.48
Overall Accuracy	0.57	0.54	0.60	0.54

TABLE XIV. OVERALL ACCURACY FOR SOURCE$_{LOW}$ VS SOURCE$_{HIGH}$

Approach / POI	BOW	NLP	BOW+LO	NLP+LO
SOURCE$_{LOW}$	0.57	0.57	0.57	0.57
SOURCE$_{HIGH}$	0.57	0.54	0.60	0.54

Note: SOURCE$_{LOW}$ (Genting as training source) SOURCE$_{HIGH}$ (Cameron as training source)

VI. CONCLUSION

This research aims to help the tourist to easily digest the vast availability of opinion by categorizing the reviews. Specifically, we improve of opinion type detection via keyword expansion to address the out-of-vocabulary (OOV) and limited labeled data issues.

From this study, we found that WordNet's labels of semantic relations are useful for the research of feature expansion. This is validated from the experiment that shown that our proposed feature expansion approach is able to improve opinion type detection with reasonable accuracy. Better accuracy can be seen for BOW+LO (in Table XIII) as well as when SOURCE$_{HIGH}$, i.e. when larger sentences, compared to the one with lesser sentences, SOURCE$_{LOW}$ are used as training data. This result suggests that the former could yield more keywords/features to be expanded and trained.

In summary, opinion type detection is important as it helps to automatically categorize customer review according to opinion type. This is convenient for customer and it could improve the way of how information can be selected to reach its users by filter the information they need. Designing a good opinion type detection framework is challenging as it involves solving problems at various stages ranging from training reviews collection, features selection, classification of reviews, and building model. In order to verify all these stages, the proposed feature expansion has been evaluated with real user reviews and data collection. A positive outcome in terms of performance accuracy is achieved from the evaluation and this motivates us to move forward to further investigation the potential of other semantic relationships in the adapted LO as future work.

REFERENCES

[1] Z. Kavasoglu, "Personalized Summarization Of Customer Reviews Based On User's Browsing History," International Journal on Computer Science and Information Systems, vol 8(2), pp. 147–158, 2013.

[2] A. Khan, B. Baharudin, L. H. Lee, & K. Khan, "A review of machine learning algorithms for text-documents classification," Journal of advances in information technology, vol 1(1), pp. 4–20, Feb. 2010.

[3] Q. Li, S. Li, S. Zhang, & J. Hu, "A Review of Text Corpus-Based Tourism Big Data Mining," Applied Sciences, vol. 9, pp 1–27, Aug 2019.

[4] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes and Donald Brown, "Text Classification Algorithms: A Survey," In Information 2019, vol 10, pp. 1-68, 2019.

[5] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, S. Emrich, & M. Sedlmair, "More than Bags of Words: Sentiment Analysis with Word Embeddings," Communication Methods and Measures, pp. 1931-2458, Apr. 2018.

[6] Sneha M Nakade & Sachin N Deshmukh, "Finding Semantic Orientation of Reviews Using Unsupervised PMI Algorithm," International Journal of Science and Research (IJSR), vol 5(2), pp. 2107–2110, Feb. 2016.

[7]   M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews," Proceedings of the 19th National Conference on Artifical Intelligence, vol. 4, no. 4, pp. 755–760, No2004.

[8]   K. Vivekanandan, & J. S. Aravindan, "Aspect-based opinion mining: A survey," International Journal of Computer Applications, vol 106(3), pp. 21–26, Nov. 2014.

[9]   S. Abeysinghe, I. Manchanayake, C. Samarajeewa, P. Rathnayaka, M. J. Walpola, R. Nawaratne, T. Bandaragoda & D. Alahakoon, "Enhancing Decision Making Capacity in Tourism Domain Using Social Media Analytics," 2018 18th International Conference on Advances in ICT for Emerging Regions, ICTer 2018 – Proceedings, pp. 369–375, Sept. 2018

[10] S. M. Rezaeinia, A. Ghodsi, & R. Rahmani, "Improving the accuracy of pre-trained word embeddings for sentiment analysis," arXiv 2017, arXiv:1711.08609.

[11] Muhammad Afzaal, Muhammad Usman, and Alvis Fong, "Tourism Mobile App With Aspect-Based Sentiment Classification Framework for Tourist Reviews," IEEE Transactions on Consumer Electronics, vol. 65, pp 233–241, 2019.

[12] K. Soo-Min and H. Eduard, "Automatic Detection of Opinion Bearing Words and Sentences," Information Sciences Institute University of Southern California, vol 10, pp. 60–66, Apr. 2005.

[13] L. Wu, S. C. H. Hoi and N. Yu, "Semantics-Preserving Bag-of-Words Models and Applications," in IEEE Transactions on Image Processing, vol. 19, no. 7, pp. 1908-1920, July 2010.

[14] M.F. Porter, "An algorithm for suffix stripping," Program, vol 14(3), pp. 130–137, July 1980.

[15] Bruce M. Hill, "Posterior distribution of percentiles: Bayes' theorem for sampling from a population," Journal of the American Statistical Association, vol 63, pp. 677–691, Nov. 2014.

[16] Z. Qu, X. Song, S. Zheng, X. Wang, X. Song, & Z. Li, "Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification," Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China. pp. 677–680, 2018.

[17] L. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, "Deep Learning," MIT Press: Cambridge, MA, USA, vol 1, 2016.

[18] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network framework for dimensionality reduction," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, pp. 490-497, 2014.

[19] G. Bo, and H. Xianwu, "SVM Multi-Class Classification," Journal of Data Acquisition & Processing, vol 3, pp. 47–52, 2006.

[20] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "Foundations of Machine Learning," MIT Press: Cambridge, MA, USA, 2012.

[21] R. Farzi, and V. Bolandi, " Estimation of organic facies using ensemble methods in comparison with conventional intelligent approaches: A case study of the South Pars Gas Field, Persian Gulf, Iran," In Modeling Earth Systems and Environment, vol 2(105), pp. 1-13, 2016.

[22] A. Genkin, D. D. Lewis, & D. Madigan, "Large-scale Bayesian logistic regression for text categorization," Technometrics, vol 49(3), pp. 291–304, Aug 2007.

A Novel Approach for Computer Assisted Sleep Scoring Mechanism using ANN

Hemu Farooq*[1], Anuj Jain[2], V.K. Sharma[3], Iflah Aijaz[4], Sheikh Mohammad Idrees*[5]

Department of Electronics and Communication Engineering, Bhagwant University, Rajasthan, India[1, 3]
School of Electronics and Electrical Engineering, Lovely Professional University, Punjab, India [2]
Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India[4]
Department of Computer Science (IDI), Norwegian University of Science and Technology, Norway[5]

Abstract—**Sleep analysis and its categories in sleep scoring system is considered to be helpful in an area of sleep research and sleep medicine. The scheduled study employs novel approach for computer assisted automated sleep scoring system using physiological signals and Artificial neural network. The data collected were recorded for seven hour, 30 second epoch for each subject. The data procured from the physiological signal was controlled and prepared to expel degenerated signals in order to extract essential data or features used for the study. As, it is known human body distributes its own electrical signals which is needed to be eliminated and these are known as artifacts and they are needed to be filtered out. In this study, signal filtering is achieved by using Butterworth Low-Pass filter. The features extracted were trained and classified using an Artificial Neural Network classifier. Even though, it is a highly complicated concept, using same in biomedical field when engaged with electrical signals which is obtained from body is novel. The accuracy estimated for the system was found to be good and thus the procedure can be very helpful in clinics, particularly useful for neurologist for diagnosing the sleep disorders**.

Keywords—Sleep scoring stages; EEG; EMG; EOG; artificial neural network

I. INTRODUCTION

A. Sleep Scoring System and its Classification

Sleep is the emotional, mental and physical condition of body which is naturally recurring, distinguished by altered consciousness, comparatively inhibited of nearly sensor activities, inhibition of all voluntary muscles, and reduced interactions with surroundings [1]. Sleep is a highly conserved behavior across animal evolution. It is that condition in which a person goes from conscious to sub conscious state. Sleep is considered to be an essential element in many physiological processes which includes consolidation of memories and the processing of experiences. The significance of sleep is emphasized by the indications experienced by those who suffer from the disorders of sleep. Individuals suffering from disorders of sleep like insomnia, narcolepsy, nocturnal breathing etc. do not get adequate sleep and thus need to be treated [2].

Scoring mechanism procedure is done by the inspection and examination of polysomnography (PSG test) which can be obtained in the presence of sleep specialist. It is considerably essential feature in the area of sleep medicine and research which is going on sleep scoring. On the other hand, PSG (Polysomnography) is a test which involves transmission of records of various biological signals like EOG (Electrooculogram), EEG (Electroencephalogram), EMG (Electromyogram). All of recorded signals collected known as epoch which are obtained for short span of time are inspected carefully which designates different sleep classification [3].

Normally, sleep scoring system categorizes sleep into different types of classification such as Waking, Rapid Eye Movement (REM), Non-Rapid Eye movement (NREM). Waking is the stage when the individual prepares for sleep. In this stage, person is at leisure and in a calm position, also Electromyography of a person conceivably is elevated or in average mode. NREM are further divided into four stages i.e., stage 1 which occurs after little time is spent in waking stage and Electromyography is diminished, stage 2 in which subject is in the condition of sleep and EMG value decreases, stage 3 in which EMG gets reduced due to relaxed muscles, stage 4 which records the moderate EMG and subject is termed asleep. REM is the stage in which EMG demonstrates muscle tones called as twitches and is distinguished by random or rapid movements of eye [4].

B. Artificial Neural Network (ANN)

(ANN) Artificial Neural Network is significantly engaged computing approach. Artificial Neural Network are being used for a different kind of purposes like recognition of image, pattern recognition, analyze and classification. It has become popular for the widely array of fields in the world of computing. The designing ease of balanced equation, execution of time and good accuracy has made (ANN) artificial neural network extremely functional classifier in terms of calculating the large amount of data [5] [14].

The modeling and design of (ANN) artificial neural networks is influenced from human neural network. The human brain consists of cells known as neuron whose main purpose is processing, assembling and distribution of electrical signals while an artificial neuron is information processing unit that is similar in operation of biological neural network. Arithmetically, it can be stated as operation performed by an activation function which receives input x and weight w. It can be written as:

$$net = \sum_{i=0}^{n} w_i\, x_i \tag{1}$$

Where x_i is the input to neuron and w_i is weight related to input. An activation function f_{act} is applied to net giving as a

---

*Corresponding Author

result the output of neuron O and amplitude of neuron in output is limited by this function [6].

The basic characteristics of neural networks when acquired for computational cause are its capability to adapt as well as characteristics of parallel-distributed memory and non-algorithm. It comprises of three operational layers i.e. input, output and hidden layers respectively as shown in Fig. 1. Also, it has been seen that here can be possibility of additional middle layers which performs an assemblage of various tasks. The middle layers made up of cells whose main function is to perform summing which is a product of preceding layer given out by output [7]. The resolution of hidden layers is the censorious work in terms of figuring and outlining the neural networks. The calculation of middle layers and its activation function are done moderately by artificial neural network by employing hit and trial method. This phase of procedure is known as training phase and the trained function of artificial neural network is required as per computation.

Fig. 1. An Artificial Neuron.

II. MATERIAL METHODS

The implementation is based on Artificial Neural Network classifier using MATLAB software. Fig. 2 shows the approach to carry out the work using ANN.

Fig. 2. Stepwise Proposed Work.

The research evaluates the following performance metrics to verify the superiority of the scheduled work employing Artificial Neural Network (ANN).

III. RESULTS

The standard data base of physiological signals was collected from different hospitals especially from the sleep laboratories in front of sleep specialists. The real-time Electroencephalograph signals were acquired using electrodes on surface of skin which are standard. Electromyography signals were acquired by utilizing electrodes kept below the persons chin which are in two in number. The Electrooculography signals were obtained by utilizing electrodes placed either below or above the eye or to the right or left of the eye. By analyzing the biomedical signals different stages of sleep was understood. The different steps of the research are as follows:

A. Signal Processing

The initial signal procured from EMG must be controlled which is needed for the utilization of ANN. The signs are prepared to expel degenerated signals, extract important data and to foresee values. Biomedical signs are frequently undermined by different waves which are not related in the slightest degree. Biological or specialized issues are the purpose behind these waves. A human body delivers its own electrical exercises and it can be seen utilizing the accompanying gadgets.

- ECG artifacts causes spike in the QRS complex stage.

- EOG artifacts appears by movements of eye and blinking.

- EMG are the high-frequency artifacts.

The most well-known technique to eliminate these artifacts is the technique of filtering utilizing high-pass channels, band-pass channels, and low-pass channels, contingent upon the artifact itself. The motivation behind preprocessing the signal is to enhance the signals of EMG by normalizing. EMG flag is fragmented into epoch of 30 seconds each with every epoch corresponding to various rest stages. In this, the signal filtering is achieved using Butterworth Low-pass channel (LPF) having 4 Hz to 4000 Hz of cut-off frequencies so that artifacts and noises are eradicated from the EMG signals.

B. Feature Extraction: Time Domain Analysis Entropy

Histogram of each epoch is used for computing the corresponding entropy.

$$Entr_{EEG} = -\sum_{j=1}^{N} \frac{n_j}{n} \ln \frac{n_j}{n} \tag{2}$$

y(i) is measured for the signal which is y inside each epoch and n is the sample numbers.

In order to calculate the histogram, bin number is termed as N.

y(i) samples whose values are in the jth bin range so that the number of values is denoted by nj.

The 75th percentile

The 75% of the values which lies below it are the one-third quartile or upper quartile.

$$card\left\{\frac{x(i)}{x(i)} < prc_{75}\right\} = \frac{75n}{100} \tag{3}$$

y(i) is measured for the signal which is y inside each epoch.

And the card is referred as the number of elements in set.

Skewness

Skewness is defined as distribution of random variable having asymmetry probability with reference to its mean value. The skewness ranges between positive, negative or not defined values. It is seen if the values of left side are further distant than that of the values of right side, then skewness value will be negative, likewise if the right side values are further remote than that on the left side values then skewness value is positive and if the value is zero then the distribution is symmetric.

$$skewness = \frac{M3}{M2\sqrt{M2}} \tag{4}$$

Where

Mean is denoted by M3 and standard deviation by M2

$$M_K = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^K \tag{5}$$

x(i) is the ith x value, n= sample size

Kurtosis

Likewise to skewness, kurtosis communicates the state of likelihood conveyance. It quantifies the commitment of remote qualities. Kurtosis of an ordinary distribution is equivalent to 3. At the point, it is seen that when the values of kurtosis is lesser than 3, it is termed as platykurtic distribution because outliers contribution becomes lesser than in normal distribution. Then again, when its is higher than 3, the outliers contribution is greater than ordinary dispersion and circulation is termed as leptokurtic distribution.

$$kurtosis = \frac{M4}{M2M2} \tag{6}$$

Where; Mean is denoted by M4 and standard deviation by M2

Hjorth Parameters

The three distinct parameters which are calculated from epochs having long 1s or longer, utilizing the variance as a parameter for measuring the signal activity are Hjorth parameters.

$$var = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - x)^2 \tag{7}$$

- Activity

$$A = var(x) \tag{8}$$

- Mobility

$$M_x = \frac{var(x\prime)}{var(x)} \tag{9}$$

where $x\prime$ stands for the first derivative of x.

- Complexity (Form Factor)

$$FF = \frac{M(x\prime)}{M(x)} = \frac{\frac{var(x")}{var(x\prime)}}{\frac{var(x\prime)}{var(x)}} \tag{10}$$

where x'' is the second derivative of x.

Variance

It is a number which indicates group variability of an individuals are wide. If the observation of each individual greatly varies, then variance becomes large and otherwise low with reference to its mean group. In a sample it is essential to differentiate population variance. Mathematically, variance population is the distribution from each data point to the square of mean in the population. It is given by:

$$\sigma^2 = \sum(X_i - X)^2 / N \tag{11}$$

where population variance denoted by σ^2,

population mean denoted by X,

ith element from the population denoted by X_i and number of elements in the population by N.

Root Mean Square (RMS)

The Root Mean Square is defined as the square root of squares of set of quantities or numbers. The Root Mean Square value of a continuous - time waveform is described as the square root of the arithmetic mean of the squares of the values. For a set of n values $(x_1, x_2, \dots x_n)$, the RMS values can be computed using the formula:

$$x_{rms} = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots\cdots x_n^2)} \tag{12}$$

Mean Absolute Deviation (MAD)

The quantitative data of sample is univariate and is measured using the mean absolute deviation. It is also referred as calculation of sample estimated by Mean Absolute Deviation featured as population parameter. The Mean Absolute Deviation computed for a univariate data set $(X_1, X_{2,\dots} x_n)$ is the median of absolute deviations from the median of the data. Mathematically it is represented as.

$$MAD = median\left(|X_i - \overline{X}|\right) \tag{13}$$

The Standard Deviation

It is a parameter which measures the dataset dispersion relative to its mean. It is computed as the square of variance by deciding the variation amid each data point analogous to its mean. If the data points are far from the mean value then the deviation will be higher in the dataset hence higher the range of dataset higher is the deviation.

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})} \tag{14}$$

Where n, the number of samples x(i) for the signal which is measured which is x within the epoch, the mean value of the signal x is represented by $\bar{x}$:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \tag{15}$$

C. Frequency Domain Analysis

Analyzing the spectrum of the signal is useful for describing the signal with the help of its components of frequency. The time shift and magnitude of the frequency components will provide the characteristics of the signal and hence allowing further classification of the signals.

Fourier Transform

Fourier transforms (FT) are commonly used in expressing the spectrum of a continuous signal. But it can be used only in case of continuous signals and not for discrete signals.

$$F(\omega) = F\big(f(t)\big) = \int_{-\infty}^{\infty} f(t)\, e^{-j\omega t} dt \qquad (16)$$

where $F(\omega)$ is the signal's spectrum, $F(f(t))$ is Fourier Transform of the function $f(t)$, and ω is frequency.

D. Estimation of Power Spectrum

EEG being randomly distributed signal, its procedure can't be anticipated. Despite the fact that the signal itself must be accessible for its investigation, diverse strategies for examination are utilized. Randomly distributed signal procedure creating the flag is vital to investigate. Well-known characteristics of various recognition that are anticipated with definite likelihood enable us to foresee the subsequent recognitions of the randomly distributed signal procedure and different techniques can be made to additionally function with signals created by this kind of procedure. Power spectrum parametric and non-parametric strategies can be used for forecasting.

E. Nonparametric Methods

The use of the received data is sufficient for defining the non-parametric methods, and there is no need for setting up the models to generate the signals. The techniques of periodogram and correlogram are the most common nonparametric methods. Thus, the most commonly used non-parametric method is periodogram and is defined as:

$$S_{ff}(\omega) = E\left\{\frac{1}{N}\left|F_\omega(\omega)\right|^2\right\} \approx \frac{1}{M}\sum_{\omega i=\omega l}^{\omega M}\frac{1}{N}|F_{\omega i}(\omega)^2| \qquad (17)$$

where number of realizations is M and number of samples in N. The power spectrum can be estimated from a single realization itself.

F. Parametric Methods

Parametric methods generally involve the creation of models for the starting of the signal. The model will later characterize the both its signal as well as its spectrum. Reduction of the data for representing the spectrum more realistically is done by the models. But choosing these models along with its order is a difficult task.

G. Artificial Neural Network

NNs, by and large, are designated from biological neural networks, however with significantly more straightforward structures which have learning and responding qualities. Some different attributes, are realized with designing methodologies rather than neuropsychological ones. The central qualities of NNs presently set in motion are versatile, parallel disseminated memory and non-algorithmic. ANN are as of now, utilized in investigation, order, pattern recognition, classification and examination of functionality of systems and signals. They additionally have been very fruitful in analyzing the bio signals to a vast degree, as well as the investigation and elucidation of Electroencephalograph amid subjective interest, sleep or rest. An Artificial Neural Network comprises of input and output layer that are one in number and hidden layer/s including which are one or more than one in number. The hidden layer includes cell that sum up the former layer multiplied with a weighting vector in the output through neurons. Every node or cell will then produce an output with transfer function as non- linear function which is referred to as the activation function.

The ANN can be utilized as classifier, with respect to the positive after effects of different investigations on this matter of subject. For classification, Artificial Neural Network classification was done by software called as Matrix Laboratory (MATLAB). Sleep epochs range used differs, based on the different stages of sleep combination being classified, training and testing of data were always used. The data obtained was segmented into training set, validating set and the testing set which constitutes 60%, 20%, 20% of total data respectively. There were three kinds of data classification seen by using this approach. The first approach of categorization varies only between the Wakeful stage, N1 stage, and REM stage, while the second alternative categorizes W, N1, N2, and REM sleep stage, and the third categorization scores all the stages of sleep. Input and target matrix are the most important representation of neural network.

The training set used in training of Artificial Neural Network is the input matrix and for adjusting the neural network depending on the differences amidst the target and output of the neural network is the target set. The obtained data were segmented into three sets, namely, training, validating and testing set using the function divider and are in the ratios of 60 % for training set and remaining 40 % of the testing set respectively. The sleep experts marked the correct stage of column representing zeroes and ones as one in the target matrix.

For creating the Artificial Neural Network, the function tool is opened in the generator in MATLAB software. The variables for the input as well as the target are chosen along with distributing the testing as well as training sets. The Levenberg-Marquardt algorithm optimizes the weights and bias update is chosen due to its high speed for the function of training the network. For creating the structure of ANN, the selected hidden neurons are required and the output is the variable net within all the network parameters. Fig. 3 show the structure of ANN.

The output represents the adjusted network (net), number of the learning epochs and momentary performance (perf). The performance of the network is set with the help of minimum square error method. For statistical evaluation of performance and accuracy of the confusion matrix for the ANN a comparison of the results and the targets is illustrated in Table I. The actual class is represented by the row and predicted class by the columns.

Fig. 3. The Structure of the ANN.

TABLE I. SIMPLE CONFUSION MATRIX

		Target Class	
		1	**0**
Output class	1	True positive	False positive
	0	False negative	True negative

The confusion matrix in MATLAB can be plotted using the plotconfusion function. The number of appropriate classifications into class 1 is represented by true positive, classification number incorrectly assigned to class 1 is the false positive, the case numbers where the sleep epoch is incorrectly anticipated belongs to class 0, is false negative and the times number where the sleep epoch is not correctly assigned to class 1 is true negative. Table II presents the explanation of a confusion matrix.

TABLE II. AN EXPLANATION OF CONFUSION MATRIX

Output class	W	Correctly classified	Incorrectly classified	Incorrectly classified	Positive Predictive values False discovery rate
	N1	Incorrectly classified	Correctly classified	Incorrectly classified	Positive Predictive values False discovery rate
Target class	REM	Incorrectly classified	Incorrectly classified	Correctly classified	Positive Predictive values False discovery rate
	True positive rate False Negative rate	True positive rate False Negative rate	True positive rate False Negative rate	Overall Accuracy Overall Inaccuracy	
	W	N1	REM		

Sensitivity, specificity, and efficiency of the neural network is evaluated for the statistical assessment of the accuracy of the ANN.

Sensitivity: It is the capability of determination of patient cases correctly. To evaluate the sensitivity, the proportion in patient cases for true positive is calculated. Mathematically, it can be written as:

TP/ (TP + FN) = Sensitivity

That is the Number of true positive assessment / Number of all positive assessment.

Specificity: It is the capability of determination of healthy cases correctly. To evaluate the specificity, the proportion in healthy cases for true negative is calculated. Mathematically, it can be written as:

TN/ (TN + FP) = Specificity

That is the Number of true negative assessment / Number of all negative assessment.

Accuracy: It's the capability of determination in order to differentiate the patient and healthy cases correctly. To evaluate the accuracy of a test, the proportion of all evaluated cases in true positive and true negative is calculated. Mathematically, it can be written as:

(TN + TP) / (TN+TP+FN+FP) = Accuracy

That is the Number of correct assessments / Number of all assessments.

Further the proposed system of performance is evaluated using precision, recall, F1 score.

Precision: The precision is termed as the ratio of correctly predicted positive observation to the total number of positive observations. High precision is related to low false positive rate. In other words, it can be written as:

TP/ (TP + FP) = Precision

Recall: The recall is the ratio of correctly predicted positive observations to all the observations in the actual class. In other words, it can be written as:

TP/ (TP + FN) = Recall

F1 Score: It is the weighted average of recall and precision. Hence it considers both false negative as well as false positive values. F1 is more useful when compared with accuracy when the distribution class is uneven. Mathematically, this can be stated as:

F1 Score = 2 * (Recall * Precision) / (Recall + Precision)

Fig. 4, 5, 6 and 7 shows the whole confusion matrix of proposed neural network training state, error histogram, gradient and performance, respectively. The three sets of collaborate confusion matrices that is training, testing and validation matrix are actually based on total confusion matrix. This total confusion matrix plot shows 100% correct classification for the proposed system. The representation of error histogram plot shows the error of this proposed system is very close to zero. At the beginning of training, cross entropy error and the plot performance is maximum. For the scheduled system, epoch 11 has shown best validation performance, and the cross-entropy error is very close to zero. Table III presents the results of statistical parameters while the comparative analysis of different methods is shown in Table IV.

Fig. 4. Confusion Matrix of the Proposed Model.

Fig. 5. Error Rate.

Fig. 6. Gradient.

Fig. 7. Performance.

TABLE III. RESULTS OF THE STATISTICAL PARAMETERS

	W	REM	NR1	NR2	NR3
variance	12.6690	8.3126	7.7941	7.4460	2.3177
RMS	3.5593	2.8831	2.7918	2.77287	1.5224
MAD	2.1429	1.3522	1.4349	1.5406	0.7406
SD	0.0036	1.6922e-05	-0.016	-0.020	-2.2597e-04

TABLE IV. COMPARATIVE ANALYSIS OF THE EXISTING METHODS WITH THE PROPOSED TECHNIQUE

Tittle	Publication Year	Stage Classification	Classification Method used	Accuracy
Extracted features for sleep stage classification using data driven methods. [8]	2007	W, S1, S2, S3, S4, REM	ANN	80 %
Comparison methods between visual and automatic analysis in normal patients: staging using neural network method [9]	1996	W, S1, S2, S3, S4,	ANN	82.34%
Computer-assisted sleep staging [10]	2001		Clustering	76.8 %
Elman-type feedback SOM for sleep stage classification [11]	2010	W, S1, S2, S3, S4	ANN	72.20%
Artificial Neural Network employing EEG, EMG and EOG [12]	2010	REM, S1, S2, S3, S4,	ANN	74.7%
Proposed technique	-	W, S1, S2, S3, REM	ANN	83.33%

The research evaluates the following performance metrics to verify the superiority of the scheduled work using Artificial Neural Network (ANN).

1) Accuracy=83.3333
2) Specificity=100
3) Sensitivity=80
4) Precision=50
5) Recall=100
6) F1-score=66.6667
7) gmean = 89.4427

These metrics are evaluated based on the true positive, true negative, false positive and false negative.

The following table gives the results of Variance, Root mean square, Mean absolute deviation and standard deviation.

The following table gives the comparative analysis of the proposed method and the existing research.

IV. Discussion

Using physiological signals through the classifier Artificial Neural Network has proved an effective approach of determining different stages of sleep with good accuracy. Non- invasive biological measurement method [13] that defines relationship between sleep stages and physiological signal through a developed classifier via mathematical tool has been achieved effectively. The accuracy obtained by Tagluk et al. [12] on scoring capability was obtained to be 74.7% which was considered good enough when contrasted to that acquired by Hanaoka et al. [15]. However, the motive of present study is to determine sleep stages using physiological signals (EEG, EMG, EOG) recorded simultaneously and fed to Artificial Neural Network.

V. Conclusion

The motive of the research was not to look at the outcomes obtained by comparison techniques, however to present a strategy that utilizes EEG, EMG, EOG signals to be given to a pre- prepared Artificial Neural Network classifier with a high precision. The current study proposes computer assisted sleep scoring system using ANN. The data base was obtained from different hospitals in presence of sleep analyst. The signals obtained were pre processed in order to EMG signals by the process of normalization. The extracted features were classified using ANN and the accuracy evaluated by performance metrics is 83.33%.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

[1] HJ. Park, JS. Oh, DU. Jeong, KS. Park "Automated Sleep Stage Scoring Using Hybrid Rule- and Case-Based Reasoning.", Computers and Biomedical Research , 33, pp. 330-349, 2000.

[2] David T. Krausman, Bel Air, Richard P. Allen, "Sleep Scoring apparatus and method. Google Patents," 12 April 2005.

[3] Penzel, T., Conradt, R., Computer based sleep recording and analysis. Sleep Med. Rev. 4:131-138, 2000.

[4] Tagluk, M. E., Sezgin, N., and Akin, M., "Estimation of Sleep Stages by an Artificial Neural Network employing EEG, EMG, EOG", Journal of Medical System, 2010, 34(4), pp.717-725.

[5] Tagluk, M. E., Sezgin, N., and Akin, M., "Estimation of Sleep Stages by an Artificial Neural Network employing EEG, EMG, EOG", Journal of Medical System, 2010, 34(4), pp.717-725.

[6] Mousmita Sarma, Kandarpa Kumar Sarma, "Fundamentals consideration of ANN", Phoneme-Based speech segmentation using hybrid soft computing frame work, April 2014; 47-75.

[7] Kulkarni, P., Ade R., "Incremental learning from unbalanced data with concept drift and missing features: a review", International journal of data Min knowledge management process, 2014; 15-29.

[8] Zoubek, L., Charbonnier, S., Lesecq, S., Buguet, A., & Chapotot, F. (2007). Feature selection for sleep/wake stages classification using data driven methods. Biomedical Signal Processing and Control, 2(3), 171-179.

[9] Schaltenbrand, N., Lengelle, R., Toussaint, M., Luthringer, R., Carelli, G., Jacqmin, A., ... & Macher, J. P. (1996). Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. Sleep, 19(1), 26-35.

[10] Agarwal, R., & Gotman, J. (2001). Computer-assisted sleep staging. IEEE Transactions on Biomedical Engineering, 48(12), 1412-1423.

[11] Shimada, T., Tamura, K., Fukami, T., & Saito, Y. (2010, July). The effect of using Elman-type feedback SOM for sleep stage diagnosis. In Complex Medical Engineering (CME), 2010 IEEE/ICME International Conference on (pp. 165-170). IEEE.

[12] Tagluk, M. E., Sezgin, N., & Akin, M. (2010). Estimation of sleep stages by an artificial neural network employing EEG, EMG and EOG. Journal of medical systems, 34(4), 717-725.

[13] Watanabe, T., and Watanabe, K., Noncontact method for sleep stage estimation. IEEE Trans. Biomed. Eng. 51 (10)1735–1748, 2004. doi:10.1109/TBME.2004.828037.

[14] Idrees, S. M., Alam, M. A., Agarwal, P., & Ansari, L. (2019, April). Effective Predictive Analytics and Modeling Based on Historical Data. In International Conference on Advances in Computing and Data Sciences (pp. 552-564). Springer, Singapore.

[15] Hanaoka, M., Kobayashi, M., and Yamazaki, H., Automated sleep stage scoring by decision tree learning, 23rd Annual EMBS International Conference, 1751–1754, 2001.

Cloud of Things (CoT) based Parking Prediction

Nazish Razzaq[1], Muhammad Asaad Subih[2], Madiha Khatoon[3], Dr. Amir Razi[4], Dr. Babur Hayat Malik[5]
Nimra Ashraf[6], Tehseen Kausar[7], Rashida Tarrar[8], Muhammad Usman Sabir[9], Syed Izaz ul Hassan Bukhari[10]
CS and IT Department
University of Lahore, Gujrat Campus

Abstract—**Cloud computing with an amalgamation of the internet of things (IoT) typically gave birth to an ideal field called Cloud of things (CoT). CoT maintains revolutionary services in every domain, but it has instantly become a rising star in smart transportation because a well-organized facility might present a challenge for dealing with the exponentially expanding people living in smart cities. Lack of management in transport can cause distress among people and nowadays parking has come to be one of the critical issues faced by the public daily. In this paper, we present a parking availability prediction model implemented within a geo-fence ranging from 100-150 meters based on cloud, IoT, and GIS. In contrast to present models, instead of offering no space or parking is full; our model accurately determines the ETA of a vehicle and checks the potential/chance of parking availability. It also calculates the time for the next parking space if the existing parking space accessibility is zero. Moreover, our model includes all the exogenous factors like weather, time zone conditions to gain prediction accuracy.**

Keywords—*Cloud computing; internet of things (IoT); parking; prediction; availability; Estimated Time of Arrival (ETA); Geofence; Geographic Information System (GIS)*

I. Introduction

Transportation establishes an essential foundation for our present society. The transportation system performs a vital role in personal mobility and for the financial development of all countries. Modern society has been confronting more environmental problems today, i.e. traffic congestions, higher fuel costs, pollution, and parking [2]. Rapid growth of traffic and limited transport resources has caused urban traffic pressure on society [16]. Modern vehicles nowadays are prepared with innovative implanted technologies like sensors, GPS, and communication capabilities [18]. Cloud computing and the internet of things (IoT) has provided various services to minimize traffic congestion as well as management problems [19]. The progress in cloud computing and the internet of things (IoT) have given a chance to a new field called Cloud of Things (CoT) which adopts both technologies to improve the expanding transportation issues, like significant traffic, blockage, and vehicle security [16] [17]. CoT further uses Geographic Information System (GIS) for managing transportation. The GIS helps to manage and monitor the land information [3].

In the transportation field, parking is considered a key segment [3]. Due to the immense growth in the world's population, there has been a massive increase in the usage of public and private vehicles [1]. Parking has remained a standard problem due to the constant increase in the use of vehicles everywhere like airports, malls, offices, hospitals, etc.

[8]. Traffic congestion due to widespread conveyances on the road is a distressing hassle at an international level and it has been developing exponentially. The primary cause of congestion is automobile parking [3].

When drivers are unable to find parking spots, they get frustrated and leave their cars in the middle of roads leading to problems such as "traffic congestion or Illegal parking" [2]. The Global Parking Survey, carried out by International Business Machines back in 2011, showed that on an international scale 20 drivers struggle while finding parking spaces daily, where 6 out of 10 give up immediately, and they are looking for parking make up about 30% of the city traffic [1] [6]. With the evolution of technology in the past few years, many means of transports and parking industries have made use of these technological advancements to create parking systems. With the use of cloud computing, IoT, sensors, GIS, and simultaneous data gathering, modernized parking systems allow drivers to get information about available parking space anytime anywhere [7]. These smart parking structures further assisted with RFID, WSN, and NFC makes it easier for a driver to find an available spot [3].

In this article, we propose a model for parking space occupation prediction in a particular fenced area using the Cloud of Things (CoT) which represents the combination of various advanced technologies. We also present a simulating evaluation of the subsequent model with a proper view.

The remainder of this paper is organized as follows. The next section discusses technological background automobile services. Section III presenys the past work done on parking prediction. Section IV describes our evaluation methodology. Section V demonstrates the simulative view of our model. The results are displayed in Section VI. Finally, Section VII concludes this paper.

II. Technology Background

A. Cloud Computing

It represents an innovative on-demand model that delivers hundreds of services over the web with a pay-as-you-go scheme [17]. Cloud computing has guaranteed the accessibility of information to everybody through the on-demand model. It enables the end-users to store, access, and process information over the web. It provides all the facilities without demanding intense investments [18]. Gartner Group and International Data Corporation (IDC) consider cloud computing a peak IT trends with high business potential ranging from $40-100 billion yearly [20]. Cloud computing could be a bundle of three distinct services named as

Infrastructure as a Service, Platform as a Service and Software as a service [18]. It is exemplary a benefit in various areas but it has helped a lot to revolutionize the shape of transport services within the automobile industry [19]. Cloud-based urban control systems help in traffic administration and control, road routing, vehicle safety, and parking [20].

B. Internet of Things

It is a promising technology to revolutionized human life with remote monitoring by connecting everything to the internet [16]. IoT helps in tracking, controlling, and monitoring remote gadgets associated with the internet. Things in IoT refer to any object which needs to be tracked i.e. lights at home, cars on the road, products in stock [17]. IoT practices three emerging technologies to track things i.e. Identification which uses RFID bar code to recognize the items and their status. Sensing and Communication use WSN to track the varying category of the things and then transfer this information [16] [18]. IoT resolved traffic congestion, accident detection, road safety, and parking problems with wireless remote sensing [19].

C. Geoinformation System

It is a geographical structure used to capture, store, control, and observe a broad range of spacial land information [5]. The geographical material of a particular area can be accessed with the help of GPS, WSN to organize, analyze, and interpret according to the specific choice of the problem. GIS helps in getting parking area information via sensors and makes it easier for drivers to find parking space [3].

D. Technology Innovation in Automobile Services

Cloud computing emerged with IoT has caused the transformation into the automobile field. In many scenarios, cloud computing and IoT work together for better business processes and operations to achieve customer satisfaction [17]. This innovative and versatile field is better identified as the Cloud of Things (CoT). Cloud computing with an accomplice of the IoT gathers data from remote areas with wireless sensors. At that time information can be utilized by the providers to assist drivers to discover the unoccupied spot remotely [19]. The mixture of technologies is the middleware pattern of advanced parking systems in the present era [20].

Table I shows how parking traditions changed over time. In traditional models finding parking space was a manual process and now due to advanced models of automobile, it has become easier to find the parking space ahead of time.

Fig. 1 and Fig. 2 show the mixture of two latest technologies cloud computing and the internet of things anonymously known as the cloud of things (CoT). CoT works as middleware that operates some or all features of both technologies. Cloud provides database services while IoT helps in a better interactive interface. Collectively they perform a better job than individually.

TABLE I. TRANSFORMATION OF AUTOMOBILE MODEL

Service	Traditional Model	Advance Model
Parking	- Driver roams around the parking lot to find the unoccupied spot. - Manual receipt of the parking. - If parking is full, no chance of advance parking. - Driver can not get information about parking occupancy until s/he reaches into lot.	- Track vehicle movement and road conditions. - Assist driver with real time parking lot information. - If the parking is full, driver gets alert ahead of time. - Assist driver to find the parking spot based on preference.

Fig. 1. Innovative Technology Architecture.

Fig. 2. Cloud of Things (CoT) Mechanism.

III. Related Work

Parking is the key feature of transportation, if the parking lot lacks anything; it can shake up everyday life [4]. Many techniques are invented to overwhelm this critical issue. But all these systems are not real-time which means parking can only be available if the driver has a reserve spot in advance [5] [9] [10]. The problem was first highlighted in 2007. The authors addressed, "Automation and Modernization of Car Parking Management" and proposed a framework based on the queuing model. Markov Chain approach operating the two merges technologies WSN and RFID. This approach contains information such as occupied parking spots, parking, and the arrival rate of the vehicle. The parking rate reflects the lot's current state and the arrival rate; ETA calculation is needed for prediction. Since it operates a similar queue-based structure so the stream of inward vehicles is a counting procedure. Due to the finite capability of the lot and predictability is calculated through information exchanges among vehicles in a row. Markov Chain senses and tracks the vehicle whenever it enters a detecting field and transmits the signals of availability to the driver. This represents the individual number of vehicles parked in the lot simultaneously. When the lot is full but the line maintains some vehicles, it rejects the parking request and informs parking is unavailable. Markov Chain additionally includes guidance about parking availability, automatic payment, security and vandalism detection, etc. [10].

The Global Positioning System (GPS) technology is identified as the rising star for predicting parking availability; not only it provides navigation but also helps in discovering the closet lots, the existing position of the vehicle, and instruct towards the desired stop. The availability via GPS navigation system asks the driver his current location and destination. For computing the provisional distance of the vehicle by using the shortest path on which the vehicle will reach the target. Based on past occupation, information of the car parks, and estimated arrival time the probability of zones is predicted. These prediction approaches use the finest granularity of data and consider every bit of data in prediction at the given instant of time [11].To improve urban infrastructure and location-based services observation, and modeling human movement in an urban environment remains a crucial factor. Analyzing and uncovering human behavior patterns results in better predictive models. In Spain: Barcelona, a bicycling system based on spatiotemporal data of the bicycle stations and their parking patterns for 13 weeks were implemented. To gather the social and geographic viewpoints of the city and predict future bicycling station usage behavior, which compares to human movement in the city. This system used Bayesian Networks to gather the spatiotemporal data of human behavioral patterns Barcelona's public into three clusters i.e. outgoing, incoming, and flat. The outgoing cluster contains the order of people leaving for work from seven to eight AM and lunch breaks from two to three PM. The incoming group supports the data in operating time from seven to eight AM and early leaving time from one to two PM. The flat cluster station periodically is 66% full but mostly 15% available due to Barcelona's incline topography because people avoid parking at higher altitudes above on sea levels. This system predicts short and medium-term parking availability with 80%

accuracy. Where the short-term forecast is up to 5 minutes and the long-term forecast is up to 2 hours or 120 minutes utilizing 10-12 weeks of historic information. Shared bicycling usage is considered as a source of movement, and the flow of the city [12]. Exogenous factors like (weather, daytime) affect the mobility and use of resources [12] [13].

Barcelona shared bicycling model, further expanded in the city of Ireland: Dublin named the Dublin bike system but used General Additive Model (GAM) a class of the algorithms incorporating exogenous variables as well. Dublin bike system targeted 550 bikes and 44 stations which typically work from 05:00 AM till 00:30 AM every day. GAM is a two-stage prediction model that discovers the accessibility demands on various time stamps and calculates the allocation of remaining time for another possible parking space if current accessibility is zero. It hits the highest point on weekdays through morning and evening rush hours from 08:00 AM to 05:00 PM as well as during lunch hours 12:30 PM. On the ends of the week, the request lies from 07:00 AM till lunchtime and diminishes a short time later. The climate data is collected half-hourly and categorized into (rainy, foggy, normal) temperature ranges 1-21 ℃ for bike riders. This system calculates the prediction for three-time scales: 1) short term: prediction up to 5 minutes, 2) medium-term: up to 1 hour, 3) long-term: up to 24-hours. This system also provides the journey planner to the driver and helps him chooses the desired route from multiple routes according to his time constraints [13].

Intelligent Parking Assist (IPA) works on a similar Poisson process and communicates through the vehicle to infrastructure (V2I) and infrastructure to vehicle (I2V) for receiving information from the environment. It finds a reasonable spot based on all individual's preference and determine that behavior. The most suitable spot is determined using the chi-square goodness best fit method. The optimization prediction algorithm of IPA is probabilistic and works in real-time and on historical data. This framework maintains an advantage over all past models that; it operates on-street, off-street garages, also free spaces to estimate parking accessibility on the base of drivers expected arrival time to his/her target [14].

The cloud-based intelligent system is a three-tier architecture which dominates on all systems because it not only encourages the driver to discover, pick, and prevent the 'most suitable' accessible parking lot that is driving a car in a specific region, but also provide the driver with navigation directions for reaching this zone too. At the first tier, a user interface provides information about cloud-based services. The second tier is the communication layer; which possesses various wireless technologies and acts as a link between the user interface and the cloud layer for the trade of the data. Cloud Tier contains information storage space and computing resources for the parking service. It collects the 'big data' of possible lots, parking region, car area, user area, and profiles, etc. Each parking lot contains sensors whenever a car enters its proximity, sensors communicate with the cloud layer to check the lot occupation level then transmit the available spots (if any) direction to the driver [15].

Table II represents the comparsion of all the previous parking models, how this problem was solved with the help of technologies and the limitations of previous models. Not only these models solved this problem but also had limitations in their working.

TABLE II.   COMPARISON OF PREVIOUS PARKING METHODOLOGIES

Sr#	Name of paper	Technique	Limitation
1	Predicting parking lot occupancy [10].	Markov Chain Approach	**1.** It is based on the exchange of information among vehicles in a queue for parking occupancy which causes connectivity and storage issues. **2.** Vehicles in the queue are not guaranteed to get parking every time.
2	Intelligent GPS-based navigation system [11].	GPS based system	**1.** Parking availability is dynamic and can't be deterministic which means accessible spots at the period of decision-making cannot be guaranteed to be available at the arrival time. **2.** This model is unanalyzed in the situation when several vehicles search for the lot in real-time.
3	Predicting the pulse of the city through shared bicycling [12].	Prediction via sensing methodology	**1.** This model doesn't consider circumstantial features like weather, time-zone, season, special events. **2.** It only works during day hours.
4	Predicting waiting times for shared bicycles and parking lots [13].	Generalized Additive method	**1.** There is no facility of reserving space in advance. **2.** This model does not work for cars.
5	Intelligent parking assists [14].	Prediction Algorithm	**1.** Error ratio in prediction on average is 2.8% calculated by the Monte Carlo algorithm.
6	Cloud-based parking services for smart cities [15].	Cloud-based parking system	**1.** This model covers a partial area for prediction.

IV. METHODOLOGY

This section contains an algorithm for the prediction of parking space occupation. Our model works on real-time parking availability data based on geofence. Geo-fencing allows the specification of geological areas and produce entry/exit events when a device spans the border [5]. This model uses the blend of multiple technologies like cloud computing, the internet of things, GPS, within a certain geofence. The minimum range of geo-fence is considered as 100-150 meters around a parking lot. The driver gets inform whenever his/her vehicle crosses the threshold of the geofence zone and can request parking. If the parking request is made then the system checks the current state of the parking lot occupancy from the CoT database, calculates the no of vehicles that are about to leave the lot is greater than 10 minutes, and lastly free parking spots available at that timestamp. Our model tells the driver whether there is potential for parking availability or not. The time customers park their vehicles, both driver and owner, are notified with information about that particular spot through push notifications. Additionally, when a driver leaves a parking spot and drives away, it indicates that free space should be available again, so there is no need for drivers to roam around again and again to find parking [9].

Fig. 3 shows the flow of proposed model. This model is a two way process, it either entertains the parking request or gives exit call if there is no potential of availability. This model takes the request and responds according to the ETA or given time parameters of the demand. This model has an innovative feature which is determination 10 minutes stamp of ETA, this features enhances the quality of finding parking availability as it makes possible that maximum vehicles get the chance of ocuupaying the parking space.

The proposed model also calculates the waiting time prediction if parking is not available instantly. This model includes all exogenous factors, i.e. weather and day-hours.

Fig. 4 shows how parking requests are made and how the data is stored. The data streams of parking requests are stored in the form of tuples. The necessary parameters which are needed to allocate space are also ingested in tuples. In data contextualization, assignment of the parking spot is done on the basis of traffic conditions and ETA of the vehicle.

A. Dissemination of Parking IoT Information

In our approach, four values for each parking lot are considered. These are timestamp, the capacity of the lot, the number of parking spots currently engaged, and the ETA of the vehicle.

B. Prediction of Availability

Our model predicts three kinds of availability predictions using signs given below:

- indicates that the parking spot is available in the lot.

- shows that parking is unavailable at that time but calculates the ETA of the vehicle to check whether the

parking can be provided or not. If the ETA is greater than 10 minutes and there is a vehicle that will empty its occupied spot within these minutes. There is a chance of availability.

- ![P] notifies that parking is full.

Fig. 3. A Proposed Methodology for Parking Prediction.

Fig. 4. Automated Parking Prediction on CoT-GIS Platform.

C. Prediction of Waiting Time

This model calculates the waiting time for the next convenient spot if the current predictions indicate no spot is vacant at the moment. We assume that the arrival time-varying intensity. The intensity is typically extraordinary during active times such as morning or evening rush hours so there is a tremendous fluctuation in parking requests. In contrast, it is low late at night or early morning at the end of the week [13].

D. Weather Factors

Since our model is implemented using a combination of technologies so the weather would not affect the predictability. In most scenarios weather, is considered as the constant.

Fig. 5. Constant External Factors in Parking Prediction.

Fig. 5 shows that the projected model considers all external variables while predicting the parking availability. The CoT framework considers all the inputs from the user as well as external factors, i.e. weather, time-zone, distance calculation, and provide results with high accuracy.

V. SIMULATION

This section contains the stimulative view and evaluation of the proposed model. This model does not require any technology but makes better use of embedded technologies such as GPS and WSN for parking prediction. GPS and WSN also do not need high power or cost. Fig. 6 below shows the states of simulation for smart parking in the geo-fenced area whereas Fig. 7 shows the mechanism of simulative states. The detailed simulation represents states of the vehicle which are looking for parking within a smart city and geo-fence. State 1 indicates the detection of a vehicle in a particular area through wireless sensors. State 2 portrays the user request for parking. State 3 directs the parking directions to the driver if there is any unoccupied space. State 4 asks the driver for parking fees which is optional and depends on the parking lot administration. State 5 checks the identity of the driver via license for safety purposes. Finally, state 6 concludes the workflow and gives access to the driver towards the parking spot. In case, if the parking is not available, it enables users to book parking in advance within the next 24 hours. The new addition to the presented model is that with the help of CoT it provides potential parking which means if any vehicle from the parking lot is about to exit to less than 10 minutes, the requesting vehicle/driver can get that spot. This feature differentiates our model from all other classical models.

Fig. 6.   Block Diagram of Simulative Parking States.

Fig. 7.   Parking Availability Prediction Simulation.

A. State 1

This state shows how certain vehicle sinks in a particular geofenced area and its detection. Fig. 8 shows how the sunk vehicle can request a parking spot.

Fig. 8.   Vehicle Sinks in a Virtual Geofence.

B. State 2

The vehicle entered in a geo-fence can request parking in a near-by parking lot within a geo-fence area only. Fig. 9 shows the parking request made by the vehicle in a geo-fenced area.

C. State 3

If the user has requested parking and parking lot log has processed the request which results either in parking directions or parking is not available at the spot. Fig. 10 shows how the proposed model gives parking direction and its coordinates if parking is available.

D. State 4

This phase depends on parking lot administration if they charge fees the requester will pay the specific amount either to parking guard or via banking app based on driver's preference. Fig. 11 shows the payment methods for parking.

E. State 5

In this state, the parking guard will permit the driver by checking his/her driving license for safety purposes. Fig. 12 shows the basic inspection of vehicle entered in a parking lot which is done manually by guards as well as machines.

F. State 6

After the checking of the vehicle, the driver will get access to the parking spot along with parking coordinates. Fig. 13 concludes the whole procsess of parking by using coordinates to park in the allocated space.

Fig. 9.   Vehicle Requests Parking within Geofence.

Fig. 10. Vehicle Directing Towards the Parking Spot.

Fig. 12. Parking Permit by Parking Administration.

Fig. 11. Two Methods for Parking Payment.

Fig. 13. Parking Area Access to a Vehicle Parked.

VI. Results

The parking simulation is implemented on one of the parking lots of Pakistan named as Fatima Jinnah /Capital / F-9 Park situated in F-9, Islamabad, Islamabad Capital Territory, Pakistan coordinated as 33.7018° N, 73.0228° E. This parking lot has a parking capacity of 440 vehicles at a time. The parking trends of this parking lot over 7 days of the week and 24 hours per a day: (1) During working days (Monday to Friday) the parking demands lie from 12 PM to 6 PM which is off timing from working areas i.e. office, universities, etc. (2) During weekends (Saturday and Sunday) parking demands rise from 3 PM to 6 PM with little decrease till 9 PM since it's a park so people on holidays come here to spend time with their families. In order to test our model we built simulation on the below mentioned scenario.

Result Dicussion of Demo: A car enter in a geofence area. A map view of the parking lot is show to the driver. The model shows all possibilities of parking shown in Fig. 17. If the driver wants to get a parking spot s/he can request and Fig. 18 will be the result of hisparking request. System will show free, occupied and potentially available spots along with parking block, time and charges. The driver will further be able to get the direction coordinates and parking track of his assigned parking spot to park the vehicle.

Fig. 14 depicts the map view of F-9 Park's parking lot whereas Fig. 15 shows the demand of parking in F-9 Park seven days a week along with peak times and peak days.

Fig. 16 shows the 3D view of F-9 Park's parking lot which is simulated on Anylogic Simulation Software.

Fig. 19 and Fig. 20 shows how vehicle used assigned parking coordinates and followed the directions along with the route to park at allotted spot or block.

Fig. 15. Map and Ariel view of F-9 Park Parking loT.

Fig. 16. 3D view of F-9 Park Parking Lot Sketched on AnyLogic Personal Learning Edition.

Fig. 14. Parking Trends and Demand of F-9 Park.

Fig. 17. Parking Spots Available at Requested Timestamp.

Fig. 18. Parking Spot Availability at Requested Timestamp.

Fig. 19. Direction Coordinates of the Assigned Parking Spot.

Fig. 20. Parking Route and Vehicle Parked at Assigned Block.

VII. CONCLUSION

In this article, we present a model that helps in finding parking using CoT, GPS, WSN, and RFID within a specific geo-fence. Our proposed algorithm represents a step towards enhancing the ability of a Cloud-IoT based vehicle navigation system. It provides the probability of parking accessibility in the parking lot with their distance from the destination. Compared to other existing strategies, our approach considers

exogenous external components such as variable (weather, time-zone), which leads to progress execution. If parking is inaccessible, our model calculates the waiting time. We illustrated the approach on data from city center (Islamabd) car parking. Using the test bed implementation as starting point our early results are reassuring, but the proposed model needs to be analyzed in circumstances when several vehicles look for the parking space. Our proposed model can be deployed on both existing and upcoming parking systems for use in real world. Further, the future traffic management systems would gain an improved and reliable framework for parking problems. The internet of things and cloud computing development offers rise to many modern potential results in areas of smart cities and smart transportation. With the progress of the internet of things (IoT), cloud computing, and other techniques. It is predicted that future traffic management systems will utilize reasonable cost and high efficiency to serve the public with more helpful, fast, accurate, and quality traffic administrative results.

REFERENCES

[1] G. Watene, C. Ndegwa, and D. Musiega, "A GIS-Based Parking Management and Dissemination System," Proceedings of Sustainable Research and Innovation Conference, pp. 117-124, 2014.

[2] Yogesh, Tayade, and MD. Patil, "Advance prediction of parking space availability and other facilities for car parks in smart cities, "International Research Journal of Engineering and Technology, vol. 3, no. 5, pp. 2225-2228, 2016.

[3] N. Ya'acob, A.M Azize, and N.M.R.N.Z Alam, "Parking system using geographic information system (GIS)," IEEE Conference on Systems, Process and Control (ICSPC), pp. 12-16, 2016.

[4] S.S Adhatarao, O. Alfandi, A. Bochem, and D. Hogrefe, "Smart Parking System for Vehicles," IEEE Vehicular Networking Conference (VNC), pp. 189-190, 2014.

[5] F. Caicedo, C. Blazquez, and P. Miranda," Prediction of parking space availability in real-time," Expert Systems with Applications, vol. 39, no. 8, pp. 7281-7290, 2012.

[6] Y. Asakura and M. Kashiwadani," Effects of parking availability information on system performance: a simulation model approach," Proceedings of VNIS'94-1994 Vehicle Navigation and Information Systems Conference. IEEE, pp. 251-254, 1994

[7] Levy, Nadav, and I. Benenson, "GIS-based method for assessing city parking patterns." Journal of Transport Geography, vol. 46, pp. 220-231, 2015.

[8] M. Silva, G. Martín, M. Gould, R. Montoliu, J.T Sospedra, and J. Huerta, "An occupancy simulator for a smart parking system: Developmental design and experimental considerations," ISPRS International Journal of Geo-Information, vol. 8, no. 5, pp. 212, 2019.

[9] Rinne, Mikko, S. Törmä, and D. Kratinov, "Mobile crowdsensing of parking space using geofencing and activity recognition," 10th ITS European Congress, Helsinki, Finland, pp. 16-19, 2014.

[10] Caliskan, Murat, A. Barthels, B. Scheuermann, and M. Mauve, "Predicting parking lot occupancy in vehicular ad hoc networks," IEEE 65th Vehicular Technology Conference-VTC2007-Spring, pp. 277-281, 2007.

[11] Pullola, Sherisha, P. K. Atrey, and A.E Saddik, "Towards an intelligent

GPS-based vehicle navigation system for finding street parking lots," IEEE International Conference on Signal Processing and Communications, pp. 1251-1254, 2007.

[12] Froehlich, J. Edward, J. Neumann, and N. Oliver, "Sensing and predicting the pulse of the city through shared bicycling." Twenty-First International Joint Conference on Artificial Intelligence. 2009.

[13] Chen, Bei, F. Pinelli, M. Sinn, A. Botea, and F. Calabrese, "Uncertainty in urban mobility: Predicting waiting times for shared bicycles and parking lots," 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), pp. 53-58, 2013.

[14] Rajabioun, Tooraj, B. Foster, and P. Ioannou, "Intelligent parking assist," 21st Mediterranean Conference on Control and Automation. IEEE, pp. 1156-1161, 2013.

[15] Ji, Zhanlin, I. Ganchev, M. O'Droma, and X. Zhang, "A cloud-based intelligent car parking services for smart cities," 2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS). IEEE, pp. 1-4, 2014.

[16] Leng, Ying, and L. Zhao, "Novel design of intelligent internet-of-vehicles management system based on cloud-computing and internet-of-things." Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology. IEEE, vol. 6, pp. 3190-3193, 2011.

[17] Qin, Erwa, Y. Long, C. Zhang, and L. Huang, "Cloud computing and the internet of things: Technology innovation in automobile service," International Conference on Human Interface and the Management of Information. Springer, pp. 173-180, 2013.

[18] K. Ashokkumar, B. Sam, R. Arshadprabhu and Britto, " Cloud-based intelligent transport system," Procedia Computer Science, vol. 50, pp. 58-63, 2015.

[19] P.S. Saarika, K. Sandhya and T. Sudha," Smart transportation system using IoT," International Conference on Smart Technologies For Smart Nation (SmartTechCon). IEEE, pp. 1104-1107, 2017.

[20] Cao, Hung, and M. Wachowicz, "The design of an IoT-GIS platform for performing automated analytical tasks," Computers, Environment, and Urban Systems, vol.74, pp. 23-40, 2019.

xMatcher: Matching Extensible Markup Language Schemas using Semantic-based Techniques

Aola Yousfi[1], Moulay Hafid El Yazidi[2], Ahmed Zellou[3]

ENSIAS, Mohammed V University

in Rabat, Morocco

Abstract—Schema matching is a critical step in data integration systems. Most recent schema matching systems require a manual double-check of the matching results to add missed matches and remove incorrect matches. Manual correction is labor-intensive and time-consuming, however without it the results accuracy is significantly lower. In this paper, we present xMatcher, an approach to automatically match XML schemas. Given two schemas S_1 and S_2, xMatcher identifies semantically similar schema elements between S_1 and S_2. To obtain correct matches, xMatcher first transforms S_1 and S_2 into sets of words; then, it uses a context-based measure to identify the meanings of words in their contexts; next, it captures semantic relatedness between sets of words in different schemas; finally, it uses WordNet information to calculate the similarity values between semantically related sets and matches the pairs of sets whose similarity values are greater than or equal to 0.8. The results show that xMatcher provides superior matching accuracy compared to the state of the art matching systems. Overall, our proposal can be a stepping stone towards decreasing human assistance and overcoming the weaknesses of current matching initiatives in terms of matching accuracy.

Keywords—*Schema matching; matching accuracy; semantic similarity; semantic relatedness; WordNet*

I. Introduction

A. Motivation and Background

Schema matching aims at identifying semantic correspondences called matches [1], [2] in multiple schemas. It is critical for applications that manipulate data across different data sources because - if done correctly - it gives the end user a unified view over sources. We use an example to illustrate the schema matching problem. Let S_1 (Listing 1) and S_2 (Listing 2) be two XML schemas describing academic conferences. Our goal is to identify the matches in Fig. 1.

Although it is often desirable to define manually an integrated schema that represents all sources, this is often impossible for two main reasons: (1) the huge number of sources; and (2) the continuous updates. Thus, plenty of *automatic* schema matching systems have been developed (we refer the reader to [3], [4], [5], [6] for recent surveys and some state of the art matching systems). However, the term *automatic* is quite relative because even when humans do not help during the matching process, they help at the end correcting the results: adding missed matches and removing erroneous matches. Therefore, improving the accuracy of the output matches can significantly reduce humans' workload, and avoid possible mistakes humans might make. Also, it can save a considerable amount of time by leaving merely few results to correct.

```
1  <?xml version="1.0"?>
2  <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
3  <xs:element name="conference">
4   <xs:complexType>
5    <xs:element name="conference_name" type="xs:string"/>
6    <xs:element name="publication">
7     <xs:complexType>
8      <xs:element name="title" type="xs:string"/>
9      <xs:element name="author_name" type="xs:string"/>
10    </xs:complexType>
11   </xs:element>
12  </xs:complexType>
13 </xs:element>
14 </xs:schema>
```

Listing 1: S_1

```
1  <?xml version="1.0"?>
2  <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
3  <xs:element name="conference">
4   <xs:complexType>
5    <xs:element name="name" type="xs:string"/>
6    <xs:element name="paper">
7     <xs:complexType>
8      <xs:element name="title" type="xs:string"/>
9      <xs:element name="author_name" type="xs:string"/>
10    </xs:complexType>
11   </xs:element>
12  </xs:complexType>
13 </xs:element>
14 </xs:schema>
```

Listing 2: S_2

```
conference.conference_name <=> conference.name
conference.publication.title <=> conference.paper.title
conference.publication.author_name <=> conference.paper.author_name
```

Fig. 1. Matches between S_1 and S_2

Furthermore, the state of the art schema matching systems often reach a very moderate (sometimes poor) matching accuracy [2], and require loads of manual assistance to help correct the matching results [2]. In this paper, we will introduce a new schema matching system that will overcome these limitations as it is designed to achieve a high matching accuracy without any human assistance.

B. Challenges

Valuable as it is, producing high accuracy matches is also very difficult. First, schemas often use different naming conventions, e.g. *conference_name* (see Listing 1) and *name* (see Listing 2), or totally different words, e.g. *publication* (see Listing 1) and *paper* (see Listing 2). Second, schema elements are not fully independent from each other. For example, nested

elements in XML schemas. Third, a word can have multiple meanings. Finally, given a word W, WordNet hierarchy [7] connects W to other words through a wide variety of relations (e.g. hypernyms, hyponyms, meronyms); contributing unevenly to the definition of W. For example, according to WordNet the word *conference* has a direct hypernym (*meeting*) and five direct hyponyms (*symposium, seminar, colloquium, Potsdam conference*, and *Yalta conference*), both combined provide a comprehensive definition than one of them combined with *conference*'s meronym (*conferee*).

C. Contributions

In this paper, we introduce xMatcher, an approach to automatically match XML schemas. The key idea of xMatcher is to match XML schemas based on their semantics and with the objective of obtaining high accuracy matches, which reduces considerably humans' workload and offers a reliable and unified view over a large number of data sources. In particular, we make the following contributions:

- We propose a context-based measure to determine the meanings of words according to their contexts.
- We propose an automatic strategy to capture semantic relatedness between sets of words in different schemas.
- We present a semantic similarity measure over WordNet to calculate the semantic similarity between semantically related sets of words.
- We evaluate our similarity measure on a popular dataset and show that it provides correct results and surpasses the state of the art semantic measures and distances.
- We evaluate xMatcher on different real-world domains and show that it produces high accuracy matches and outperforms the state of the art systems in terms of matching accuracy.

The rest of this paper is organized as follows. Section II first reviews the state of the art schema and ontology matching systems, then it presents the state of the art similarity measures. Section III defines the problem of schema matching. Section IV describes xMatcher. Section V evaluates both our similarity measure and xMatcher in terms of matching accuracy. Section VI concludes this paper and discusses future work.

II. RELATED WORK

A. Schema and Ontology Matching Systems

Although it is not in its infancy, schema and ontology matching still an active research area. Indeed, the number of approaches available for schema and ontology matching increases continuously (we refer the reader to [3], [4], [5], [6], [8] for recent surveys and some existing matching systems). Also, the number of matching systems participating in the Ontology Alignment Evaluation Initiative (OAEI[1]) is increasing significantly. Before we proceed with the description of our new matching system xMatcher, we first review the state of the art matching systems that use WordNet as the matching space (e.g. ALIN [9]), and the top matching systems that participated in the 2018 edition of OAEI (e.g. Holontology [10], DOME [11], ALOD2Vec [12], and AgreementMakerLight [13]).

Holontology [10] is a modular holistic ontology matching system based on the Linear Program for Holistic Ontology Matching (LPHOM) system. It uses a combination of several similarity measures: Levenstein, Jaccard, and Lin to match two ontologies or multiple ontologies at once after it converts them into an internal predefined format. Then, Holontology transforms the results into alignments exported by RDF.

ALIN [9] is an interactive ontology matching system which takes as input two ontologies and deliver as output a set of alignments between them. It proceeds in two major steps. (1) It generates the initial mappings. (2) It waits for the human expert feedback and changes the mappings accordingly in order to improve the accuracy of the final results. This step is repeated until the human expert has no more mapping suggestions.

DOME (Deep Ontology MatchEr) [11] is a scalable matcher which uses doc2vec and exploits large texts that describe the concepts of the ontologies. To deal with the main issue of matching similar large texts, DOME uses topic modelling such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

ALOD2Vec [12] uses as external background knowledge source the WebIsALOD database of hypernym relations extracted from the Web. It also exploits element-based information and label-based information. In order to determine the similarity score between nodes of the knowledge graph (WebIsALOD is viewed as a knowledge graph), ALOD2Vec applies RDF2Vec.

AgreementMakerLight (AML) [13] is an ontology matching system which derives from AgreementMaker [14]. AML consists of two main modules: the ontology loading module and the ontology matching module. The ontology loading module loads the ontology files along with the external resources and then generates the ontology objects. The ontology matching module main goal is to align the ontology objects generated. The ontology loading module is extensible as it allows the virtual integration of new matching algorithms.

The matching systems presented above achieve acceptable results. The goal of this paper is to surpass the aforementioned systems in terms of *Precision*, *Recall*, *Overall*, and *F-Measure* (we refer the reader to subsection V-A for a definition of these quality metrics).

B. Semantic Similarity

1) Similarity Measures and Distances: One of the many possible approaches to discover matches is to compute the semantic similarity values between schema elements which is the approach we adopted for our matching system. Semantic similarity measures are one of the biggest pressing challenges facing the improvement of schema matching. According to [15], [16], [17], [18], [19], [20], [21], [22], semantic similarity measures are grouped into four categories: edge-based measures, information content-based measures, feature-based measures, and hybrid-based measures.

- **Edge-based measures (also known as path-based measures).** They determine the similarity between two concepts by considering both the length of the path that links the concepts in the taxonomy and the position of the

---

[1]http://oaei.ontologymatching.org/

concepts in the taxonomy [15], [16], [18], [20]. Examples include the shortest path-based measure [15].

- **Information content-based measures.** The main idea of these measures is that the more information two concepts have in common, the more semantically similar the concepts are [15], [23]. Examples include Resnik [24], Jiang & Conrath [25], Lin [26], and Nababteh [27].
- **Feature-based measures.** They use the properties of the concepts in a way that the more common features two concepts have and the less non-common features they have, the more semantically similar the two concepts are [15], [16], [18], [20], e.g. Tversky [28].
- **Hybrid-based measures.** They combine all the three aforementioned categories [16], [18], [20]. Zhou's measure is an example of hybrid-based measures [29].

But since the information content-based measures perform better than other categories (information content-based measures have the highest correlation coefficients when compared to the matching results provided by human experts) [15], we decided to direct our attention to the aforementioned information content-based measures that we will compare later to our semantic similarity measure.

WordNet [7] is a lexical database for the English language created by a research team at Princeton University. It groups words into sets of synonyms called *synsets*, which are interlinked by means of semantic relationships, for instance, ***is-a*** relationship which connects a hyponym to a hypernym. And it is commonly used by semantic similarity measures. Indeed, the following measures all use WordNet as an external resource.

Resnik's measure [24] computes the Information Content (IC) of the Least Common Subsumer (LCS) of two concepts denoted by a and b as follows:

$$Sim_{Resnik}(a,b) = IC(LCS(a,b)) \qquad (1)$$

Where:

- Given a concept C, we have IC(C)= - log(p(C)).
- $p(C) = \frac{frequency(C)}{N}$ refers to the probability of C.
- N refers to the total number of nouns.

The main issue of Resnik's measure is the following: any pair of concepts having the same LCS will definitely have the same semantic similarity value [15]. Luckily, Jiang & Conrath (J&C) and Lin found out a way to overcome Resnik's problem [25], [26]. In addition to the IC of the LCS, both J&C and Lin consider the IC of each concept [25], [26]. J&C define the distance between two concepts as follows [25]:

$$Dis_{J\&C}(a,b) = IC(a) + IC(b) - 2 \times IC(LCS(a,b)) \quad (2)$$

It differs from similarity measures in a way that the higher it gets, the less similar the two compared concepts are. Typically, given J&C's distance, one can revert it to serve as a similarity measure and vice versa. Conversions are made using equation 3. In this paper, we are going to use the similarity measure.

$$Sim_{J\&C}(a,b) = \begin{cases} 1, & \text{if } Dis_{J\&C}(a,b) = 0 \\ \frac{1}{Dis_{J\&C}(a,b)}, & \text{otherwise} \end{cases} \qquad (3)$$

Lin describes the semantic similarity between two concepts as follows [26]:

$$Sim_{Lin}(a,b) = \frac{2 \times IC(LCS(a,b))}{IC(a) + IC(b)} \qquad (4)$$

The main issue with Lin's measure is the following: if the IC of LCS, a, or b is equal to 0 then the semantic similarity value is equal to 0 as well [27].

In order to deal with Lin's problem, Nababteh suggests to divide 2 times the IC of the LCS of the two compared concepts by the sum of the IC of the direct hypernym of the first concept and the IC of the direct hypernym of the second concept [27].

$$Sim_{Nababteh}(a,b) = \frac{2 \times IC(LCS(a,b))}{IC(P(a)) + IC(P(b))} \qquad (5)$$

For the time being, the aforementioned semantic similarity measures are quite successful, they remain, however, some issues that require more attention. Indeed, according to [24], [25], [26], [27], the aforementioned measures might not provide the correct results all the time since when compared to the reference similarity values on Miller and Charles' (M&C) benchmark dataset the results were not promising.

2) Schema-Based Information and Instance-based Information: The Rivalry to Dominate Schema Matching: One of the most important choices that impacts the accuracy of the results returned by a similarity measure used by a schema matching system is the information used to find out semantic correspondences between schemas. Besides the external resources, a similarity measure may utilize either schema-based information, instance-based information, or both. In Table I, we present the advantages and disadvantages of each approach.

TABLE I. PROS AND CONS OF SCHEMA- AND INSTANCE-BASED PRACTICES

	Advantages	Disadvantages
Schema-based approach	- It uses the properties of the schema elements (e.g. labels, data types, integrity constraints). - Easy to implement. - They are fast.	- It does not produce good results when the properties of the schema elements are not available.
Instance-based approach	- It exploits the data stored at a given time which provides more details about the schema elements and hence improves the accuracy of the final results.	- Unavailable data may cause the matching system to stop functioning properly and exit. - Incorrect data may lead to false matches or miss true matches. - They operate slowly. - More complicated to implement than schema-based approaches.

Based on the information presented in Table I, we decided to use schema-based information to define our solution.

III. PROBLEM STATEMENT

In this section, we present definitions related to the schema matching problem. In this paper, we consider only XML schemas and leave other data representations for future work.

Definition 1 (Entity). *Let S be an XML schema. An entity e is used interchangeably to refer to a complex type element, a simple type element, or an attribute.*

Definition 2 (Set of Words). *Let S be an XML schema and n be the number of entities $(e_1, e_2, \ldots, e_n)$ it contains. Given an entity $e_1 \in S$, the set of words generated from e_1 is defined as follows $set_{e_1} = \{W_{1,1}, W_{1,2}, \ldots, W_{1,card(set_{e_1})}\}$, where $W_{1,1}, W_{1,2}, \ldots, W_{1,card(set_{e_1})}$ are words extracted from e_1.*

Remark: All the sets of words generated from S are defined as follows $SETS = \{set_{e_1}, set_{e_2}, \ldots, set_{e_n}\}$.

Definition 3 (Semantic Relatedness). *Let S_1 and S_2 be two schemas, and $SETS_1$ and $SETS_2$ be their respective sets of words. $set_1 \in SETS_1$ and $set_2 \in SETS_2$ are semantically related if they can be used together in the same schema. For example $\{conference, paper, title\}$ and $\{conference, paper, author\}$ from Listing 2 are semantically related.*

Definition 4 (Semantic Similarity). *Let S_1 and S_2 be two schemas, and $SETS_1$ and $SETS_2$ be their respective sets of words. $set_1 \in SETS_1$ and $set_2 \in SETS_2$ are semantically similar if they share the same meaning. Also, semantically similar sets cannot be used together in the same schema. For example $\{conference, publication, title\}$ in Listing 1 and $\{conference, paper, title\}$ in Listing 2 are semantically similar.*

Remark: Let S_1 and S_2 be two schemas, and $SETS_1$ and $SETS_2$ be their respective sets of words. If $set_1 \in SETS_1$ and $set_2 \in SETS_2$ are semantically similar then they are semantically related as well, e.g. $\{conference, publication, title\}$ and $\{conference, paper, title\}$. However $set_1 \in SETS_1$ and $set_2 \in SETS_2$ are semantically related does not necessarily imply that they are similar, e.g. $\{conference, paper, title\}$ and $\{conference, paper, author\}$.

Definition 5 (Problem Statement). *Given n schemas S_1, S_2, $\ldots$, S_n. Our goal is to maximize the accuracy of the matches discovered between S_1, S_2, $\ldots$, S_n and minimize humans' workload traditionally used to correct the matching results.*

Table II lists the notations used throughout this paper.

TABLE II. SUMMARY OF SYMBOL NOTATIONS

Notation	Description
S, e, c	XML schema, *entity*, complex type element
W, DB_{abbr}	Word from WordNet entries, abbreviations database
$set, SETS,$ $SETS'$	set of words, sets of words, semantically related pairs of word sets
$set_{e_{WordNet}}$, $set_{e_{abbreviations}}$, $set_{e_{expression}}$	set of WordNet entries that correspond to words in e, set of abbreviations database entries that correspond to words in e, set of full expressions of abbreviations contained in $set_{e_{abbreviations}}$
SM, Sim, Dis	sub-measure, similarity measure, distance
F, M	relatedness matrix, similarity matrix
$card(set)$	cardinality of set

In the next section, we introduce xMatcher the solution to the schema matching problem described in Definition 5.

IV. THE XMATCHER APPROACH

The xMatcher architecture (see Fig. 2) consists of three main modules: pre-matching, matching, and post-matching. Given two XML schemas S_1 and S_2, the *pre-matching module* ($\mu : S_1 \times S_2 \rightarrow SETS_1 \times SETS_2$) uses WordNet along with a database of abbreviations and applies fuzzy string matching to generate, from each *entity* in S_1 and S_2, a set of words. The *matching module* ($\phi : SETS_1 \times SETS_2 \rightarrow [0, 1]$) then identifies semantically related sets, for which it calculates the similarity values. Finally, the *post-matching module* ($\theta : [0, 1] \rightarrow Matches$) matches the *entities* whose similarity values are greater than or equal to 0.8. It is important to note that all three modules take place prior to any user request.

Fig. 2. The xMatcher Architecture

The rest of this section describes the pre-matching module (see subsection IV-A), the matching module (see subsection IV-B), and the post-matching module (see subsection IV-C).

A. The Pre-Matching Module

Before we proceed with the matching module, a pre-matching step is required since schemas use different naming conventions. The *entity* name might be an expression that does not belong to WordNet. Examples of such non-WordNet *entities* include abbreviations, concatenation of words, and words separated by underscores. Thus, we use two components, the *sets generator* and the *entities combiner*, to produce, for each *entity*, a set of words that help clarify its meaning.

Sets generator. Given a non-WordNet *entity* e, the sets generator proceeds in three steps (see Algorithm 1). (1) It uses fuzzy string matching to extract from e words $set_{e_{WordNet}}$ that syntactically correspond exactly or approximately to WordNet entries. (2) The sets generator then uses fuzzy string matching to see if $e \setminus set_{e_{WordNet}}$ includes abbreviations $set_{e_{abbreviations}}$ that correspond exactly or approximately to DB_{abbr}, in which case it substitutes $set_{e_{abbreviations}}$ for their full expression $set_{e_{expression}}$ available in DB_{abbr}. (3) It assigns a set of words to e, such that $set_e = set_{e_{WordNet}} \cup set_{e_{expression}}$.

Entities combiner. Let c be a complex type element, e be a non-complex type element included in c, and $set_c = set_{c_{WordNet}} \cup set_{c_{expression}}$ and $set_e = set_{e_{WordNet}} \cup set_{e_{expression}}$ be their respective sets of words. We made the following observation: the more words set_e contains, the more meaning e conveys. Therefore, we decided to utilize the context of e, which is the complex elements e belongs to, as follows $set_e \leftarrow set_e \cup set_c$. Algorithm 2 summarizes this.

Next, we use the sets of words to match schemas using relatedness matrices and a semantic similarity measure.

B. The Matching Module

The matching module consists of two major components: *relatedness determinator* and *similarity calculator*. The relatedness determinator uses relatedness matrices to capture

TABLE III. RELATEDNESS MATRIX

	W_{1_1}	W_{1_2}	$\cdots$	$W_{1_{card(set_{e_1})}}$
W_{2_1}	$f_{1,1}$	$f_{1,2}$	$\cdots$	$f_{1,card(set_{e_1})}$
W_{2_2}	$f_{2,1}$	$f_{2,2}$	$\cdots$	$f_{2,card(set_{e_1})}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$W_{2_{card(set_{e_2})}}$	$f_{card(set_{e_2}),1}$	$f_{card(set_{e_2}),2}$	$\cdots$	$f_{card(set_{e_2}),card(set_{e_1})}$

Algorithm 1 SetsGenerator(S)

Input:
    S
Output:
    $SETS$

1: **for** each e in S **do**
2:     **for** each $W \in WordNet$ in e **do**
3:         $set_e \leftarrow W$
4:     **end for**
5:     **for** each $abbr \in DB_{abbr}$ in e **do**
6:         Substitute $abbr$ for its full expression
7:         Add its full expression to set_e
8:     **end for**
9: **end for**
10: **return** $SETS$

Algorithm 2 EntitiesCombiner(S, $SETS$)

Input:
    S
    $SETS$
Output:
    $SETS$

1: **for** each e in S **do**
2:     **for** each c containing e **do**
3:         $set_e \leftarrow set_e \cup set_c$
4:     **end for**
5: **end for**
6: **return** $SETS$

---

semantic relatedness between different sets of words. Then, the similarity calculator exploits WordNet hierarchy to calculate the similarity between every semantically related sets.

1) Generating relatedness matrices: Prior to computing the semantic similarity values between different sets of words, we first must identify semantically related sets. This is very important for two main reasons. First, it narrows down the total number of computations, since we will only calculate the semantic similarity values between related sets. Second, let $e_1 \in S_1$ and $e_2 \in S_2$ be two *entities*, and set_{e_1} and set_{e_2} be their respective sets of words. Let's suppose that both e_1 and e_2 are not contained in any complex type element. Missing contexts implies that set_{e_1} and set_{e_2} convey poor meanings. Thus, identifying whether they are semantically related or not will help improve considerably their meanings. To this end, the relatedness determinator proceeds in two steps (Algorithm 3 summarizes this). First, it uses equation (6) to determine the meaning of a word according to the other words in the same set. Second, it employs fuzzy string matching and words synonyms available in WordNet to identify semantically related

sets. In the following, we explain these steps in more details.

Step 1: Identifying meanings of words. Let e be an *entity* and set_e be its set of words. Given that a word $W \in set_e$ may have more than one meaning, we use $set_e \setminus W$ to identify the meanings of W.

$$\forall\, W \in set_e,\ Sense(W) = \max_{1 \leq i \leq n} \sum_{j=1}^{card(set_e \setminus W)} \sum_{k=1}^{n_j} relatedness(s_i, s_{j,k}) \quad (6)$$

Where:

- s_i and $s_{j,k}$ are the ith sense of W (meaning of W in WordNet) and the kth sense of the jth word in $set_e \setminus W$, respectively.
- n and n_j are the total number of senses of W and the total number of senses of the jth word in $set_e \setminus W$, respectively.
- $relatedness$ returns the number of overlapping phrases or words between s_i and $s_{j,k}$.

Step 2: Identifying semantically related sets of words. Let $e_1 \in S_1$ and $e_2 \in S_2$ be two *entities* and $set_{e_1} = \{W_{1,1}, W_{1,2}, \ldots, W_{1,card(set_{e_1})}\} \in SETS_1$ and $set_{e_2} = \{W_{2,1}, W_{2,2}, \ldots, W_{2,card(set_{e_2})}\} \in SETS_2$ be their respective sets of words. We use fuzzy string matching to determine the words contained in both set_{e_1} and set_{e_2}. We display the results in a relatedness matrix $F = (f_{i,j})_{\substack{1 \leq i \leq card(set_{e_1}) \\ 1 \leq j \leq card(set_{e_2})}}$ (see Table III) whose individual items are defined as follows: $f_{i,j} = (o_{1_{i,j}}, o_{2_{i,j}})$, where $o_{1_{i,j}}$ is equal to 1 if W_{1_j} or one of its synonyms and W_{2_i} or one of its synonyms appear together in set_{e_1}, and 0 otherwise. Similarly, $o_{2_{i,j}}$ is equal to 1 if W_{1_j} or one of its synonyms and W_{2_i} or one of its synonyms appear together in set_{e_2}, and 0 otherwise.

Remark: $\forall \{i,j\} \in [\![1, card(set_{e_1})]\!] \times [\![1, card(set_{e_2})]\!]$. If W_{1_j} and W_{2_i} refer to the same word then $o_{1_{i,j}} = o_{2_{i,j}} = 1$.

We generated relatedness matrices for different real-world schemas (*Airfare*, *Automobiles*, *Books*, *Car Rentals*, *Hotels*, *Jobs*, *Movies*, and *Music Records*) extracted from the Web interfaces in the TEL dataset of the UIUC Web Integration Repository[2]. We noticed that semantically related sets (provided manually) are assigned matrices that contain more ones than zeros. Thus, we made the following conclusion: we say that two sets are semantically related if and only if the occurrence of 1 in F is greater than the occurrence of 0.

---

[2]http://metaquerier.cs.uiuc.edu/repository

Algorithm 3 RelatednessDeterminator($SETS_1$, $SETS_2$)

Input:
$SETS_1$, $SETS_2$
Output:
$SETS'$

1: **for** each W in $set_1 \in SETS_1$ **do**
2:   Identify the meaning of W using equation (6) */*Similarly, we identify the meanings of words in $SETS_2$*/*
3: **end for**
4: **for** each set_1 in $SETS_1$ **do**
5:   **for** each set_2 in $SETS_2$ **do**
6:     Determine semantically related sets based on their relatedness matrix
7:     Add semantically related sets to $SETS'$
8:   **end for**
9: **end for**
10: **return** $SETS'$

Next, we calculate the similarity between semantically related sets of words.

2) Calculating similarity values between entities: The similarity calculator operates in two steps (see Algorithm 4). First, it calculates the similarity between words. Then, it uses the results to calculate the similarity between sets of words.

Algorithm 4 SimilarityCalculator($SETS'_1$, $SETS'_2$)

Input:
$SETS'_1$
$SETS'_2$
Output:
V */*Similarity values between sets of $SETS'_1$ and sets of $SETS'_2$*/*

1: **for** each set_1 in $SETS'_1$ **do**
2:   **for** each set_2 in $SETS'_2$ **do**
3:     Calculate the similarity v between set_1 and set_2 using equation (17)
4:     $V \leftarrow V \cup v$
5:   **end for**
6: **end for**
7: **return** V

Step 1: Calculating the semantic similarity between words. Given a word $W \in WordNet$, we noticed that both its hypernyms and its direct hyponyms can be used together to define it. Hence, we decided to utilize this information to determine how similar two words are. Given two words $a, b \in WordNet$, comparing a to b is equivalent to comparing $\{a, P_a, H_a\}$ to $\{b, P_b, H_b\}$. Thus, the similarity calculator calculates the similarity between a and b (7), a and P_b (8), a and H_b (9), P_a and b (10), P_a and P_b (11), P_a and H_b (12), H_a and b (13), H_a and P_b (14), and H_a and H_b (15). P_a and P_b refer to the hypernyms of a and b, respectively. H_a and H_b refer to the direct hyponyms of a and b, respectively. Note that we consider only non-shared hypernyms hence $P_a \cap P_b = \phi$.

$$SM_1(a, b) = card(s_a \cap s_b) + card(s_a \cap (b \cup Sy_b)) + card(s_b \cap (a \cup Sy_a)) \quad (7)$$

$$SM_2(a, P_b) = \sum_{i=1}^{|P_b|} card(s_a \cap s_{P_{b_i}}) + card(s_a \cap (P_{b_i} \cup Sy_{P_{b_i}})) + card(s_{P_{b_i}} \cap (a \cup Sy_a)) \quad (8)$$

$$SM_3(a, H_b) = \sum_{i=1}^{|H_b|} card(s_a \cap s_{H_{b_i}}) + card(s_a \cap (H_{b_i} \cup Sy_{H_{b_i}})) + card(s_{H_{b_i}} \cap (a \cup Sy_a)) \quad (9)$$

$$SM_4(P_a, b) = \sum_{i=1}^{|P_a|} card(s_{P_{a_i}} \cap s_b) + card(s_{P_{a_i}} \cap (b \cup Sy_b)) + card(s_b \cap (P_{a_i} \cup Sy_{P_{a_i}})) \quad (10)$$

$$SM_5(P_a, P_b) = \sum_{i=1}^{|P_a|}\sum_{j=1}^{|P_b|} card(s_{P_{a_i}} \cap s_{P_{b_j}}) + card(s_{P_{a_i}} \cap (P_{b_j} \cup Sy_{P_{b_j}})) + card((P_{a_i} \cup Sy_{P_{a_i}}) \cap s_{P_{b_j}}) \quad (11)$$

$$SM_6(P_a, H_b) = \sum_{i=1}^{|P_a|}\sum_{j=1}^{|H_b|} card(s_{P_{a_i}} \cap s_{H_{b_j}}) + card(s_{P_{a_i}} \cap (H_{b_j} \cup Sy_{H_{b_j}})) + card(P_{a_i} \cup Sy_{P_{a_i}}) \cap s_{H_{b_j}}) \quad (12)$$

$$SM_7(H_a, b) = \sum_{i=1}^{|H_a|} card(s_{H_{a_i}} \cap s_b) + card(s_{H_{a_i}} \cap (b \cup Sy_b)) + card(s_b \cap (H_{a_i} \cup Sy_{H_{a_i}})) \quad (13)$$

$$SM_8(H_a, P_b) = \sum_{i=1}^{|H_a|}\sum_{j=1}^{|P_b|} card(s_{H_{a_i}} \cap s_{P_{b_j}}) + card(s_{H_{a_i}} \cap (P_{b_j} \cup Sy_{P_{b_j}})) + card((H_{a_i} \cup Sy_{H_{a_i}}) \cap s_{P_{b_j}}) \quad (14)$$

$$SM_9(H_a, H_b) = \sum_{i=1}^{|H_a|}\sum_{j=1}^{|H_b|} card(s_{H_{a_i}} \cap s_{H_{b_j}}) + card(s_{H_{a_i}} \cap (H_{b_j} \cup Sy_{H_{b_j}})) + card((H_{a_i} \cup Sy_{H_{a_i}}) \cap s_{H_{b_j}}) \quad (15)$$

Where s_a refers to the sense of a and Sy_a refers to the *synset* (set of synonyms) of a.

We applied our measure (16) on M&C benchmark dataset several times, each time with a different combination of $SM_{1 \leq i \leq 9}$ (Given two parameters $\alpha, \beta \in [0, 1]$, $[\alpha \times \sum_{i=1}^{9} SM_i = \beta \times \sum_{i=1}^{9} SM_i = \sum_{i=1}^{9} SM_i$, (Where $\alpha = \beta = 1$)], $[\alpha \times SM_1 = 0.1 \times SM_1$ and $\beta \times \sum_{i=2}^{9} SM_i = 0.9 \times \sum_{i=2}^{9} SM_i]$, $[\alpha \times SM_2 = 0.1 \times SM_2$ and $\beta \times \sum_{i=1}^{9}_{i \neq 2} SM_i = 0.9 \times \sum_{i=1}^{9} SM_i]$ etc.). We then calculated, for each combination, the correlation coefficients between the reference results in M&C's experiment [24] and our similarity values. The process of selecting the most promising combination was based on the correlation r: eliminating combinations with weak correlation ($|r| < 0.5$), and keeping combinations with strong correlation ($0.5 \leq |r| \leq 1$).

$0.8 \times (SM_1 + SM_5 + SM_9) + 0.2 \times \sum_{i=2}^{8}_{i \neq 5} SM_i$ is the combination we decided to keep because its correlation was

the highest almost every time (in the range of $0.88 - 1$). This is due to the fact that given a and b are semantically similar, they satisfy that similar relations (a with b (SM_1), hypernyms of a with hypernyms of b (SM_5), and hyponyms of a with hyponyms of b (SM_9)) are more likely to be similar than different relations (a with hypernyms of b (SM_2), a with hyponyms of b (SM_3), hypernyms of a with b (SM_4), hypernyms of a with hyponyms of b (SM_6), hyponyms of a with b (SM_7), and hyponyms of a with hypernyms of b (SM_8)). Thus, the similarity value between a and b is calculated as follows:

$$
\begin{cases}
Sim_{words}(a,b) = 1, \text{ if } a \text{ and } b \text{ are} \\
\qquad \text{synonyms or one of them is a direct hyponym} \\
\qquad \qquad \qquad \qquad \text{of the other} \\[2mm]
Sim_{words}(a,b) = 0, \text{ if } [0.8 \times (SM_1 + SM_5 + SM_9) \\
\qquad + 0.2 \times \sum_{\substack{i=2 \\ i \neq 5}}^{8} SM_i] \times \exp \frac{\sum_{\substack{i=1 \\ SM_i \neq 0}}^{9} 1}{9} \leq 1 \\[4mm]
Sim_{words}(a,b) = \cfrac{\begin{array}{l}[0.8 \times (SM_1 + SM_5 + SM_9) \\ + 0.2 \times \sum_{\substack{i=2 \\ i \neq 5}}^{8} SM_i] \times \exp \frac{\sum_{\substack{i=1 \\ SM_i \neq 0}}^{9} 1}{9} - 1\end{array}}{\begin{array}{l}[0.8 \times (SM_1 + SM_5 + SM_9) \\ + 0.2 \times \sum_{\substack{i=2 \\ i \neq 5}}^{8} SM_i] \times \exp \frac{\sum_{\substack{i=1 \\ SM_i \neq 0}}^{9} 1}{9} + 1\end{array}} \\
\qquad \qquad \qquad \qquad \qquad \qquad \qquad \text{, otherwise}
\end{cases}
$$
$$(16)$$

Step 2: Calculating the semantic similarity between sets of words. The similarity calculator uses the similarity measure between words (16) to compute the similarity between sets of words. Given two *entities* $e_1 \in S_1$ and $e_2 \in S_2$. Let $set_{e_1} = \{W_{1,1}, W_{1,2}, \ldots, W_{1,card(set_{e_1})}\}$ and $set_{e_2} = \{W_{2,1}, W_{2,2}, \ldots, W_{2,card(set_{e_2})}\}$ be their respective sets of words. The similarity calculator uses equation (17) to calculate the similarity between set_{e_1} and set_{e_2}.

$$
Sim_{sets}(set_{e_1}, set_{e_2}) = \frac{1}{min(card(set_{e_1}), card(set_{e_2}))}
$$
$$
\times \left(\sum_{i=1}^{card(set_{e_1})} max(m_{i,j})_{1 \leq j \leq card(set_{e_2})} \right) \quad (17)
$$

Where $M = (m_{i,j})_{\substack{1 \leq i \leq card(set_{e_1}) \\ 1 \leq j \leq card(set_{e_2})}}$ is the similarity matrix. Its individual items are defined as follows $m_{i,j} = Sim_{words}(W_{1_i}, W_{2_j})$.

Next, we define the matches based on the similarity values.

C. The Post-matching Module

We applied our similarity measure (17) on the semantically related sets of words from the TEL schemas. The results formed a set of similarity values, each represents the similarity between two sets. The process of selecting the threshold value was based on reference matches we defined manually in order to identify the range of similarity values generated for semantically similar sets. We noticed that most matching sets have a similarity value greater than or equal to 0.8. Hence, we defined the threshold value 0.8 beyond which the pair of *entities* must be matched.

The post-matching module consists mainly of one major component, namely the *matches generator*, which uses the

threshold value to eliminate *entity* pairs with very low similarity values, and match only pairs with high similarity values (≥ 0.8). Algorithm 5 summarizes this.

Algorithm 5 MatchesGenerator($SETS_1'$, $SETS_2'$, V)

Input:
  $SETS_1'$
  $SETS_2'$
  V
Output:
  $Matches$
1: **for** each v in V **do**
2:   **if** ($v \geq 0.8$) **then**
3:     $Matches \leftarrow Matches \cup (set_1, set_2)$
4:   **end if**
5: **end for**
6: **return** $Matches$

V. EXPERIMENTAL RESULTS

We conducted extensive experiments to evaluate xMatcher based on a real implementation. We focused on evaluating two major issues. (1) We verified the accuracy of the results of our similarity measure, by evaluating the correlation coefficient and the Mean Square Error. (2) We examined the accuracy of the matches generated by xMatcher, by evaluating *Precision*, *Recall*, *Overall*, and *F-Measure*.

A. Experimental Setup

Datasets: First, we experimented our measure on M&C dataset [24], which contains thirty word pairs (see Table IV). We then experimented xMatcher over the *Conference Track* used in OAEI 2018 and available on the Web[3]. The *Conference Track* involves 16 ontologies describing the domain of organizing academic conferences. It has been used by the research community for over 13 years. It has 21 reference alignments composed from 7 out of 16 real domain ontologies.

Implementation: In addition to our measure, we implemented four measures and distances Resnik, J&C, Lin, and Nababteh over WordNet. Then, we implemented xMatcher. Finally, since xMatcher was initially developed to take as input XML schemas and since the *Conference Track* includes ontologies, we implemented the converting process presented in [30] to transform ontologies into XML schemas.

Measures: For semantic similarity values (produced by all five measures), we used the correlation coefficient and Mean Square Error (MSE) to compare the returned results with the reference results [24]. The correlation coefficient measures how strong the relationship is between the returned values and the reference results. MSE measures the average of the squares of the errors between the returned values and the reference results. The lower the MSE is, the better.

For matching results, we used the previously published results produced by twelve ontology matching systems (SANOM [31], AML [13], LogMap [32], XMap [33], KEPLER [34], ALIN [9], DOME [11], Holontology [10], FCAMapX [35], [36], LogMapLt [32], ALOD2Vec [12], and Lily [37]) that

---

[3]http://oaei.ontologymatching.org/2018/

TABLE IV. Semantic similarity values by word pair

Word pair	M&C	Resnik	J&C	Lin	Nababteh	Our measure
Automobile / Car	0.98	0.9962	1	1	1	1
Journey / Voyage	0.96	0.9907	0.9165	0.8277	0.857335	1
Gem / Jewel	0.96	1	1	0.2434	0.31453	1
Boy / Lad	0.94	0.9971	0.8613	0.6433	1	1
Coast / Shore	0.925	0.9994	0.9567	0.96	1	1
Asylum / Madhouse	0.9025	1	0.9379	0.769	0.879	1
Magician / Wizard	0.875	0.9999	1	0.1958	0.28158	1
Midday / Noon	0.855	0.9998	1	1	1	1
Furnace / Stove	0.7775	0.6951	0.593	0.2294	0.26674	0.79
Food / Fruit	0.77	0.9689	0.7925	0.0956	0.103839	0.98
Bird / Cock	0.7625	0.9984	0.8767	0.7881	0.930014	1
Bird / Crane	0.7425	0.9984	0.815	0	0.850943	0.95
Implement / Tool	0.7375	0.9852	0.977	0.914	1	1
Brother / Monk	0.705	0.8722	0.6656	0	1	1
Crane / Implement	0.42	0.8722	0.6526	0	0.513459	0.73
Brother / Lad	0.415	0.8693	0.6775	0.24	0.29735	0.62
Car / Journey	0.29	0	0.5883	0	0	0
Monk / Oracle	0.275	0.8722	0.6203	0.1828	0.191595	0.75
Food / Rooster	0.2225	0.5036	0.5885	0.0762	0.095302	0.66
Coast / Hill	0.2175	0.9867	0.8487	0.127	0.19414	0.49
Forest / Graveyard	0.21	0	0.484	0.1119	0.1706	0.61
Monk / Slave	0.1375	0.8722	0.6962	0.2011	0.34281	0.25
Coast / Forest	0.105	0	0.5179	0	0	0.2
Lad / Wizard	0.105	0.8722	0.6905	0.2241	0.34155	0.2
Cord / Smile	0.0325	0.8044	0.5845	0	0	0.11
Glass / Magician	0.0275	0.5036	0.5699	0.0663	0.09335	0
Rooster / Voyage	0.02	0	0.4168	0	0	0
Noon / String	0.02	0	0.4329	0	0	0

participated in OAEI 2018 over the *Conference Track*. We used *Precision* (18), *Recall* (19), *Overall* (20), and *F − Measure* (21) [38] to evaluate the returned matches based on nine combinations of evaluation variants with crisp reference alignments: *ra1-M1*, *ra1-M2*, *ra1-M3*, *ra2-M1*, *ra2-M2*, *ra2-M3*, *rar2-M1*, *rar2-M2*, and *rar2-M3* (*ra1* is the original reference alignment; *ra2* is an extension of ra1; and *rar2* is an updated version of *ra2* that deals with violations of conservativity). *ra1-M1*, *ra2-M1*, and *rar2-M1* are used to evaluate only alignments between classes; *ra1-M2*, *ra2-M2*, and *rar2-M2* are used to evaluate only alignments between properties; and *ra1-M3*, *ra2-M3*, and *rar2-M3* are used to evaluate both alignments between classes and properties.

$$Precision = \frac{Correct\ Matches}{Correct\ Matches + Incorrect\ Matches} \qquad (18)$$

(18) is the probability of correct matches among the matches returned by a matching system.

$$Recall = \frac{Correct\ Matches}{Missed\ Matches + Correct\ Matches} \qquad (19)$$

(19) is the probability of correct matches returned by a matching system among the reference matches.

$$Overall = Recall \times (2 - \frac{1}{Precision}) \qquad (20)$$

(20) quantifies the amount of manual post-effort necessary to remove false matches and add missed matches.

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (21)$$

(21) is the harmonic mean of *Precision* and *Recall*.

B. Results and Discussion

1) Semantic Similarity Measure: Experiment Results: We first applied our measure, Resnik, J&C, Lin, and Nababteh on M&C dataset (see the results in Table IV). We then used the results to calculate the correlation coefficient and MSE (see the overall results in Table V and the details about correlations in Fig. 3(a), Fig. 3(b), Fig. 3(c), Fig. 3(d), and Fig. 3(e)).

TABLE V. Comparison between some state of the art similarity measures and our measure

Measure	Correlation coefficient	MSE
Resnik	0.6671	0.1373
J&C	0.8363	0.1018
Lin	0.6852	0.1188
Nababteh	0.7654	0.0699
Our measure	0.9102	0.0453

The findings indicate a strong positive correlation ($+0.9102$) between our measure and the reference results. They also indicate that our measure obtained the smallest MSE (0.0453) compared to the other measures. Thus, our measure outperforms the state of the art measures, showing that on the one hand information content-based measures cannot provide high accuracy results; on the other hand combining different WordNet information (hypernyms, direct hyponyms, senses, and *synsets*) is a good plus to obtain high accuracy results.

Fig. 3. Regression lines for (a) Resnik vs. M&C; (b) J&C vs. M&C; (c) Lin vs. M&C; (d) Nababteh vs. M&C; and (e) Our measure vs. M&C

Fig. 4. Accuracy of the Matches

2) xMatcher: Experiment Results: We first generated the matches using xMatcher. We then calculated, for all matches for which there is a reference alignment, *Precision*, *Recall*, *Overall*, and *F-Measure* nine times, each time with a different reference alignment. Fig. 4(a), Fig. 4(b), Fig. 4(c), Fig. 4(d), Fig. 4(e), Fig. 4(f), Fig. 4(g), Fig. 4(h), and Fig. 4(i) present the new and previously published results.

On the one hand, the previously published results indicate noticeable changes in *Precision*, *Recall*, *Overall*, and *F-Measure*: overall, they achieved good matching accuracy when evaluated based on *ra1-M1*, *ra1-M3*, *ra2-M1*, *ra2-M3*, *rar2-M1*, and *rar2-M3*; and low accuracy even null sometimes (Lily and ALIN) with *ra1-M2*, *ra2-M2*, and *rar2-M2*. On the other hand, xMatcher obtained high accuracy matches, outperforming all systems almost every time except from *ra1-M2* and *ra2-M2* where AML surpassed it slightly ($Precision = 1$).

While xMatcher matches both classes and properties, Lily and ALIN match only classes the reason why they failed to produce high accuracy matches with *ra1-M2*, *ra2-M2*, and *rar2-M2*; SANOM, AML, LogMap, and XMap match some but not all properties which explain their negative *Overall* with *ra1-M2*, *ra2-M2*, and *rar2-M2*; and KEPLER, DOME, Holontology, FCAMapX, LogMapLt, and ALOD2Vec match very few properties which justify their negative *Overall* and low *Precision*, *Recall*, and *F-Measure* with *ra1-M2*, *ra2-M2*, and *rar2-M2*. We can conclude that (1) SANOM, AML, LogMap, XMap, KEPLER, ALIN, DOME, Holontology, FCAMapX, LogMapLt, ALOD2Vec, and Lily work well with the reference alignments that consider classes or both classes and properties. However, they fail to match correctly with the reference alignments that consider only properties; and (2) xMatcher succeeds to achieve superior accuracy matches regardless of the reference alignment it is compared to.

Overall, xMatcher obtained the highest accuracy matches (see Fig. 4.j which displays the average matching accuracy): $Precision = 0.89$ suggests that most matches are correct;

$Recall = 0.66$ suggests that xMatcher missed only few matches; and $Overall = 0.57$ implies that xMatcher needs only a small amount of manual post-effort to correct the results.

To prove scalability of xMatcher (note that due to space limitation, we do not display the results in figures in this paper), we applied xMatcher on more datasets, for instance the TEL (Travel, Entertainment and Living) datasets which contain five different datasets that are publicly available on the Web. The Travel group includes two various domains: *Car Rentals* and *Airfare*; the Entertainment group contains two different domains as well: *Movies* and *Books*; and, the Living group involves mainly one single domain: *Jobs*. The results show once again the capability of xMatcher to reach a high matching accuracy, which proves that xMatcher is scalable.

VI. CONCLUSION

We have demonstrated that the use of WordNet combined with our semantic similarity measure is an effective way to capture semantic correspondences in XML schemas. Current matching systems are error-prone and human-dependent. Thus, we have developed xMatcher, an approach to automatically match XML schemas and provide accurate matches.

Given two XML schemas S_1 and S_2, our main idea is to first generate sets of words from S_1 and S_2, then determine

semantically related sets, and finally identify semantic correspondences between related sets. We evaluated xMatcher over the *Conference Track*. The results show that xMatcher achieves better accuracy than twelve state of the art matching systems. Future research includes the following:

- **Improving the accuracy of the matches.** An interesting direction is to achieve better correlation, MSE, *Precision*, *Recall*, *Overall*, and *F-Measure*.
- **Considering other matching quality factors.** In this paper, we focused on achieving high matching accuracy. A future direction is to propose techniques that consider other quality factors.
- **Matching other data representations.** xMatcher takes as input XML schemas. An interesting direction is to match different data representations.

REFERENCES

[1] C. Zhang, L. Chen, H. Jagadish, M. Zhang, and Y. Tong, "Reducing uncertainty of schema matching via crowdsourcing with accuracy rates," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[2] Y. Lee, M. Sayyadian, A. Doan, and A. S. Rosenthal, "etuner: tuning schema matching software using synthetic scenarios," *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 16, no. 1, pp. 97–122, 2007.

[3] L. Otero-Cerdeira, F. J. Rodríguez-Martínez, and A. Gómez-Rodríguez, "Ontology matching: A literature review," *Expert Systems with Applications*, vol. 42, no. 2, pp. 949–971, 2015.

[4] F. Ardjani, D. Bouchiha, and M. Malki, "Ontology-alignment techniques: survey and analysis," *International Journal of Modern Education and Computer Science*, vol. 7, no. 11, p. 67, 2015.

[5] L. Mukkala, J. Arvo, T. Lehtonen, T. Knuutila, *et al.*, "Current state of ontology matching. a survey of ontology and schema matching," 2015.

[6] S. Anam, Y. S. Kim, B. H. Kang, and Q. Liu, "Review of ontology matching approaches and challenges," *International journal of Computer Science and Network Solutions*, vol. 3, no. 3, pp. 1–27, 2015.

[7] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[8] D. Faria, C. Pesquita, B. S. Balasubramani, T. Tervo, D. Carriço, R. Garrilha, F. M. Couto, and I. F. Cruz, "Results of aml participation in oaei 2018.," in *OM@ ISWC*, pp. 125–131, 2018.

[9] J. da Silva, K. Revoredo, and F. A. Baião, "ALIN results for OAEI 2018," in *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018.*, pp. 117–124, 2018.

[10] P. Roussille, I. Megdiche Bousarsar, O. Teste, and C. Trojahn, "Holontology: results of the 2018 oaei evaluation campaign," CEUR-WS: Workshop proceedings, 2018.

[11] S. Hertling and H. Paulheim, "Dome results for oaei 2018.," in *OM@ ISWC*, pp. 144–151, 2018.

[12] J. Portisch and H. Paulheim, "Alod2vec matcher.," in *OM@ ISWC*, pp. 132–137, 2018.

[13] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, and F. M. Couto, "The agreementmakerlight ontology matching system," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pp. 527–541, Springer, 2013.

[14] I. F. Cruz, F. P. Antonelli, and C. Stroe, "Agreementmaker: efficient matching for large real-world schemas and ontologies," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1586–1589, 2009.

[15] L. Meng, R. Huang, and J. Gu, "A review of semantic similarity measures in wordnet," *International Journal of Hybrid Information Technology*, vol. 6, no. 1, pp. 1–12, 2013.

[16] B. Poorna and A. S. Ramkumar, "Semantic similarity measures: an overview and comparison," *International Journal of Advanced Research in Computer Science*, vol. 9, no. Special Issue 1, p. 100, 2018.

[17] M. H. El Yazidi, A. Zellou, and A. Idri, "Towards a fuzzy mapping for mediation systems," in *2012 IEEE International Conference on Complex Systems (ICCS)*, pp. 1–4, IEEE, 2012.

[18] A. Gupta, A. Kumar, J. Gautam, A. Gupta, M. A. Kumar, and J. Gautam, "A survey on semantic similarity measures," *IJIRST-International Journal for Innovative Research in Science & Technology*, vol. 3, p. 12, 2017.

[19] A. Yousfi, M. H. Elyazidi, and A. Zellou, "Assessing the performance of a new semantic similarity measure designed for schema matching for mediation systems," in *International Conference on Computational Collective Intelligence*, pp. 64–74, Springer, 2018.

[20] A. M. Abdelrahman and A. Kayed, "A survey on semantic similarity measures between concepts in health domain," *American Journal of Computational Mathematics*, vol. 5, no. 02, p. 204, 2015.

[21] M. H. E. Yazidi, A. Zellou, and A. Idri, "FMAMS: fuzzy mapping approach for mediation systems," *Int. J. Appl. Evol. Comput.*, vol. 4, no. 3, pp. 34–46, 2013.

[22] M. H. E. Yazidi, A. Zellou, and A. Idri, "Mapping in GAV context," in *10th International Conference on Intelligent Systems: Theories and Applications, SITA 2015, Rabat, Morocco, October 20-21, 2015*, pp. 1–5, 2015.

[23] F. Couto and A. Lamurias, "Semantic similarity definition," *Encyclopedia of bioinformatics and computational biology*, vol. 1, 2019.

[24] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *arXiv preprint cmp-lg/9511007*, 1995.

[25] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.

[26] D. Lin, "Principle-based parsing without overgeneration," in *31st annual meeting of the association for computational linguistics*, pp. 112–120, 1993.

[27] N. Mohammed and D. Mohammed, "New modified semantic similarity measure based on information content approach," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 17, no. 3, p. 73, 2017.

[28] A. Tversky, "Features of similarity.," *Psychological review*, vol. 84, no. 4, p. 327, 1977.

[29] Z. Zhou, Y. Wang, and J. Gu, "New model of semantic similarity measuring in wordnet," in *2008 3rd International Conference on Intelligent System and Knowledge Engineering*, vol. 1, pp. 256–261, IEEE, 2008.

[30] L. Mukkala, J. Arvo, T. Lehtonen, and T. Knuutila, "Trc-matcher and enhanced trc-matcher. new tools for automatic xml schema matching," 2017.

[31] M. Mohammadi, W. Hofman, and Y. Tan, "SANOM results for OAEI 2018," in *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018.*, pp. 205–209, 2018.

[32] E. Jiménez-Ruiz, B. C. Grau, and V. Cross, "Logmap family participation in the OAEI 2018," in *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018.*, pp. 187–191, 2018.

[33] W. E. Djeddi, S. B. Yahia, and M. T. Khadir, "Xmap: results for OAEI 2018," in *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018.*, pp. 210–215, 2018.

[34] M. Kachroudi, G. Diallo, and S. B. Yahia, "KEPLER at OAEI 2018," in *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018.*, pp. 173–178, 2018.

[35] M. Zhao and S. Zhang, "Fca-map results for oaei 2016.," in *OM@ ISWC*, pp. 172–177, 2016.

[36] G. Chen and S. Zhang, "Fcamapx results for OAEI 2018," in *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018.*, pp. 160–166, 2018.

[37] Y. Tang, P. Wang, Z. Pan, and H. Liu, "Lily results for OAEI 2018," in *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018.*, pp. 179–186, 2018.

[38] I. Kastner and F. Adriaans, "Linguistic constraints on statistical word segmentation: The role of consonants in arabic and english," *Cognitive science*, vol. 42, pp. 494–518, 2018.

Performance Analysis of a Graph-Theoretic Load Balancing Method for Data Centers

Walaa M. AlShammari[1], Mohammed J.F. Alenazi[2]

College of Computer and Information Sciences,
Department of Computer Engineering,
King Saud University, Riyadh, Saudi Arabia

Abstract—Modern data centers can process a massive amount of data in a short time with minimal errors. Data center networks (DCNs) use equal-cost, multi-path topologies to deliver split flows across alternative paths between the core layer and hosted servers, which could lead to significant overload if path scheduling is inefficient. Thus, distributing incoming requests among these paths is crucial for providing higher throughput and protection against link or switch failures. Several approaches have been proposed for path selection, mainly relying on round-robin and least-congested methods. In this paper, we propose a load-balancing method based on betweenness centrality to improve the overall performance of a data center in terms of throughput, delay, and energy consumption. For evaluation, we compare our method with baseline methods of different DCN topologies: fat-tree, DCell, and BCube. On average, the evaluation results show that our method outperforms the others. It increases throughput by 202% and 33% while reducing delay by 20% and 22%, and energy consumption by 40% and 41% compared to the round-robin and least-congested methods, respectively.

Keywords—*Data center; load balancing; path diversity; network management; load management; throughput; topology; performance metrics; betweenness centrality; flow scheduling; modeling; DCNs*

Fig. 1. Three-Tier Data Center Architecture

I. INTRODUCTION

Nowadays, studies in computer networking focus on data center networks (DCNs) and challenges involved in scheduling the paths of these networks. A DCN is a construction that links a large number of servers, switches, and routers to connected devices. Data center switches are designed to forward data between endpoints, while servers process the data [1], [2], [3]. The importance of these data centers is increasing, as a considerable number of networks are now linked. Moreover, the main functions of data centers are the analysis, processing, and storage of large data.

DCNs have two types of architecture: two- and three-tier architectures. The most commonly used architecture in current DCNs is the three-tier architecture, comprising a core layer, an aggregate layer, and an access layer, from top to bottom, as shown in Fig. 1 [3], [2], [4]. When DCNs process applications' requests, the requests first arrive at the core layer. Then, the core layer forwards the requests to the destination server across multiple paths through the aggregate and access layers. Computations by servers complete applications' requests; thus, massive requests require high-performance computing servers [5], [6].

Data center traffic can be classified into two main types: mice and elephant. Mice traffic makes up a data flow of small sizes, which is sensitive to delays (time-sensitive). An example of a mice flow is web-searching. Elephant traffic is defined as a flow that consumes more channel bandwidth. Thus, it is data-sensitive; for example, when downloading or uploading files [3], [1], [7]. Dealing with data center traffic causes a major issue known as congestion. Congestion occurs when a traffic imbalance ensues between network paths, leading to significant overload if path scheduling is inefficient—for instance, using a static approach to scheduling requests, e.g., equal-cost multipath (ECMP) or round-robin, in which requests are sequentially assigned to paths, regardless of the paths' status. Thus, some paths are overloaded, while multiple other paths are available, affecting overall DCN performance [8], [9]. Therefore, an efficient load balance of requests across multiple links (paths) must be considered to achieve the best possible data center performance. A link-based load balancing method, in general, aims to distribute requests across multiple links to avoid congestion in a single path and guide traffic to the best available destination (path). However, this task is currently one of the major challenges facing data centers [10], [11], [12], [2].

In this paper, we propose a load balancing method that utilizes a graph-theoretic betweenness centrality (C_B) metric, which attempts to forward new traffic to the least congested paths within DCN topologies. For evaluation, we compare our proposed method with baseline methods: round-robin and least-congested. We then present a comparative performance analysis between the above load-balancing algorithms and the proposed method in commonly used DCN topologies, fat-tree, BCube, and DCell, using NetworkX, a Python library for graphics and networking.

The remainder of this paper is organized as follows. In

Section II, the necessary background for the proposed method is covered, including DCN topologies, graph theory, and load balancing algorithms. Section II-B introduces the proposed load-balancing method. Section IV discusses the dataset and performance metrics. Section V presents the implementation and modeling results. Finally, Section VI presents conclusions and suggestions for future work.

II. BACKGROUND AND RELATED WORK

In this section, we present a graph-theoretic background based on centrality. Furthermore, we show the most popular DCN topologies. All topologies presented were implemented and evaluated in our study. Moreover, related studies are presented in this context.

A. Graph-Theoretic Centrality Metrics

Several metrics have been introduced to measure the centrality of a node or link in a graph: to measure the importance of a vertex or an edge [13], [14]. *Degree centrality* (C_D), the degree of a node in a graph refers to the number of links attached to that node [15], [14]. A node with a high degree of centrality has a critical position in a network, since most of the links pass through it. *Closeness centrality* (C_C), is defined as the sum of the shortest paths from one node to all other nodes in a graph. *Betweenness centrality* (C_B) is a metric that is used for nodes and edges. The C_B of a node is defined as the number of shortest paths passing through that node in a graph, while the C_B of a link refers to the number of shortest paths passing through that link in a graph [16], [14]. C_B is a vital metric since the C_B value of a link or node is changed based on graph structures. The function used to calculate the C_B of a link l in networks is defined as follows:

$$C_B(l) = \sum_{u,z \in V, l \in L} \frac{|Fu, l, z|}{|Fu, z|} \tag{1}$$

where the V refers to the number of vertices, such as source or destination vertices in a graph, and L is a set of links in a graph.

set of shortest paths from a vertex u to a vertex z, passing through an edge l.

$|Fu, l, z| \Rightarrow$ set of shortest paths from a vertex u to a vertex z, passing through an edge l.

$|Fu, z| \Rightarrow$ set of all shortest paths between vertices u and z.

B. Data Center Network Architectures

DCN topologies can be classified into three categories: switch-centric, server-centric, and hybrid structure. In this section, we present examples of switch-centric and hybrid architectures.

1) Fat-tree topology: Fat-tree topology is an example of switch-centric architecture that has only switches for communication and computing tasks. It is the most widely used topology in DCNs due to a full mesh connection. Fat-tree, as illustrated in Fig. 2b, comprises three layers: a core layer, an aggregate layer, and an edge layer [17]. Servers at the bottom of a graph are directly attached to switches in the edge layer. Each edge switch comprises an n-port, which connects an n/2 server. The rest of the ports are linked to the aggregate switches. Fat-tree topology offers multipath routing between any peer hosts, which reduces the likelihood of link or switch failures. The total number of links in the fat-tree topology is $\frac{3 \times n^3}{n}$ link, where n is the number of switches [11].

2) BCube and DCell topologies: These are hybrid structure topologies that use both switches and servers for data forwarding and computing functions. They have different architectures, but both have recursively defined structures [11], [17].

$BCube_0$ comprises n count servers, where each server has a direct link to a switch at each level ($BCube_0$ and $BCube_1$ levels) through a port. Thus, servers cannot directly communicate with each other because no direct link exists between them. Moreover, as shown in Fig. 2a, switches at the $BCube_1$ level can communicate with switches at the $BCube_0$ level through an attached server, where BCube topology can construct up to three levels (k = 3). Furthermore, the number of complete links in the BCube topology is $2 * n^2$ link, where n is the number of switches [18], [11].

Moreover, the DCell architecture, as illustrated in Fig. 2c, comprises servers, mini-switches, and links. Each mini-switch comprises n servers, which are connected through a link. In addition, servers from different DCells are directly connected through connection links, and mini-switches from different DCells can communicate through an attached server. The total number of links in the DCell topology is $\frac{3n \times (n+1)}{2}$ links, where n is the number of switches [11].

C. Path Selection Methods

In this section, we discuss the difference between the two kinds of path scheduling algorithms: static and dynamic.

In static scheduling, traffic scheduling decisions are applied at the configuration stage, and decisions taken at this stage are independent of subsequent network statuses. A round-robin algorithm is an example of a static scheduling algorithm, which distributes requests to specific servers through sequential paths. Thus, the capacity of the server or link is not considered [11], [19]. In contrast, dynamic scheduling algorithms dynamically distribute flows through optimum links by considering the current state of a network. The 'least congested' algorithm is an example of a dynamic scheduling algorithm, which forwards a new request to the least loaded or congested link [11], [20], [21], [22].

D. Related Work

Several studies have proposed approaches to improve DCN performance. In this section, we investigate related approaches.

Author in [23] used Luopan, a congestion-aware load-balancing method, to distribute flowcells (sub-flows after

BCube [1]

(a) BCube

(b) Fat-tree

(c) DCell

Fig. 2. Popular Topologies of Data Center Architectures

breaking an inserted flow) into a few available paths based on sampling. The idea is to take a few paths and distribute flowcells, relying on the least congested paths. Results show that this method outperforms the Presto algorithm, which improves the flow completion time for mice and elephant flows by 35% and 30%, respectively. However, their proposed method does not use a C_B as a metric for flow scheduling.

Zakia et al. [24] proposed a dynamic SDN-based path selection method to improve DCN load balancing. They used the Dijkstra algorithm for calculating all the shortest paths between two nodes. Then, a new flow will be assigned to the path with the least congestion and minimum cost. However, if the path has been overloaded, the path's flow will be forwarded to an alternative least-congested and lowest-cost path. However, C_B is not used in their path selection method.

Alenezi et al. [25] proposed a comprehensive comparison technique to study DCN robustness in the event of a single point of failure. They addressed this issue by increasing the number of links to balance the edge-based BC yields. Thus, they found that utilizing the BC metric and increasing the number of links improves the robustness of a network upon attacks. However, their approach lacks path scheduling, since

they used C_B to increase network robustness.

Author in [26] proposed a load-balancing method that balances the workload in SDN-based DCNs. Their method monitors bandwidth utilization and the rate of packet loss before congestion occurs. When the bandwidth utilization and packet loss rate exceed a specified threshold, flows are rerouted from congested paths to alternative, less congested paths. However, if some metrics, such as latency and throughput, reach a specified threshold, a controller will set new flow rules to distribute flows among multiple paths. However, their method uses a least-congested metric to dynamically schedule the flow and they do not use the C_B metric to schedule the flow.

Shafiee et al. [27] applied a congestion-aware load-balancing method to distribute flows to paths with minimum cost, which are paths that accumulate low cost; the estimated cost is the sum of convex functions of link utilization. Moreover, assigning a flow to paths does not rely on splitting the flow into small flowlets. Thus, they reduced the difficulties associated with TCP packet reordering. Simulation results show that their method reduced the overall cost in fat-tree and JellyFish DCNs. Still, C_B is not used in their dynamic

approach for flow scheduling.

Challa et al. [28] introduced a routing method based on software-defined networking (SDN), called CentFlow. Their method utilizes the degree centrality metric to detect the node or link with the highest utilization value and avoids using that node (link) to minimize cost, since the degree of utilization affects a node's cost in a network. Thus, CentFlow selects the node or link with the lowest cost based on the Dijkstra algorithm. However, C_B does not exist in its routing method.

To briefly summarize the above-mentioned related works, we noticed that the procedure described in [24], selects the least congested path based on the shortest-path algorithm, while that in [26], considers the least-congested path and packet loss rate metric to select an optimum path. Moreover, [27], selected a path based on the least-congested metric, while [23] followed the same approach but with partial flow into multiple flowcells, which were then distributed to the least congested paths. Furthermore, [25], uses the C_B metric as an indicator to increase the number of links and network robustness against attacks, while [28], utilizes the C_B and the degree of centrality to select the optimum paths from the available shortest paths.

The approaches discussed above use only static or dynamic approaches in scheduling flows among paths. In our method, we utilize both C_B and least-congested metrics to select the optimum path for a flow from diverse paths. We evaluate and compare the performance of our method with other load-balancing methods in widely adopted DCNs.

III. A Load-Balancing Method

In this section, we present our proposed load-balancing method, which utilizes C_B and least-congested metrics to select paths between the core and edge layers.

We propose a C_B-based dynamic load balancing method that performs load balancing in a DCN topology by measuring the minimum C_B of all links as well as the least congested path at a particular time. The method obtains all simple paths between a single source and a single destination node through which flows can be sent. Traffic flows are of the same priority and size. Among the selected paths, the path with the least C_B and congestion is selected, and traffic flow is forwarded along that path. The performance of the proposed method is evaluated in a state-of-the-art DCN topology by collecting data from links and switches. Fig. 3 illustrates a flow diagram of our load-balancing method.

To demonstrate our proposed method, we apply it to a simple topology, as shown in Fig. 4. The graph has two, three, and four nodes from top to bottom and has a link of 1 Gbps capacity connecting the nodes. For instance, host1 sends data to host2 through multiple paths. We evaluate, compute, and assign a C_B value to each edge using NetworkX, as shown in Fig. 4.

Suppose we have three paths from node 0 to node 6, where each path has a number of files; for example, path 1 ([0, 1, 3, 2, 6]) has one file, path2 ([0, 1, 4, 2, 6]) has three files, and path3 ([0, 2, 6]) has one file. The question arises as to which paths the new files are to be assigned. Our proposed method is based on a four-step process. First, the C_B values of all edges in the

Fig. 3. The Proposed Load-Balancing Method.

Fig. 4. A DCN Topology to Illustrate the Proposed Load-Balancing Method. Values that Appear at Each Edge in the Figure are the Betweenness Centrality.

paths are summed. Second, the C_B values of each path are normalized by dividing each value by the maximum C_B value. Then, the congestion at each path is normalized: the number of files in each path is divided by the maximum number of files. Table I illustrates the C_B and normalized values. Finally, each path is assigned a value equal to the sum of the normalized C_B and congestion values, and the path with the lowest value is selected. In the example shown in Fig. 4, we see that path 3 ([0, 2, 6]) has the lowest value. Therefore, the new file will be forwarded to this path.

IV. Evaluation

This section discusses our evaluation and presents the dataset of the DCN topologies and performance metrics used.

TABLE I. BETWEENNESS AND NORMALIZATION VALUES BASED ON THE
EXAMPLE IN FIG.4

Paths	Betweenness Values	Normalization
[0,1,3,2,6]	16.99	1
[0,1,4,2,6]	16.99	1
[0,2,6]	11	0.65

TABLE II. DATASET USED IN THE EVALUATION OF THE PROPOSED WORK.

Fat-tree DCN	Number of Pods	4
	Number of nodes	20
BCube DCN k-level	1	
DCell DCN level	1	
Link bandwidth	1 Gbps	
Requests number	200	
Packet size	1500 KB	

Thereafter, the evaluation results are presented and discussed.

A. Dataset

The DCN topologies mentioned in Section II are implemented and evaluated. The dataset used to evaluate the proposed method against the other load balancing methods is shown in Table II. To evaluate the fat-tree topology, we use a network of four core switches, eight aggregation switches, and edge switches. Each core switch is connected to an edge switch through aggregation switches via 1 Gbps links.

The BCube topology is implemented for level one (k = 1) with four $BCube_1$ and $BCube_0$ switches. Each $BCube_0$ switch connects to four servers, making up a total of 16 servers. Each link used to connect the layers has a capacity of 1 Gbps.

The DCell topology is implemented for level 1 and comprises five switches, each connecting to four servers through a direct link of 1 Gbps. Each server is connected to other servers in different $DCell_0$ through a 1 Gbps link.

B. Performance Metrics

In this section, we present three metrics for measuring the impact of the three load-balancing methods on the performance of DCN topologies.

1) End-to-end throughput: To calculate the throughput of a DCN topology from a single-core source to a server destination, we calculate the total time for all files to be delivered, while considering bottleneck links. Completion time is the time taken to complete a specific task [29]. Then, the total size of all the files is divided by the total completion time. The approximate completion time for each path is calculated as follows:

$$(F - 1) \times (T_d + P_d) \tag{2}$$

where F is the total number of files, Td is the transmission delay, and Pd is the propagation delay. To simplify the calculations for when congestion occurs, we assume that the files of the relevant paths are sent last. This is because we are only interested in the maximum completion time.

2) End-to-end delay: The following equation is used to calculate the end-to-end delay:

$$TotalDelay = T_d + P_d + U_d + Q_d \tag{3}$$

where T_d is transmission delay, which is the required time to transmit a packet in a channel; P_d is propagation delay, which is the time taken to forward packets across the channel; U_d is a processing delay, which refers to the time taken to process a task; and Q_d is a queuing delay, which is the packet's waiting time in a buffer [30]. In our measurements, U_d is equal to zero, and T_d can be calculated using $\frac{Request\ Size}{Link\ Capacity}$, where 1.5 GB is the request size and 1 Gbps is the link capacity. P_d can be calculated using $\frac{Channel\ Length}{Propagation\ Speed}$, where the propagation speed is the speed of light ($2 \times 10^8\ m/s$), and the link length is 10km. To calculate the queuing delay, we focus on calculating the average queuing delay for all links. Thus, we first find the total number of sent files (F) and subtract them one by one, starting with the first file, which has no waiting time. Then, we multiply the total number of files minus one by (T_d). Finally, we divide the calculated value by two, which affords us the average Q_d.

3) Energy consumption: Energy consumption is one of the biggest challenges faced by DCNs. Significant effort has been made to reduce the power consumed by data center components, while ensuring high network performance [11]. Thus, the use of the proposed load-balancing method for selecting an optimum path for sending packets is an appropriate way to reduce energy consumption [6]. Thus, a large workload in links increases energy consumption in a DCN [31], [32]. In our evaluation, we aimed to estimate energy consumption by considering the number of switches in paths, regardless of any other components in the DCN. This is because switches consume more energy than servers [11].

V. RESULTS AND DISCUSSIONS

In this section, we present and discuss the results of the three load-balancing methods applied to the three data center topologies, based on the experimental setup presented in Section IV. Our evaluation is based on three performance metrics: throughput, end-to-end delay, and energy consumption.

A. Throughput Results

The throughput results of the application of the three load-balancing methods to the three DCN topologies are shown in Fig. 5.

The throughput results in the fat-tree topology, as displayed in Fig. 5a show that our method has the best results compared to the other load-balancing methods. We notice that, for request number 25, the round-robin decreased the throughput by approximately 69% and 76% compared to the least-congested and our method, respectively. In contrast, we see that our scheme increased the total throughput by 3% and 65%, compared to the least-congested and round-robin algorithms, respectively. Furthermore, we observed that the round-robin has the worst results, achieving the lowest throughput compared to the other methods because it assigns flows to paths regardless of the paths' load status, reducing the throughput of the DCN.

Moreover, the throughput results in the BCube topology, as demonstrated in Fig. 5b, validated that our method increases throughput by approximately 92% and 296% compared to the least congested and round-robin algorithms, respectively. It is followed by the least-congested algorithm, with a throughput 106% higher than the round-robin, which achieves the worst throughput results due to its static approach.

The throughput results in the DCell topology are shown in Fig. 5c. Fig. 5c shows that our method has the highest throughput compared to the other methods. We notice that, at request number 25, the round-robin achieves a throughput of approximately 60% and 68% lower compared to the least-congested method and our own, respectively. Conversely, our method increases the total throughput by 3% and 247% compared to the least-congested and round-robin, respectively. On the other hand, we observe that the round-robin algorithm has the worst throughput results because it does not consider the load status of paths in each case, affecting the overall throughput results.

By studying all the throughput results in different DCN topologies, we can see that our proposed method outperforms the other load-balancing algorithms. This is mainly because our method selects a high-bandwidth path, while avoiding the highly congested paths and bottleneck links. On the other hand, the round-robin algorithm achieves the poorest throughput results due to its static scheduling approach which assigns flows regardless of paths' load status. The round-robin algorithm reduces the throughput results by 40%, 75%, and 71% in the fat-tree, BCube, and DCell topologies, respectively, compared to our method. Moreover, it reduces the throughput results by 38%, 52%, and 70%, respectively, compared to the least-congested algorithm.

B. End-To-End Delay Results

End-to-end delay results of the three load balancing methods in the three DCN topologies are shown in Fig. 6.

The delay results in the fat-tree topology are shown in Fig. 6a. We observe that our scheme reduces the delay by 11% compared to the least-congested and round-robin algorithms. Thus, it concisely achieves the best results, because our load-balancing method avoids congested bottleneck links. We observe that the least-congested and round-robin methods produce the same results at the end of requests. However, the round-robin algorithm incurs the highest delay at the start of requests until request number 175, which is a negative sign of load balance.

Next, Fig. 6b shows the delay results in the BCube topology. We can distinguish that our method outperforms the other load-balancing methods in terms of delay time in the BCube topology. It reduces the delay by approximately 51% and 30% compared to the least-congested and round-robin methods, respectively. Moreover, we observe that the least-congested method has the highest delay because it uses bottleneck links despite the presence of congestion.

The delay results in the DCell topology are demonstrated in Fig. 6c, which shows that our load-balancing method results in has the lowest delay compared to the other methods. Our method reduces overall delay by approximately 5% and 18%

(a) Fat-tree

(b) BCube

(c) DCell

Fig. 5. Throughput Analysis for Data Center Topologies

(a) Fat-tree

(b) BCube

(c) DCell

Fig. 6. End-To-End Delay Analysis for Data Center Topologies

(a) Fat-tree

(b) BCube

(c) DCell

Fig. 7. Energy Consumption Analysis for Data Center Topologies

compared to the least-congested and round-robin methods, respectively, from smaller to larger requests. Furthermore, the round-robin algorithm incurs the worst delay results, while increasing the number of requests because it distributes flows regardless of paths' load status.

We can conclude that our method has better delay results than the other load-balancing methods studied in several DCN topologies. This is primarily because our method bypasses highly congested links with the help of the BC value of each path, which, as mentioned earlier, directly affects network

latency. Moreover, we note that the round-robin method provides the worst results in both fat-tree and DCell topologies because it forwards requests regardless of paths' availability. As a result, the round-robin algorithm increased the network delay by 13%, 44%, and 21% in the fat-tree, BCube, and DCell DCNs, respectively, compared to our method. It increased the delay by approximately 16% in the DCell topology, while it has the same delay result in the fat-tree topology, and in the BCube topology, about 41% higher delay than the least-congested method.

C. Energy Consumption Results

Energy consumption results from the three load balancing methods in the evaluated DCN topologies are displayed in Fig. 7.

Fig. 7a shows the energy consumption results in the fat-tree topology. We observe that, in the fat-tree topology, our method affords the optimum results, while the round-robin and least-congested methods achieve the same results. Moreover, both the round-robin and least-congested methods consume up to 1000 watts, while our method consumes only 600 watts. As a result, our method reduces energy consumption by 40% in the fat-tree topology compared to the other algorithms.

Furthermore, the energy consumption results in the BCube topology are shown in Fig. 7b. We notice that the round-robin algorithm energy consumption results are approximately 12% lower than those of the least-congested algorithm and 76% higher than those of our method. Thus, we notice that the least-congested method consumes more energy than the other methods because it uses specific links with the maximum number of hops. On the other hand, the round-robin algorithm assigns flows in sequential order, regardless of the number of switches in links, while our method uses links with fewer hops because these links have the lowest C_B values. Thus, our method achieves the lowest energy consumption compared to other methods.

The energy consumption results in the DCell DCN, as illustrated in Fig. 7c, show that our load-balancing method has the best results, consuming approximately 1100 watts. It is followed by the least-congested method and then the round-robin method, which consume approximately 1650 watts and 1755 watts, respectively. As a result, we see that our method reduces energy consumption by approximately 33% and 38% compared to the least-congested and round-robin algorithms, respectively.

Our proposed method consistently improves energy consumption in DCNs, because it uses paths with the lowest C_B values, which are the paths with a low number of hops since using paths that have a large number of hops will increase energy consumption. On the other hand, the least-congested method achieves the worst energy consumption results in the BCube DCN, because it forwards flows to the least congested paths, regardless of the number of hops in those paths. In addition, the round-robin consumes more energy in the DCell topology, increasing the energy consumption by 46% and 6% compared to our method and the least-congested method, respectively, in the DCell topology. Moreover, we notice that the least-congested method and round-robin have the same results in the fat-tree topology, both consuming more energy than our method.

To summarize these results, we studied and analyzed all the performance results in the different DCN topologies investigated. We found that our load-balancing method outperforms the other methods. On average, the throughput results show that our method achieves a total throughput that is 202% and 33% higher compared to the round-robin and least-congested algorithms, respectively. Moreover, it reduces network delay by 20% and 22%, and energy consumption by 40% and 41% compared to the round-robin and least-congested methods, respectively. This is because our method bypasses bottleneck links, while considering the lowest edge C_B values. On the other hand, we notice that the round-robin algorithm has the worst throughput results, because it assigns flows sequentially, regardless of the paths' load status. We also observed that the least-congested algorithm has the worst results in delay and energy consumption, since it assigns a flow to a path despite the bottleneck links and the number of hops in that path.

VI. CONCLUSION

Improving the performance of data center systems would allow DCNs to be scalable and more customizable, helping to connect more devices around the world. This paper presented a comparative performance analysis of some load-balancing methods in three DCN topologies: fat-tree, BCube, and DCell. We proposed a load-balancing method based on BC. Our method was compared to other load-balancing algorithms in different DCN topologies using a one-to-one traffic scenario.

We conclude that our method outperforms the existing load-balancing methods in all the three DCN topologies investigated. It improves throughput by 202% and 33%, while decreasing delay by 20% and 22%, and energy consumption by 40% and 41% compared to the round-robin and least-congested methods, respectively. The evaluation and DCN topology graphs were implemented using NetworkX. In our future studies, we will evaluate the performance metrics in different traffic scenarios. Moreover, this study will be evaluated in a simulation program, i.e., ns3.

REFERENCES

[1] F. Estrada-Solano, O. M. Caicedo, and N. L. S. Da Fonseca, "Nelly: Flow detection using incremental learning at the server side of sdn-based data centers," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1362–1372, 2020.

[2] H. Geng, H. Zhang, X. Shi, Z. Wang, and X. Yin, "Efficient computation of loop-free alternates," *Journal of Network and Computer Applications*, vol. 151, p. 102501, 2020.

[3] M. A. S. Saber, M. Ghorbani, A. Bayati, K.-K. Nguyen, and M. Cheriet, "Online data center traffic classification based on inter-flow correlations," *IEEE Access*, vol. 8, pp. 60401–60416, 2020.

[4] L. Wang, X. Wang, M. Tornatore, K. J. Kim, S. M. Kim, D. Kim, K. Han, and B. Mukherjee, "Scheduling with machine-learning-based flow detection for packet-switched optical data center networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, pp. 365–375, April 2018.

[5] J. Ye, L. Feng, Z. Xie, J. Huang, and X. Li, "Fine-grained congestion control for multipath tcp in data center networks," *IEEE Access*, vol. 7, pp. 31782–31790, 2019.

[6] Y. Wang and S. You, "An efficient route management framework for load balance and overhead reduction in sdn-based data center networks," *IEEE Transactions on Network and Service Management*, vol. 15, pp. 1422–1434, Dec 2018.

[7] W. Wang, Y. Sun, K. Zheng, M. A. Kaafar, D. Li, and Z. Li, "Freeway: Adaptively isolating the elephant and mice flows on different transmission paths," in *2014 IEEE 22nd International Conference on Network Protocols*, pp. 362–367, IEEE, 2014.

[8] R. Dewanto, "Improved load balancing on software defined network-based equal cost multipath routing in data center network," *JURNAL INFOTEL*, vol. 10, no. 3, pp. 157–162, 2020.

[9] B. He, D. Zhang, and C. Zhao, "Hidden markov model-based load balancing in data center networks," *The Computer Journal*, 2019.

[10] H. Sufiev, Y. Haddad, L. Barenboim, and J. Soler, "Dynamic sdn controller load balancing," *Future Internet*, vol. 11, no. 3, p. 75, 2019.

[11] J. Zhang, F. R. Yu, S. Wang, T. Huang, Z. Liu, and Y. Liu, "Load balancing in data center networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 20, pp. 2324–2352, thirdquarter 2018.

[12] S. M. Shetty and S. Shetty, "Analysis of load balancing in cloud data centers," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–9, 2019.

[13] E. Bergamini, P. Crescenzi, G. D'angelo, H. Meyerhenke, L. Severini, and Y. Velaj, "Improving the betweenness centrality of a node by adding links," *Journal of Experimental Algorithmics (JEA)*, vol. 23, pp. 1–32, 2018.

[14] M. J. F. Alenazi and J. P. G. Sterbenz, "Comprehensive comparison and accuracy of graph metrics in predicting network resilience," in *2015 11th International Conference on the Design of Reliable Communication Networks (DRCN)*, pp. 157–164, March 2015.

[15] A. Elibol and N.-Y. Chong, "Topology graph pruning for optical mapping methods using edge betweenness centrality," in *2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 954–959, IEEE, 2019.

[16] F. Jamour, S. Skiadopoulos, and P. Kalnis, "Parallel algorithm for incremental betweenness centrality on large graphs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, pp. 659–672, March 2018.

[17] H. M. Helal and R. E. Ahmed, "Performance evaluation of datacenter network topologies with link failures," in *2017 7th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO)*, pp. 1–5, April 2017.

[18] Z. Han and L. Yu, "A survey of the bcube data center network topology," in *2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 229–231, May 2018.

[19] J. Li, X. Chang, Y. Ren, Z. Zhang, and G. Wang, "An effective path load balancing mechanism based on sdn," in *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 527–533, Sep. 2014.

[20] A. Kabbani, B. Vamanan, J. Hasan, and F. Duchene, "Flowbender: Flow-level adaptive routing for improved latency and throughput in datacenter networks," in *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pp. 149–160, 2014.

[21] Y.-L. Lan, K. Wang, and Y.-H. Hsu, "Dynamic load-balanced path optimization in sdn-based data center networks," in *2016 10th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, pp. 1–6, IEEE, 2016.

[22] N. S. B. Saeed and M. J. Alenazi, "Utilizing sdn to deliver maximum tcp flow for data centers," in *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, ICISS 2020, (New York, NY, USA), p. 181–187, Association for Computing Machinery, 2020.

[23] P. Wang, G. Trimponias, H. Xu, and Y. Geng, "Luopan: Sampling-based load balancing in data center networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, pp. 133–145, Jan 2019.

[24] U. Zakia and H. Ben Yedder, "Dynamic load balancing in sdn-based data center networks," in *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 242–247, Oct 2017.

[25] M. J. F. Alenazi and J. P. G. Sterbenz, "Evaluation and comparison of several graph robustness metrics to improve network resilience," in *2015 7th International Workshop on Reliable Networks Design and Modeling (RNDM)*, pp. 7–13, Oct 2015.

[26] F. S. Fizi and S. Askar, "A novel load balancing algorithm for software defined network based datacenters," in *2016 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom)*, pp. 1–6, Sep. 2016.

[27] M. Shafiee and J. Ghaderi, "A simple congestion-aware algorithm for load balancing in datacenter networks," *IEEE/ACM Transactions on Networking*, vol. 25, pp. 3670–3682, Dec 2017.

[28] R. Challa, S. Jeon, D. S. Kim, and H. Choo, "Centflow: Centrality-based flow balancing and traffic distribution for higher network utilization," *IEEE Access*, vol. 5, pp. 17045–17058, 2017.

[29] A. Rabih, A. Ghandour, Y. N. Shnaiwer, and S. Al-Ghadhban, "Rate-aware instantly decodable network codes for heterogeneous cellular networks," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, pp. 1–5, IEEE, 2019.

[30] C. Lee, C. Park, K. Jang, S. Moon, and D. Han, "Dx: Latency-based congestion control for datacenters," *IEEE/ACM Transactions on Networking*, vol. 25, pp. 335–348, Feb 2017.

[31] O. Popoola and B. Pranggono, "On energy consumption of switch-centric data center networks," *The Journal of Supercomputing*, vol. 74, no. 1, pp. 334–369, 2018.

[32] L. Gyarmati and T. A. Trinh, "How can architecture help to reduce energy consumption in data center networking?," in *Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking*, pp. 183–186, ACM, 2010.

Extending Shared-Memory Computations to Multiple Distributed Nodes

Waseem Ahmed

Department of Computer Science

Faculty of Computing and Information Technology

King Abdulaziz University, Jeddah, Saudi Arabia

Abstract—With the emergence of accelerators like GPUs, MICs and FPGAs, the availability of domain specific libraries (like MKL) and the ease of parallelization associated with CUDA and OpenMP based shared-memory programming, node-based parallelization has recently become a popular choice among developers in the field of scientific computing. This is evident from the large volume of recently published work in various domains of scientific computing, where shared-memory programming and accelerators have been used to accelerate applications. Although these approaches are suitable for small problem-sizes, there are issues that need to be addressed for them to be applicable to larger input domains. Firstly, the primary focus of these works has been to accelerate the core kernel; acceleration of input/output operations is seldom considered. Many operations in scientific computing operate on large matrices - both sparse and dense - that are read from and written to external files. These input-output operations present themselves as bottlenecks and significantly effect the overall application time. Secondly, node-based parallelization limits a developer from distributing the computation beyond a single node without him having to learn an additional programming paradigm like MPI. Thirdly, the problem size that can be effectively handled by a node is limited by the memory of the node and accelerator. In this paper, an Asynchronous Multi-node Execution (AMNE) approach is presented that uses a unique combination of the shared-file system and pseudo-replication to extend node-based algorithms to a distributed multiple node implementation with minimal changes to the original node-based code. We demonstrate this approach by applying it to GEMM, a popular kernel in dense linear algebra and show that the presented methodology significantly advances the state of art in the field of parallelization and scientific computing.

Keywords—*GPU; OpenMP; shared memory programming; distributed programming; CUDA*

I. INTRODUCTION

Many applications in engineering and scientific computing involve operations on dense and sparse matrices [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. At the core of many of these applications is the General Matrix-Matrix multiplication (GEMM) operation, which is regarded as one of the most widely used high performance kernels. GEMM is also used in graph theory and in the design of recommender systems and machine learning algorithms like page rank and logistic regression. It is used by many kernels in dense linear algebra (DLA) and has been part of many fast DLA libraries [12], [13], [14].

It is generally described by the operation $C = \alpha \cdot op(A) \times op(B) + \beta \cdot C$, where A, B and C are matrices and α and β are scalar constants between 0 and 1. The *GEMM* kernel is composed of three basic matrix-based operations - matrix-matrix multiplication, matrix-matrix addition and scalar-matrix multiplication. Among these three operations, the matrix-matrix multiplication makes GEMM computationally expensive. Moreover, when large dense matrices are involved, the GEMM operation, in addition to being computationally expensive, becomes I/O intensive as well and making the operation holistically efficient becomes more challenging.

Although this operation could naively be implemented using a simple three-nested loop with algorithmic complexity in $O(n^3)$, it has been the subject of a lot of research over the last few decades. Indeed, areas related to the optimization of this operation and its parallelization using improvised algorithms and software paradigms and hardware architectures are still active areas of research [12], [9], [15], [3], [16], [17], [18], [19], [20]. More recently, with the advent of easily available yet powerful workstations equipped with sophisticated co-processors, researchers have parallelized GEMM on these single-node workstations using accelerators like Graphics Processing Units (GPUs), Many Integrated Cores (MICs) and Field Programmable Gate Arrays (FPGAs) [19], [21], [12], [17].

When parallelizing GEMM on a single node, shared memory programming paradigms like OpenMP and heterogeneous computing paradigms like CUDA, OpenACC and OpenCL have been the dominant choice among developers. But as scientists continue to expand their problem domain, either to investigate larger problem sizes or to obtain finer results, developers will face three limitations when developing applications for a single-node execution. Firstly, is the cap on the maximum speedup that can be theoretically achieved on a single node that is a function of the core/compute element per node. Secondly, as the size of matrices grow ($>10^5 \times 10^5$), they become too large to be accommodated in the limited memory of a single node with a moderately sized RAM. For example, a single dense matrix of size $50k \times 50k$ used to store elements of double floating-point precision will require approximately 18.6GB of RAM. This makes the GEMM operation for these large matrices non-trivial to implement on a node with 32GB of memory or a high-end GPU with 16GB of device memory. Thus, when dealing with large matrices, parallelization on a single node will not yield the desired speedup because of memory constraints. Thirdly, the I/O operations that deal with reading these matrices from files into memory and writing them back from memory to file will become prohibitively expensive and will become the main bottleneck in the GEMM

operation.

One way to address this problem is to distribute the computation to multiple nodes. But this will require the developer to port the existing code written for a single-node to a distributed architecture using a distributed-memory paradigm like MPI. Moreover, this exercise is not trivial - the process of rewriting and porting scientific applications from one paradigm (or architecture) to another has a prohibitively expensive cost [22]. Besides the learning curve associated with the learning of a new programming paradigm, testing and debugging will be an extra burden that the developer will have to bear, in addition to addressing problems of load balancing and optimization *[9]*. Also, with the proliferation of platforms, architectures and programming paradigms in the HPC arena, it is indeed challenging for an application developer to command sufficient expertise to fully exploit the unique architectural features of multiple platforms and different programming paradigms.

Thus, as researchers continue to expand their problem domains to larger ones, a more holistic approach will be needed to address these categories of applications. This paper presents Asynchronous Multi-node Execution (AMNE) that attempts to address the afore-mentioned challenges associated with parallelizing a class of applications known as embarrassingly-parallel applications [23].

The rest of the paper is organized as follows: the next section presents the motivation for this approach. Section 3 describes the AMNE approach that forms the core of this paper. Pseudo-replication using the launcher script is explained in the section following Section 4. Evaluation and results of applying the approach to parallelize GEMM are presented next followed by the related work section. Conclusion is presented in the last section.

II. MOTIVATION

Consider a GEMM operation on a set of matrices. The total time taken to execute it sequentially on a single node comprises the time costs of both its core computation and I/O where the I/O costs could be disk-access related or network related. This can be expressed as

$$t_{seq} = t_i + t_{op} + t_o \qquad (1)$$

where t_i represents the time taken to read the input matrices *A* and *B* from file into memory, t_{op} represents the time taken to execute the core GEMM operation on matrices *A* and *B* to produce a resultant matrix *C*, and t_o represents the time taken to write matrix *C* from memory to file. Various approaches have been proposed in literature to reduce t_{op}, ranging from scalar optimizations like loop transformations [24], [25] and algorithmic replacement [26], [27] to parallelization of the core operation on multiple compute elements [18]. Applying any of these strategies will significantly reduce t_{op}. However, there is limited work in literature that provides strategies or methods to reduce t_i and t_{op} on a single node or on multiple nodes; details on explicitly reducing t_i or t_o are generally absent in work related to parallelization of GEMM.

So, if we assume t_i and t_o remain unchanged after the optimization/parallelization process, the speedup obtained on a single node ($S_{n=1}$) can be expressed as

$$S_{n=1} = t_{seq}/t_{par} = t_{seq}/(t_i + t_{op_{n=1}} + t_o) \qquad (2)$$

where $t_{op_{n=1}} \leq t_{op}$ is the time taken to execute the optimized or parallelized core operation (kernel) on a single node. Theoretically, if $t_{n=1} \rightarrow 0$, $S_{n=1}$ reduces to

$$S_{n=1} = 1 + t_{op}/(t_i + t_o) \qquad (3)$$

From this it can be seen that to obtain higher speedups, the second term in Eq. 3 needs to be maximized. But for large matrices, both t_i and t_o can become very large. For example, it takes about ~2745 secs to read a matrix of size $50k \times 50k$ from file to memory and to write it back from memory to file. This implies that to obtain higher speedups, a dedicated effort needs to be made to reduce both t_i and t_o. However, because of hardware limitations, reduction of t_i and t_o is indeed challenging on a single node regardless of which programming paradigm is used.

The primary focus of this paper is to help developers in addressing this challenge (to reduce t_i and t_o) while easily extending their single-node code implementation to a multiple-node implementation with minimal code change and with a zero or low learning curve for the developer. The next sections further elaborate this strategy.

III. ASYNCHRONOUS MULTI-NODE EXECUTION

The approach described in this paper, referred to as asynchronous multi-node execution (AMNE), seeks to address the two limitations of single-node execution in handling large matrix operations highlighted in the previous sections, namely, 1) to overcome the memory constraint of a single-node, 2) to reduce t_i and t_o and 3) to have minimal code changes tot he original single-node code when doing so. To address the first challenge, the computation is sub-divided or *sliced* into smaller sub-problems and the computation distributed over multi-nodes. This is a classical approach and is a well researched and documented strategy adopted by various researchers working with MPI and other distributed computing paradigms. However, unlike a master-slave relation that exists between processes in a distributed-computing programming paradigm like MPI, the launched instances in AMNE are independent of each other and no communication or synchronization exists between them. This is applicable to any embarrassingly-parallel application implemented using the AMNE approach.

Slices, as defined in this approach are unlike *tiles* and *blocks* used in the *tiling* and *batching* approach done at the node or GPU level. In the latter, partial (intermediate) results of the resultant matrices are generated by multiple compute elements and has to be aggregated with other partial results to obtain a *block(s)* of the final resultant matrices. In AMNE, on the other hand, the generation of a *slice* (or a partial-slice) of the resultant matrices is done by a single compute element. Additionally, in tiling and blocking, the tiles and blocks are in-memory representations of the matrices. A *slice*, on the other hand, is an out-of memory representation.

These multiple slices are operated on by independent, albeit similar, instances that are launched on multiple dedicated nodes in the cluster. To enable this, the single-node code

is to be transformed such that it can be controlled at run time to operate on any particular slice. Once that is done, a *launcher* script is used to initiate an asynchronous multi-node execution (AMNE) of the code instances, where the independent instances are asynchronously launched on multiple nodes in the cluster to operate on their designated *slices*.

These steps are further elaborated in the following sub-sections:

A. Slicing

Deciding the number of *slices* needed for a given application and defining their sizes and shapes, is a prerequisite for applying AMNE. In case of a homogeneous cluster, the slice sizes could be uniform. For example, consider a 20-node homogeneous cluster with each node having 32GB of memory. Also, consider the GEMM operation to be performed on this cluster for a configuration of $(80k \times 40k \times 80k)$. For this GEMM configuration, matrix A of size $80k \times 40k$ and matrix B of size $40k \times 80k$ need to be loaded into memory from file and elements of resultant matrix C of size $80k \times 80k$ need to be generated in memory. Assuming a double takes 8 bytes for storage, matrices A and B require $\sim 23.8GB$ memory each and matrix C requires $\sim 47.63GB$, a total of $\sim 95.27GB$. Clearly, it cannot be performed efficiently in a single pass on a single-node because of memory constraints of the node. To solve the problem in one pass, the GEMM computations are distributed across the 20 nodes in the cluster, with each node assigned a block(s) of matrix C for independent calculation. To enable this, both matrices A and B need to be sliced; these slices of matrices A and B are complete, in the sense that they can independently generate a complete *block(s)* of the resultant matrix C.

To decide the slice dimensions of matrices A and B, consider a slice of matrix A with dimensions $(l \times m)$ and a slice of matrix B with dimensions $(m \times n)$, to generate a block of resultant matrix C of dimensions $(l \times n)$. The choice of the variables l and n should be such that they satisfy the following relation

$$f(l, m, n) = [(l \times m) + (m \times n) + (l \times n)] \times 8] \leq R \times 2^{30} \quad (4)$$

where R refers to the node memory in GBytes. This simplifies to $f(l, 40k, n) \leq 4295$ for the example above where $m = 40k$ and memory, $R = 32GB$. Uniformly distributing the computation load on the cluster of 20-nodes, we get values $l = 4$ and $n = 4$ that also satisfy Eq. 4. These values of l and n yield dimensions of $(4k \times 40k)$, $(40k \times 4k)$ and $(4k \times 4k)$ for a slice of matrix A, a slice of matrix B and a block of resultant matrix C, respectively.

Slicing, in the case of a heterogeneous cluster, is more intense and depends on the memory constraints of each node or device on the node, in case an accelerator (GPU/MIC) is being used on the node. For example, a quad-socket multi-core node with 32GB RAM and an accelerator (GPU or MIC) with 16GB of device memory may be assigned a bigger slice (s_{gpu}) when compared to a slice (s_n) allocated to a single-socket dual-core node with no accelerator and with 4GB of RAM. In this case, the slices sizes will be considerable different with $s_{gpu} \gg s_n$.

Fig. 1. Domain Decomposition

One way to obtain slices is to fix a value for *l*, and heuristically obtain values of *n* based on experience or based on empirically obtained node peak performance values. This could be repeated for other values of *x* such that optimum performance is obtained.

There are two extreme scenarios when solving for appropriate values of *l* and *n* in Eq. 4.

1) If dimensions of the matrices are very large or the cluster size is small or the memory/node is small, then Eq. 4 may not be satisfied. This implies that a full slice of either matrix *A* or matrix *B* or slices of both matrices cannot be accommodated in the node memory along with a block of matrix *C*. In this scenario, a slice may need to be read piece-wise (tiled) into memory by a node; this will mean that two or more passes will be required to calculate each block of matrix *C* on a node.

2) If the matrices are small or the cluster size is big, then uniformly distributing the GEMM computation on all the nodes may not result in the best speedup. In this scenario, it would be better to limit the computation to fewer than the maximum nodes in the cluster. This is further explained later in this section.

There is no single one-size-fits-all strategy for deciding a slice size and shape and how they should be distributed among the nodes. The slices can be assigned to nodes in one of five ways as shown in Fig. 1 and as described below

1) One slice assigned to only one node - This is the case where the entire matrix *C* (as a single slice) is calculated by a single node as illustrated in Fig. 1(a). This is, also, the default method adopted by single-node developers using OpenMP, CUDA, OpenACC or OpenCL

2) Same-sized slices assigned to all nodes - In a cluster of *n* nodes, the calculation of matrix *C* is divided uniformly among *n* nodes. For example, in a cluster of $n = 12$ nodes, the calculation of matrix C is uniformly divided among all 12 nodes in the cluster, with each node calculating only 1/12th portion of matrix C. One slicing option that satisfies Eq. 4 is shown in Fig. 1(b)

3) Different-sized slices assigned to all nodes - In a heterogenous cluster of *n* nodes, calculation of matrix C is non-uniformly divided among the nodes in the cluster. This is shown in Fig. 1(c) where nine non-uniform slices are divided among $n = 9$ nodes in a cluster

4) Optimal distribution - Small GEMMs cannot fully exploit parallelism. Thus, in a homogeneous cluster, for relative smaller matrices, slices are assigned to only some nodes in a cluster such that nodes reach their peak capacity and the total time to execute (t_{max}) is minimized. For example, a tile size of $5120 \times 5120 \times 5120$ achieves 93% of peak performance (15TFlops) on an NVIDIA Volta 100 GPU [21]. In this option, some nodes in the cluster are left unassigned. This is illustrated in Fig. 1(d) where only six nodes of the cluster $(n > 6)$ compute their assigned slices of matrix C and the other nodes are left unassigned

5) Non-uniform optimal distribution - This is a combination of the previous two cases where non-uniform slices of matrix C are assigned to only some nodes in the cluster such that all nodes reach their peak capacity and the total time to execute (t_{max}) is minimized. Some nodes in the cluster are left unassigned. This is illustrated in Fig. 1(c) where only nine nodes of the cluster $(n > 9)$ compute their assigned slices of matrix C and other other nodes in the cluster are left unassigned.

B. Out-of-Memory Slicing

Traditionally, *tiles* represented in the tiling and blocking approaches have been in-memory representations of matrix partitions. Slices, as described in this paper, on the other hand, are out-of-memory representations of a matrix partition. They are defined externally and on a *shared-file system*.

In a *shared-file system* present in most HPC clusters, every node in the cluster has access to a common file system. All nodes in such clusters see a single, unified file system regardless of whether the file system physically resides on a single node or is spread over thousands of individual storage servers. An example of one such shared file system is *Lustre*, which was developed for extreme-scale compute clusters [28], [29], and is currently deployed in more than 60% of the top 100 supercomputers *[30]*. Some features of *Lustre* that are directly relevant to this paper are mentioned below:

1) It can support tens of thousands of client systems with petabytes (PB) of storage and handle hundreds of gigabytes per second (GB/sec) as I/O throughput

2) All clients in the cluster see a *single, coherent, synchronized namespace* [31]; the file system support concurrent read/write by multiple clients with data coherency maintained between all clients by the Lustre distributed lock manager (LDLM) [31].

3) Files on Lustre are stored as one or more objects, with each object stored on a separate Object Storage Target (OST); this allows several clients to write to different parts of the same file simultaneously while allowing other clients to read from the file

4) Lustre supports file striping, where a file may be stored on multiple (≤ 2000) OSTs . This considerably increases the I/O bandwidth (aggregate) by a factor of almost 2000 for a single file when accessed by multiple nodes.

In the approach presented in this paper, all files that contain the input data (matrices *A* and *B*) are not stored locally but on such a shared-file system. Similarly, the output data (matrix *C*) is written to a file on the shared-file system. This implies that every node in the cluster can independently and simultaneously read the input data without having it to be explicitly broadcasted to the slave-nodes by a master-node as is the case in many MPI-based approaches. Also, the distributed storage of a file as multiple OSTs and the file striping feature of the shared-file system makes access to the files much faster when simultaneously performed by the nodes.

Fig. 2. Calculation of a Block of Matrix C

Once slices have been defined, the developer chooses the storage format in one of two ways

1) All slices of the input matrices stored in separate single files on the shared-file system. In the case of GEMM, this means all elements of matrix A are stored in a single file. And all elements of matrix B stored in a separate single file.

2) Each slice of the input matrices stored as a separate file on the shared-file system. In the case of GEMM, this means that if matrix *A* has been divided into twenty slices, there will be twenty separate files used to store elements of matrix *A*, with each file corresponding to elements of a particular slice.

Once slices have been stored as files in either of the two ways described above, they need to be accessed by their respective nodes. For example, consider a GEMM operation illustrated in Fig. 2 where a node has been assigned a complete block of C (shaded) for calculation. This node requires only the elements in the shaded slice of matrix *A* and the elements in shaded slice of matrix *B* to calculate elements in the shaded block of *C*, as indicated in the figure. The other nodes may also access the same slices of matrices *A* and *B* for their calculations, but these accesses are independent of each other as mentioned previously in this section.

These simultaneous access allowed by the shared-file system to the input data by the nodes in the cluster,

considerably brings down the sequential access times, t_i and t_o in Eq. 3. Rewriting Eq. 3 for multiple nodes ($n > 1$) we get

$$S_{n>1} = t_{seq}/t_{par} = t_{seq}/(t_{i_{n>1}} + t_{op_{n>1}} + t_{o_{n>1}}) \qquad (5)$$

It was empirically observed that $t_{i_{n>1}} \rightarrow t_i/n$ and $t_{o_{n>1}} \rightarrow t_o/n$ for various values of slice dimensions and sizes. Theoretically, for an embarrassingly parallel application, and for an infinite cluster size, when $t_{op_{n>1}} \rightarrow 0$ then $S_{n>1} \rightarrow n$ and attains a close-to-peak linear speedup.

C. Code Modification

The main objective of this step is to ensure that code designed specifically for a single-node execution is extended to one that can be executed on a distributed-memory cluster.

Based on decisions taken in the previous step, each node will need to calculate only its specific portion of the slice. Code will, thus, need to be modified such it can be executed in a *slice*-specific fashion on a node in the cluster. When coding for a heterogeneous implementation, there is an additional challenge at this stage as multiple versions of the code need to be explicitly managed. For example, CPU-only code and code for a CPU+GPU execution will be different and have to be handled differently when being compiled and executed by the *launcher* script (described later). This could be managed in one of two ways:

1) same code for all nodes (controlled using pre-processor directives and conditional compilation to differentiate the CPU-only and CPU+GPU codes);
2) different code for different nodes (by explicitly separating the CPU-only and CPU+GPU codes).

Both options are supported in this approach and are managed by the job script (described later).

The first objective in this step is to ensure that all operation-specific variables are made amenable to modification at run time and are not hard-coded. This modification of variables at run-time could be done either through a configuration file or using command-line parameters. These variables, referred to as run-time parameters (RTP) in the rest of this paper, ensure that the same code can be used to operate on different slices and that it's execution behaves in a pure SPMD fashion. The second step is to ensure that all files needed for the read and write operations, including configuration files, are stored on the shared-file system.

To illustrate, consider a GEMM operation coded for a single-node execution using OpenMP, OpenACC or CUDA. In the original single-node implementation of the code, the dimensions of the matrices A, B and C, i.e. $(l \times m), (m \times n)$ and $(l \times n)$, respectively, and names of filenames that store data for these matrices may be hard-coded and may not be amenable to modification at run-time. In this case, code is modified such that these variables can be modified at run time as a RTP. The RTP in this case are the three-tuple $< l, m, n >$ and the full path-names for matrices A, B and C. Also, if the files that contain the elements of matrices A and B are stored locally on the node, they need to be moved to the shared-file

system. Similarly, the file that is generated for matrix C, also has to be placed on the shared-file system.

IV. JOB EXECUTION AND LAUNCH

A. Job Script Preparation

Once the original single-node code has been modified in the previous steps, it needs to be compiled on the node that it needs to be executed on and then launched as an executable on that node. This is done using a job script specifically prepared for each defined *block* of matrix C. The responsibility of the job script is to load the modules, libraries and the environment necessary for successfully compiling the code and executing it on the node. For example, code based on CUDA and cuBLAS needs to be compiled on a node in the cluster that has one or more Nvidia GPUs installed on it. Prior to compilation, the appropriate compilers and libraries, in this case, the appropriate version of the *nvcc* compiler and the cuBLAS libraries, need to be loaded into the environment on the node where the code is to be compiled. After creating the required environment, the job script has to compile the code using the specified compiler flags and links to libraries. If the compilation is successful, the executable is to be launched on that node with it's *node*-specific RTP supplied as command-line arguments or through a configuration file. Each *block* of the output has its own job script. These job scripts can either be manually created or automatically generated if the slices are large in number.

The collection of these job scripts is then submitted to the cluster's job scheduler by a *launcher* script, which queues the jobs based on the job's priority and the requested type and number of nodes in the cluster; the jobs are launched as and when the requested nodes becomes available.

The application is said to have completed execution when all the jobs in the collection have successfully terminated.

B. Pseudo-Replication

AMNE results in a *pseudo-replication* of the instances on the cluster. In pseudo-replication, concurrent instances of an executable are asynchronously launched on multiple nodes in the cluster with each instance executing to create a distinct portion of output. Although the AMNE behavior may initially appear to be similar to *task replication* used in distributed systems [32], [33], [34], [35], it differs in many aspects as described below:

1) In task replication, all replicated instances of the executable are identical and they work on the same data. The job scheduler in the cluster launches the same executable multiple times. In pseudo-replication, on the other hand, the replicated instances work on different data. The job script is different for each instance and is dynamically generated depending on the size of the cluster and other parameters.
2) In task replication, only the result from the first successful execution is used; all other executing instances of the executable are either explicitly terminated or ignored. This results in poor and inefficient resource utilization as the outputs from the (other) slower replicated instances are completely

ignored. In pseudo-replication, results generated by all the replicated instances are used and no launched instance is prematurely terminated. This results in better resource utilization of the cluster resources.

3) The primary objective of task replication is to address latency in nodes, to improve quality of service or to improve application fault tolerance [34]. Parallelization of tasks is not an objective. On the other hand, in pseudo-replication, parallelization of tasks is the primary objective and is explicitly desired.

4) In task replication, the total execution time of an application is calculated as the time taken by the fastest successfully completed instance ($t = min(t_1, t_2, \ldots, t_n)$, where t_i denotes the time take by node i to complete a task). On the other hand, the total execution time of an application in pseudo-replication is calculated as the time taken by the slowest successfully completed instance ($t = max(t_1, t_2, \ldots, t_n)$).

Pseudo-replication also differs from an MPI-based execution. In applications coded using MPI, identical multiple instances are concurrently launched on nodes in the cluster using the *mpirun* command from the shell. These multiple instances work together as one coherent entity (based on a master-slave model) with active message-passing based communications between them. The instances in pseudo replication, on the other hand, are launched independently using a launcher script and work in the form of a loose asynchronous fork-join model. The instances are completely independent of each other and no form of direct or indirect communication exists between them.

V. EVALUATION AND RESULTS

In this section, the proposed approach is evaluated for a GEMM operation on large matrices and results presented.

A. Experimental Testbed

All experiments for this work were carried out on the *Aziz* supercomputer at King AbdulAziz University. *Aziz* was ranked 359th in the top500 list in the June 2015 ranking with a peak performance of 228.6TFlops [36].

The Aziz cluster has 380 standard compute nodes with 96GB of memory and 112 high memory compute nodes with 256GB of memory. All nodes have a dual-socket Intel Xeon E5-2695 12-core processor running at 2.40 GHz. Nodes with GPU-based accelerators have 96GB of RAM and are equipped with Nvidia's Tesla-based K20 GPUs, which consists of 2496 CUDA cores and 5GB of device RAM.

The approach was evaluated on two types of nodes of Aziz - The first set consisted of nodes with 256GB of RAM with no accelerators, and the second set consisted of nodes with 96GB of RAM with K20 GPUs. These two sets are referred to as *setA* and *setB*, respectively, in the rest of this section.

Code was written in C/C++ and used Intel's MKL and NVidia's cuBLAS libraries. Code that used Intel's MKL libraries was compiled using Intel's *icc* version 17.0.2 and CUDA code was compiled using NVidia's CUDA 9.2. Jobs on *Aziz* were scheduled using the PBS Pro job 12.2.0 scheduler [37].

TABLE I. SIZE OF MTX FILES ON DISK

	M	N	Size of file
1	10000	10000	1.8 GB
2	20000	20000	7.5 GB
3	30000	30000	17.0 GB
4	40000	40000	31.0 GB
5	50000	50000	48.0 GB
6	100k	100k	84GB

B. Benchmarks

Although the approach is applicable to any embarrassingly parallel application, the GEMM operation is specifically used here to illustrate and evaluate the approach. The GEMM operation, described by $C = \alpha \cdot op(A) \times op(B) + \beta \cdot C$, is an operation on two input matrices A and B to produce a resultant matrix C. The values one and zero and were used for the scalars α and β, respectively.

All matrices used in this work were generated as double-precision elements as files on disks in the MTX format, the format which is used for storing large sparse matrices in the SuiteSparse Matrix Collection [38]. These MTX files are very large even for a single large matrix. Table I serves as an approximate indicator to the size of files when floating-point value matrices of size ($M \times N$) are stored in the MTX file format. It is apparent from the size of the files, that operations involving large matrices cannot be easily executed in one-pass on nodes with moderately sized RAMs or on GPU-based accelerators with small device memory.

For experiments on setA, matrices that corresponded to a GEMM configuration of ($50k \times 50k \times 50k$) were used and for experiments on setB, matrices that corresponded to a GEMM configuration of ($10k \times 10k \times 10k$) were used.

All file were stored on Aziz on the Fujitsu Exabyte File System (FEFS) which is a scalable parallel file system based on Lustre and designed for Fujitsu HPC clusters [39]. Elements of matrix A and matrix C were stored in the row-major order and elements of matrix B were stored in column-major order to enable easier slicing.

C. Slicing

For each experiment, matrices *A* and *B* were divided into *slices* based on the number of nodes and the type of nodes available in the cluster. Since the type of nodes in both sets was homogeneous and the matrices used in the GEMM operation were sufficiently large, the number of slices (*s*) was set equal to the number of nodes.

The *shapes* (dimensions) of the slices, however, can be decided in various ways. For example, a matrix of size $40k \times 40k$ could be sliced across eight homogeneous nodes in various shapes - as eight equal-sized slices of $20k \times 10k$ elements or eight equal-sized slices of $5k \times 40k$ elements, etc. based on the memory configuration of the node and other criteria. On a heterogeneous cluster, the *slice* sizes and shapes assigned to nodes could be different.

The largest slice used for a single-node implementation on setA was for GEMM configuration ($50k \times 50k \times 50k$). For setB, the slice size was limited to a GEMM configuration ($10k \times$

TABLE II. SLICES USED FOR EXPERIMENTS ON SETA

No.	number of slices	shape <l,m,n>
1	1	<50k,50k,50k>
2	2	<25k,50k,50k
3	5	<10k,50k,50k>
4	10	<5k,50k,50k>
5	20	<2.5k,50k,50k>
6	50	<1k,50k,50k>

TABLE III. SLICES USED FOR EXPERIMENTS ON SETB

No.	number of slices	shape <l,m,n>
1	1	<10k,10k,10k>
2	2	<5k,10k,10k>
3	4	<2.5k,10k,10k>
4	5	<2k,10k,10k>
5	10	<1k,10k,10k>
6	20	<0.5k,10k,10k>

$10k \times 10k$) for a single-node implementation to ensure slices for both matrices A and B simultaneously fit in the memory of the GPU (NVidia K20). The shapes of slices used in this work for experiments on setA and setB are given in Tables II and III, respectively.

D. Code Modification

Two different node-based code versions were used to implement the *GEMM* operation - one was based on Intel's MKL [14] and the other was based on CUDA and its cuBLAS libraries [13]. Both code versions consisted of four main code sections, that performed the following four main operations:

1) Allocation of memory for the matrices on the host and/or device.
2) Reading the input matrices A and B from files into (host and/or device) memory.
3) Execution of the core GEMM (kernel) operation on matrices A and B to produce matrix C.
4) Writing the resultant matrix C from (device and/or host) memory onto to files.

Both Intel's MKL library and NVidia's cuBLAS provide single-node implementations for GEMM that is optimized for their respective target platforms. They are *cblas_dgemm()* and *cublasDgemm(),* respectively, for matrices with double-precision elements. *cblas_dgemm()* expects the two input matrices A and B to be provided in the row-major format and produces matrix C, also in the row-major format. On the other hand, cublasDgemm() expects the two input matrices A and B to be provided in the column-major format and produces matrix C, also in the column-major format. Thus, the second section in code, responsible for reading data from files on disk into memory, in their respective implementations, had to be slightly modified to accommodate for this slight variation. Similarly was the case for the last section in code responsible for writing the resultant matrix to the file.

There were six run-time parameters (RTP) required for the GEMM implementation - three configuration parameters l, m and n, which represented the dimensions $l \times m$, $m \times n$ and $l \times n$ of the slices of matrices A, B and C, respectively and three

file names with the full pathnames where the slices/blocks of matrices A, B and C are or will be stored on the FEFS shared-file system. The filenames associated with blocks of the resultant matrix C are kept unique to enable a block to be differentiated from other blocks of matrix C. All relevant sections in code were modified to accept these six RTP from the command-line instead of a configuration file. The third section in code responsible for executing the kernel operation consisted of the GEMM function call in the library. The function calls in both versions of code (MKL and cuBLAS) were modified to reflect the RTPs.

In general, all hard-coded references to L, M and N in both versions of code were changed, where needed, to reflect ones provided as RTP at run time from the command line.

E. Launcher and Job Scripts

The Aziz cluster consists of nodes with different configurations and each of this configurations has a unique queue associated with it. Thus, a job placed on the queue of the PBS scheduler [37] may need to be configured individually. Additionally, as each *block* generation is unique, it is controlled through a unique job script. For each defined *block* of matrix C, the following details have to be specified in the job script for the node responsible for generating the *block*.

1) Node environment required for compilation (PBS directives, compiler version and libraries to load, number of threads, etc.)
2) Command for compilation (compiler path, optimization flags, linker options, etc. specific to the slice).
3) Absolute pathname of files on the FEFS shared-file system where the node-specific slices need to be read from or written to.
4) Command for execution along with the node-specific RTP.

For a homogeneous cluster, these details will be similar for all slices. For a heterogenous cluster, the details will vary depending on the node configuration on which the block is to be generated. The job details can either be supplied by an external configuration file or hard coded into the launcher script. The launcher script then loops through the job details for each slice, dynamically generates a job script and submits the jobs collectively to the PBS job queue for execution on the cluster. The algorithm for the *launcher* script is described in Algorithm listing 1. It basically consists of a loop which is executed n times, with each iteration creating a node-specific job script (*job.i*) based on the node configuration and block allocated to it. After the job script has been generated for the node, it is then submitted to the appropriate queue on Aziz using the *qsub* command of PBS. The launcher script terminates once all jobs have been successfully submitted to the PBS queue, which then schedules them based on the priority of the jobs and the availability of resources on Aziz. It should be noted that in any multi-user cluster environment, the availability of nodes is largely non-deterministic and influences all approaches similarly, and is not considered as a parameter in evaluating the efficacy of any approach or algorithm in a multi-user distributed-computing environment.

Algorithm 1 Launcher script

```
struct block {
    rtp    // run-time Parameters
    env    // compiler environment
    compile // compile command
    exec // launch command
    queue    // Q name to submit on Aziz
}

def prepare_job(i, b):
    s = get_env(b.env)
    s += compile_cmd(b.compile)
    s += exec_cmd(b.exec, b.rtp)

    write_file(jobfile.i, s)
end_def

b[] = init(N) // init block details
for i in N
do
    prepare_job(i, b[i])
    submit_q(jobfile.i, b[i].queue)
done
```

F. Results and Analysis

Experiments to evaluate the proposed approach were done on both setA and setB by varying *n*, the number of nodes in the virtual cluster. Slices were uniformly distributed between the nodes in both the experiments. Readings in seconds were recorded using the *omp_get_wtime()* function for multiple runs and different *slice* size combinations as tabulated in Tables II and III. Due to the similarity that exists between single-node instance execution times on the cluster for AMNE, multiple readings for single-node instance executions were recorded on a single node for all slice and cluster sizes ($n > 1$).

Individual times were recorded for the following: 1) read operation, i.e. to read a slice of matrix A and a slice of matrix B from the shared-file system into memory; 2) GEMM operation on slices of matrices *A* and *B* to produce a block of matrix *C,* and 3) write operation, i.e. to write the slice of matrix *C* from memory to the shared-file system. When a GPU was used, transfer times for moving data between host and device and back were included in the GEMM operation time and not in the read and write. Timings for the experiments were recorded as $t = max(t_1, t_2, ...t_n)$ which represented the time taken to perform the GEMM operation by the slowest node in the virtual cluster on its allocated slices where t_i is the time taken in seconds to execute the operation on node *i,* and $\forall i, 1 \leq i \leq n$.

The read and write operations in both the experiments were performed by a single thread on the node and these operations were not parallelized at the node-level. All slices of matrices A and B, required for the experiments, were stored as separate files on the shared-file system. Blocks of resultant matrix C were also generated as separate files on the shared-file system by the multiple instances of the AMNE. Slices in all cases were stored in the MTX format, as stated earlier.

Timings for the read, write and the core GEMM operations were recorded separately and plotted as shown in Fig. 3. The total time represented in the figure is the sum of the time taken for the read, write and the core GEMM operation. Also, all readings in the figure represent readings recorded with minimal change to the original code; no extra node-based optimizations were performed to the code in any of the experiments.

Fig. 3(a) shows the readings for experiments on setB using a K20 GPU for different six different values of *n*, the virtual cluster size for a GEMM configuration of $10k \times 10k \times 10k$. The size of the configuration was chosen so as to comfortably enable a single-pass execution when *n=1*. The code was compiled using CUDA 9.0 and used the *cublasDgemm()* function from the cuBLAS library. Fig. 3(b) shows the readings for experiments on setB but without using any accelerator. Six different values of *n*, were used in this case as well for a GEMM configuration of $(10k \times 10k \times 10k)$. The code in this case was compiled using *icc* and used the *cblas_dgemm()* function from Intel's MKL. Fig. 3(c) shows the readings for experiments on setA without using any accelerator. Six different values of *n*, were used in this case as well for a GEMM configuration of $(50k \times 50k \times 50k)$. The code in this case was compiled using *icc* and used the *cblas_dgemm()* function from Intel's MKL. A GPU was not used in the experiments on setA as the matrix size was too large to fit in the device memory without making a substantial change to the original code/algorithm and node-based optimization was not an intent of this research.

Speedups for all operations with $n > 1$ were calculated relative to the timings obtained for a single-node execution ($n = 1$). It can be observed from the Figure, that the speedup for both the read and write operations varies almost linearly with the increase in the number of nodes. However, for the core GEMM operation, a small dip was observed when the sizes of the matrices became smaller. In the case of the GPU this was expected as the data transfer time (communication) between device and host start to dominate the computation time. This has also been documented in [21] which notes that the efficiency of the Dense Linear Algebra (DLA) libraries for smaller sizes of matrices is not as high as for larger matrix sizes. Thus, having slices that lie beyond the dip would not be profitable. This gives a clear indication that the optimal size of slices has to be decided, either based on developer experience or using results from prior auto-tuner software runs for the application. However, in this experiment, as the time taken to execute the core operation was negligible when compared to the time taken to read and write data, the total speedup for the operation was not affected much and showed an almost linear speedup as well. No such dip was observed in readings for the experiment on setA as the slice sizes were sufficiently large to enable optimized parallelization.

For these GEMM configurations, node-scalability performance was observed to be almost linear with *n*. Extrapolating from the graphs in the figure it can stated that $S_{n>1} \rightarrow n$ in Equation 5.

G. Comparison with other Approaches

Since this work is a multi-node implementation, comparisons with any single-node approaches would be

Fig. 3. Speedups obtained by Asynchronous Batching on Cluster Nodes

unfair as the speedup obtained using this approach will always be much higher for large matrices. Also, the size of matrices that can be handled on single-nodes in a single-pass is constrained by node-memory. On the other hand, using the presented approach, matrices of very large sizes can be handled comfortably in a single-pass on reasonably large clusters.

When compared to other multi-node methods, this approach performs equally well if not better. This is because in most MPI-based and distributed-computing approaches, reading data from files is mainly done by a single process [15]. Although the overall operation time could be reduced by overlapping communication and computation using clever heuristics [40], [41], it is still an extra exercise. Else, dealing with large matrices would mean that the speedup would be constrained by the large values of t_i and t_o in Equation 5. Also, the criss-cross message passing between processes to aggregate the partial (intermediate) results, creates an increase in the network activity on the cluster. The approach proposed in this paper has an added advantage of simplicity and increased developer productivity over MPI-based approaches, as coding, testing, debugging and workload balancing using CUDA or OpenMP single-nodes are relatively simpler operations than their equivalents on traditional distributed-memory clusters using MPI [42], [10].

VI. RELATED WORK

The optimal distribution of computations across compute elements to reduce an application's overall execution time is a decades-old, well-researched and well-documented problem [43], [44], [45]. A few common approaches used by researchers when dealing with large matrices in applications is to change the mathematical formulation of the operation[46], [20], to use secondary memory [18], [20], to pipeline the operations [13], to operate on data in multiple passes, or to distribute these operations over compute elements either on a single node or across multiple nodes.

The approach adopted on distributed platforms, is generally based on a master-slave process model where blocks of data are broadcast by a node to other nodes in the cluster [15]. Another popular approach is to use *tiling* followed by *batching*, where *tiling* refers to the partitioning of the matrices into tiny blocks or tiles, while *batching* refers to the assignment of these tiles to threads or computing elements for computation.

Batching has been used on GPUs in combination with tiling to parallelize GEMM [12], [21]. Here, the GEMM operation is broken down into smaller GEMMs, the computation of which is then distributed among the available compute elements on a single-node (CPU or GPU). The compute elements then generate partial results of the resultant matrix *C,* which are then combined to obtain the final resultant matrix. Here, blocks are generally of uniform size and are in-memory representation of the matrices. Also, data reuse is an important concern in efficient batching as memory is shared between computing elements. In AMNE on the other hand, *slicing* (a combination of tiling and batching) is an out-of-memory operation and data reuse at the inter-node cluster level is not a concern. Also, in these approaches, the reduction of the time involved in I/O when dealing with large matrices is not explicitly addressed.

The number of nodes in an exascale system is expected to grow to more than 50 times of what it was in 2010 [47]. When programming using OpenMP or CUDA, parallelism is restricted to the node on which the program executes and cannot be extended beyond that node. Also, it becomes increasingly difficult for DLA libraries like PLASMA and MAGMA to handle large-scale matrix multiplication on a single-node due to hardware resource limitations [15]. To extend parallelism beyond a single node, a pure distributed programming paradigm like MPI could be used. But some researchers prefer a hybrid combination over a pure MPI implementation [48] and there have been many who have used multiple programming paradigms on heterogeneous platforms in various combinations [4], [11], [49], [48], [8], [10], [22], [50], [3].

However, for a developer who is only proficient in using single-node-based programming models like OpenMP or CUDA, learning to code, test and debug using an additional distributed memory paradigm, like MPI, could be a big challenge. On the other hand, not utilizing the potential parallelism inherent in a large distributed cluster, also, is not desirable. But to efficiently combine programming paradigms like MPI and OpenMP/CUDA requires the developer to have intricate knowledge of the multiple hardware architectures and programming paradigms to fully exploit the combined platform's capability. For example, besides the knowledge about the number and type of streaming processors (SM) and cores on the GPU, the size and memory hierarchy on the device, registers per thread, the CUDA version that it supports and its compute-capability, support for unified memory, the

optimal dimensions for the grid and block combination to use, etc. which are required to achieve optimized performance on the GPU, a developer will be required to know the network-topology of the cluster, communication modes between processes, sophisticated algorithms that relate to efficient overlapping of computation and communication, [40], [41], etc. when using MPI. This leads to lower developer productivity.

To handle large matrices by overcoming memory-constraints of a single-node or a GPU, out-of core computations [51], [1], [7], [9] have been used which seek to decompose a matrix into smaller pieces and operate on them in multiple passes; data resides in disk and has to be explicitly moved in and out of memory for the passes. While this addresses the problem of handling large matrices on memory-constrained single-nodes and accelerators, the achievable speedup is constrained by the sequential passes that the algorithm has to make.

VII. Conclusion

Single-node programming paradigms like OpenMP, CUDA and OpenACC have recently gained popularity among researchers in Engineering and scientific computing and optimized libraries like Intel's MKL, NVidia's cuBLAS, and PLASMA are available to them that have been optimized for execution on single-nodes. However, as HPC moves to Exascale, and the number of nodes available in a cluster is set to dramatically increase, researchers will continue to expand their problem domains, either to investigate larger problem sizes or to obtain finer results. These domain sizes will be much larger than can be comfortably handled on resource constrained single-nodes. In such cases, efficient ways of handling the large I/O associated with these large applications and distributing the computations beyond a single-node and across multiple nodes will be needed. For the latter, porting single-node implementations to multiple-nodes will be needed which is a large and non-trivial exercise and will require time and developer expertise to fully exploit the unique architectural features of multiple platforms and different programming paradigms.

In this paper, an Asynchronous Multi-node Execution (AMNE) approach was proposed that works in combination with existing node-based optimizations while addressing the afore-mentioned challenges. AMNE combines shared-file storage commonly available in HPC clusters, pseudo-replication and out-of-memory matrix slicing in a unique way, which reduces the I/O time considerably with minimally-required code changes which positively influence developer productivity. AMNE was evaluated for the GEMM operation on large matrices. Matrices, stored as files on the shared-file system, were partitioned into slices at the node-level and slices allocated to nodes such that a block of the resultant matrix could be independently calculated by a node. The calculations were then independently run in an AMNE fashion to generate the final resultant matrix. Different combinations of nodes were used with different *slice* shapes and sizes. Results showed that the speedup obtained on multi-nodes using AMNE for the overall GEMM operation, including both I/O and computation times, over single-node execution time was almost linearly scalable with the number of nodes allocated to it in the cluster.

Although the approach was evaluated for GEMM on large matrices in this paper, AMNE can be easily extended to any HPC application that falls in the category of embarrassingly-parallel applications.

References

[1] M. G. Awan, F. Saeed, and F. Saeed, "An out-of-core GPU based dimensionality reduction algorithm for big mass spectrometry data and its application in bottom-up proteomics," in *in Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, August 2017.

[2] P. Czarnul, "Parallelization of large vector similarity computations in a hybrid CPU+GPU environment," *Journal of Supercomputing*, pp. 768–786, 2018.

[3] M. Kreutzer, J. Thies, M. Röhrig-Zöllner, A. Pieper, F. Shahzad, M. Galgon, A. Basermann, H. Fehske, G. Hager, and G. Wellein, "GHOST: Building blocks for high performance sparse linear algebra on heterogeneous systems," *International Journal of Parallel Programming*, vol. 45, no. 5, pp. 1046–1072, 2017.

[4] V. Lončar, L. E. Young-S., S. Škrbić, P. Muruganandam, S. K. Adhikari, and A. Balaž, "OpenMP, OpenMP/MPI, and CUDA/MPI C programs for solving the time-dependent dipolar Gross-Pitaevskii equation," *Computer Physics Communications*, vol. 209, pp. 190 – 196, 2016.

[5] F. Rabbi, C. Daley, H. Aktulga, and N. Wright, "Evaluation of directive-based GPU programming models on a block Eigensolver with consideration of large sparse matrices," Lawrence Berkeley National Laboratory, Tech. Rep., 2020.

[6] F. Yu, P. Strazdins, J. Henrichs, and T. Pugh, "Shared memory and GPU parallelization of an operational atmospheric transport and dispersion application," in *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2019, pp. 729–738.

[7] M. Chillarón, G. Quintana-Ortí, V. Vidal, and G. Verdú, "Computed tomography medical image reconstruction on affordable equipment by using out-of-core techniques," *Computer Methods and Programs in Biomedicine*, vol. 193, p. 105488, 2020.

[8] X. Guo, J. Wu, Z. Wu, and B. Huang, "Parallel computation of aerial target reflection of background infrared radiation: Performance comparison of OpenMP, OpenACC, and CUDA implementations," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 4, pp. 1653–1662, 2016.

[9] D. Zheng, D. Mhembere, V. Lyzinski, J. T. Vogelstein, C. E. Priebe, and R. Burns, "Semi-external memory sparse matrix multiplication for billion-node graphs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 5, May 2017.

[10] D. S. Henty, "Performance of hybrid message-passing and shared-memory parallelism for discrete element modeling," in *Proceedings of the 2000 ACM/IEEE Conference on Supercomputing*. Washington, DC, USA: IEEE Computer Society, 2000. [Online]. Available: http://dl.acm.org/citation.cfm?id=370049.370069

[11] S. R. Miri Rostami and M. Ghaffari-Miab, "Finite difference generated transient potentials of open-layered media by parallel computing using OpenMP, MPI, OpenACC, and CUDA," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 10, pp. 6541–6550, 2019.

[12] J. Dongarra, S. Hammarling, N. J. Higham, S. D. Relton, P. Valero-Lara, and M. Zounon, "The design and performance of batched BLAS on modern high-performance computing systems," *Procedia Computer Science*, vol. 108, pp. 495 – 504, 2017.

[13] Nvidia, "Cuda toolkit documentation," *https://docs.nvidia.com/cuda/cublas/index.html*.

[14] Intel, "Developer reference for intel math kernel library," *https://software.intel.com/content/www/us/en/develop/*.

[15] R. Gu, Y. Tang, C. Tian, H. Zhou, G. Li, X. Zheng, and Y. Huang, "Improving execution concurrency of large-scale matrix multiplication on distributed data-parallel platforms," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 9, pp. 2539–2552, 2017.

[16] K. Li, Y. Pan, and S. Q. Zheng, "Fast and processor efficient parallel matrix multiplication algorithms on a linear array with a reconfigurable pipelined bus system," *IEEE Transactions on Parallel and Distributed Systems*, vol. 9, no. 8, pp. 705–720, Aug 1998.

[17] R. Lim, Y. Lee, R. Kim, and J. Choi, "OpenMP-based parallel implementation of matrix-matrix multiplication on the Intel Knights landing," in *Proceedings of Workshops of HPC Asia*, 2018, pp. 63–66.

[18] M. Marques, G. Quintana-Orti, E. S. Quintana-Orti, and R. A. van de Geijn, "Solving large dense matrix problems on multi-core processors," in *2009 IEEE International Symposium on Parallel Distributed Processing*, May 2009, pp. 1–8.

[19] J. Shen, Y. Qiao, Y. Huang, M. Wen, and C. Zhang, "Towards a multi-array architecture for accelerating large-scale matrix multiplication on FPGAs," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5.

[20] J. Alman and V. V. Williams, "Limits on all known (and some unknown) approaches to matrix multiplication," in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 2018, pp. 580–591.

[21] X. Li, Y. Liang, S. Yan, L. Jia, and Y. Li, "A coordinated tiling and batching framework for efficient GEMM on GPUs," in *Proceedings of the 24th Symposium on Principles and Practice of Parallel Programming*, 2019, pp. 229–241.

[22] M. Martineau, S. McIntosh-Smith, M. Boulton, and W. Gaudin, "An evaluation of emerging many-core parallel programming models," in *Proceedings of the 7th International Workshop on Programming Models and Applications for Multicores and Manycores*, ser. PMAM'16. New York, NY, USA: ACM, 2016, pp. 1–10. [Online]. Available: http://doi.acm.org/10.1145/2883404.2883420

[23] K. Hwang, *Advanced computer architecture with parallel programming*. McGraw-Hill, 1993.

[24] P. Boulet, A. Darte, G. A. Silber, and F. Vivien, "Loop parallelization algorithms: From parallelism extraction to code generation," *Parallel Computing*, vol. 24, no. 3-4, 1988.

[25] V. Sarkar, "Optimized unrolling of nested loops," *International Journal of Parallel Programming*, vol. 29, no. 5, 2001.

[26] J. Ansel, C. Chan, Y. L. Wong, M. Olszewski, Q. Zhao, A. Edelman, and S. Amarasinghe, "Petabricks: A language and compiler for algorithmic choice." *ACM SIGPLAN Conference on Programming Language Design and Implementation, Dublin, Ireland,*, vol. 44, pp. 38–49, June 2009.

[27] J. Ansel, "Petabricks: a language and compiler for algorithmic choice," Master's thesis, MIT, 2009.

[28] P. Dickens and J. Logan, "Towards a high performance implementation of MPI-IO on the Lustre file system," in *Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008. Part I on On the Move to Meaningful Internet Systems:*, 2008, pp. 870–885.

[29] P. M. Dickens and J. Logan, "A high performance implementation of MPI-IO for a Lustre file system environment," *Concurrency and Computation: Practice and Experience (CCPE)*, vol. 22, pp. 1433–1449, September 2009.

[30] L. Grandinetti, G. Joubert, and M. Kunze, *Big Data and High Performance Computing*, ser. Advances in Parallel Computing. IOS Press, 2015.

[31] Lustre, "Lustre file system," *http://lustre.org/documentation/*.

[32] W. Cirne, F. Brasileiro, D. Paranhos, L. F. W. Góes, and W. Voorsluys, "On the efficacy, efficiency and emergent behavior of task replication in large distributed systems," *Parallel Comput.*, vol. 33, no. 3, pp. 213–234, April 2007.

[33] G. D. Ghare and S. T. Leutenegger, "Improving speedup and response times by replicating parallel programs on a snow," in *Proceedings of the 10th International Conference on Job Scheduling Strategies for Parallel Processing*, ser. JSSPP'04. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 264–287.

[34] Z. Qiu, J. F. Pérez, and P. G. Harrison, "Tackling latency via replication in distributed systems," in *Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering*, March 2016.

[35] D. Wang, G. Joshi, and G. Wornell, "Efficient task replication for fast response times in parallel computation," in *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, 2014, pp. 599–600.

[36] Top500, "Top500 list - june 2015," *https://www.top500.org/list/2015/06/*.

[37] Altair, "PBS professional," *https://www.altair.com/pbs-professional/*, February 2020.

[38] T. A. Davis and Y. Hu, "The University of Florida sparse matrix collection," *ACM Trans. Math. Softw.*, vol. 38, no. 1, December 2011.

[39] Fujitsu, "Fujitsu Exabyte file system (FEFS)," *https://www.fujitsu.com/downloads/TC/sc11/fefs-sc11.pdf*.

[40] S. Ghosh, J. R. Hammond, A. J. Peña, P. Balaji, A. H. Gebremedhin, and B. Chapman, "One-sided interface for matrix operations using MPI-3 RMA: A case study with Elemental," in *2016 45th International Conference on Parallel Processing (ICPP)*, 2016, pp. 185–194.

[41] X. S. Li and J. W. Demmel, "SuperLU DIST: A scalable distributed-memory sparse direct solver for unsymmetric linear systems," *ACM Trans. Math. Softw.*, vol. 29, no. 2, pp. 110–140, June 2003.

[42] L. Dagum and R. Menon, "OpenMP: An industry standard API for shared memory programming," *IEEE Computational Science and Engineering*, pp. 46–55, January 1998.

[43] D. S. Hirschberg, A. K. Chandra, and D. V. Sarwate, "Computing connected components on parallel computers," *Commun. ACM*, vol. 22, no. 8, pp. 461–464, August 1979.

[44] A. P. Reeves, "Parallel pascal: An extended pascal for parallel computers," *Journal of Parallel and Distributed Computing*, vol. 1, no. 1, pp. 64 – 80, 1984.

[45] H. Hussain, S. U. R. Malik, A. Hameed, S. U. Khan, G. Bickler, N. Min-Allah, M. B. Qureshi, L. Zhang, W. Yongji, N. Ghani, J. Kolodziej, A. Y. Zomaya, C.-Z. Xu, P. Balaji, A. Vishnu, F. Pinel, J. E. Pecero, D. Kliazovich, P. Bouvry, H. Li, L. Wang, D. Chen, and A. Rayes, "A survey on resource allocation in high performance distributed computing systems," *Parallel Computing*, vol. 39, no. 11, pp. 709 – 736, 2013.

[46] D. Merrill and M. Garland, "Merge-based parallel sparse matrix-vector multiplication," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2016, pp. 58:1–58:12.

[47] B. Klenk and H. Fröning, "An overview of MPI characteristics of exascale proxy applications," in *High Performance Computing*, J. M. Kunkel, R. Yokota, P. Balaji, and D. Keyes, Eds. Springer International Publishing, 2017, pp. 217–236.

[48] R. Rabenseifner, G. Hager, and G. Jost, "Hybrid MPI/OpenMP parallel programming on clusters of multi-core SMP nodes," in *2009 17th Euromicro International Conference on Parallel, Distributed and Network-based Processing*, Feb 2009, pp. 427–436.

[49] F. J. M.-Z. . J. R. P. Alonso, R. Cortina, "Neville elimination on multi- and many-core systems: OpenMP, MPI and CUDA," *Journal of Supercomputing*, pp. 215–225, 2011.

[50] D. D. Nikolić, "Parallelisation of equation-based simulation programs on heterogeneous computing systems," *PeerJ Computer Science*, 2018.

[51] L. Dongha, O. Jinoh, and Y. Hwanjo, "OCAM: Out-of-core coordinate descent algorithm for matrix completion," *INFORMATION SCIENCES*, vol. 514, pp. 587 – 604, April 2020.

Deep Learning Approach for Forecasting Water Quality in IoT Systems

Nguyen Thai-Nghe[1]
College of ICT
Can Tho University
Can Tho City, Vietnam

Nguyen Thanh-Hai[2]
College of ICT
Can Tho University
Can Tho City, Vietnam

Nguyen Chi Ngon[3]
College of Engineering Technology
Can Tho University
Can Tho City, Vietnam

Abstract—**Global climate change and water pollution effects have caused many problems to the farmers in fish/shrimp raising, for example, the shrimps/fishes had early died before harvest. How to monitor and manage quality of the water to help the farmers tackling this problem is very necessary. Water quality monitoring is important when developing IoT systems, especially for aquaculture and fisheries. By monitoring the real-time sensor data indicators (such as indicators of salinity, temperature, pH, and dissolved oxygen - DO) and forecasting them to get early warning, we can manage the quality of the water, thus collecting both quality and quantity in shrimp/fish raising. In this work, we introduce an architecture with a forecasting model for the IoT systems to monitor water quality in aquaculture and fisheries. Since these indicators are collected every day, they becomes sequential/time series data, we propose to use deep learning with Long-Short Term Memory (LSTM) algorithm for forecasting these indicators. Experimental results on several data sets show that the proposed approach works well and can be applied for the real systems.**

Keywords—*Forecasting model; deep learning; Long-Short Term Memory (LSTM); water quality indicators*

I. Introduction

Global climate change and Global Warming effects have caused many problems to the farmers and the managers. Climate change not only causes great impacts on all regions, resources, environment and socio-economic, but also affects water resources, agriculture and rural development, health and the coastal areas. Economic losses due to the impact of climate change combined with the costs of recovering from the damage reduce the economic growth of many countries and globally. Especially, Agricultural/Aquaculture countries have been suffering great losses because of Global climate change. For example, fishery is one of the key economic sectors of Vietnam's agriculture, with billions of dollars in exports. It also creates jobs for millions of workers, contributing to stabilizing social security and developing the country. Shrimp is a seafood product in the top export of billion US dollars. However, shrimp nursery is too difficult because they are particularly sensitive to water. Polluted water environment is a great concern for numerous fishermen and Agricultural managers. However, these farmers do not have enough technique and media to forecast environment factors, so they cannot manage risks which they are facing. Thus, how to manage and monitor the water quality, especially in aquaculture and fisheries such as fish raising management is really necessary.

Recently, techniques in the Internet of Things (IoT) can help users to build the monitoring systems. IoT refers to the millions or even billions of physical devices around the world currently connected to the internet, collecting and sharing data together. With advancements in internal processors and wireless networks, we now can make everything more proactive and intelligent [1]. There are many applications of IoT such as automatic door systems (smart-home), aircraft and self-driving cars (smart-city), etc. which have become a common part of IoT. This adds a level of digital intelligence to passive devices, allowing us to communicate real-time data without human involvement, effectively integrating the digital and physical world. When something is connected to the internet, it is able to send or receive information, or both. Leveraging IoT techniques, the ability to send or receive this information makes everything smarter, and that is also our target.

This work[1] proposes an architecture for the IoT systems to monitor and forecast the water quality (such as indicators of salinity, temperature, pH, and dissolved oxygen - DO) in aquaculture and fisheries. These data are collected every day and ordered by time, they can be considered as time series or sequential data. Previous works show that deep learning with LSTM (Long-Short Term Memory) could be an appropriate method for time series data [2], thus, we propose deep learning with LSTM models for forecasting these water quality indicators.

The remaining of this study is presented as follows. We exhibit some state-of-the-art approaches in sensor data processing in Section II. Section III introduces the proposed architectures for the forecast problems. In Section IV, we present the data sets used in the experiments and do some investigations, comparisons, explain the experimental results. Finally, we discuss and summarize our work in Section V.

II. Related Work

A complete IoT system relates to several sides such as sensor networks, hardware circuits, energy, etc., and software for analysis. One of important tasks in the environmental monitoring IoT system is the data analyzing and forecasting methods. In this work, we focus on reviewing the literature which relates to the forecasting methods. There are several researches in this field.

The authors in [3] introduced a method to develop a service oriented architecture for the weather information systems which forecast weather using these data mining techniques

---

[1]Corresponding author: Nguyen Chi Ngon

(e.g., Artificial Neural Network and Decision tree). The study in [4] proposed a deep neural network for weather forecasting with a set of weather-related variables. The study in [5] used recurrent neural networks (RNN) with its Long Short-Term Memory (LSTM) architecture to predict the ambient temperature (TA). The prediction is based on meteorological data (e.g., temperature, humidity) retrieved from IoT stations and send them to the basic station with LoRa protocol. They formulate the TA prediction problem as a time series regression problem. The proposed approach uses two types of hidden layers: LSTM layer and full connected dense layer. Experimental results show that the model work well.

In [6], the water, temperature, pH and DO levels are measured and integrated with aerating and water supply pumps using Arduino. The user could receive information at predetermined intervals on preferred communication. The authors tested on a sample of two days measurements of temperature, pH and DO levels. Results show that with this integration system, farmer need not hire worker at their site, consequently drive down operating costs and improve efficiency.

The authors in [7] introduced an online water quality monitoring system for intensive fish culture in China. Based on historical data, the system forecasts the water quality using artificial neural networks (ANNs) and control the water quality in time to reduce catastrophic losses. Their results demonstrate that multi-parametric, long-distance and online monitoring for water quality information can be accurately acquired and predicted.

The authors in [9] proposed the detection and monitoring of emotion applications. The authors introduce a deep learning approach for emotion classification through an iterative process by adding and removing large number of sensor signals from different modalities. They uses Convolutional Neural Network and Long Short-term Memory Recurrent Neural Network on the raw sensor data. The results show that their proposed approach is effective in human emotion classification when large number of sensors input is utilised and the hybrid models outperform traditional fully connected deep neural network.

In [10], the authors proposed a deep learning and image-based model for air quality estimation. The model extracts feature information from scene images captured by camera equipment and then classifies them to estimate air quality levels. The authors tested on several methods such as Support Vector Machines and Deep Residual Network on a high-quality outdoor air quality data set. Experimental results show that the proposed method produces more accurate results than other methods.

The authors in [11] used deep-learning technologies in the environmental field to predict the status of pro-environmental consumption. The authors predicted the pro-environmental consumption index based on Google search query data, using a recurrent neural network (RNN) model. They compared the prediction accuracy of the RNN model with that of the ordinary least square and artificial neural network models. The RNN model predicts the pro-environmental consumption index better than any other model.

In [12], the authors presented the Public agencies aiming to enforce environmental regulation have limited resources to achieve their objectives. They demonstrate how machine-learning methods can inform the efficient use of these limited resources while accounting for real-world concerns, such as gaming the system and institutional constraints. They predict the likelihood of a facility failing a water-pollution inspection and propose alternative inspection allocations that would target high-risk facilities.

The authors of [13] developed a web-based system which monitor the pH and salinity values. The workings of the monitoring and control system send data on pH and salinity sent by the sensor and then stored and in database hosting. They said that for the value of a dangerous salt content worth less than 160 and more than 210, if for a pH of less than 6.5 and more than 7.5.

Darmalim et al. in [14] proposed an IoT solution to automatically monitor the environmental factors. The system had five sensors to measure each parameter. It is developed using a Pythonframework with a web application to present information from the IoT device. The work in [16] presented a feasible model for the daily average temperature on the area of Zhengzhou and apply it to weather derivatives pricing. Other researches can be found in [8], [17], [18], [19], [20].

Although many researches have been done in the topic of environmental forecasting, using LSTM for forecasting water quality indicators is not seen recently and this problem still has room for improvements. In this work, we propose an architecture for the IoT systems to monitor water quality (such as indicators of salinity, temperature, pH, and DO). Since these data are collected every day (ordered by time), we propose deep learning with Long-Short Term Memory (LSTM) models for forecasting these time series data.

III. PROPOSED METHOD

A. System Architecture

The system architecture is presented in Fig. 1. In this system, we have implemented the sensors for monitoring water quality indicators such as salinity, temperature, pH, and dissolved oxygen (DO) in the fish/shrimp ponds. Data from sensors is daily transferred to a cloud database (in this work, we have used the cloud database provided by the DigitalOcean[2]). This data is automatically transferred to a server for analyzing and forecasting.

Using this system, the users can visualize all of the indicators via mobile devices or pc/laptop. The system automatically sends the messages to the users (farmers) when the value of an indicator exceeds the limit (threshold) value. Moreover, based on historical data, the system can forecast the values of each indicator for the next dates (or other selected time).

The system infrastructure is implemented by another group in our team. In this paper, our task is to develop a forecasting model which can provide the forecast values to send early warning to the users. Thus, in the rest of this work, we focus on developing the forecasting models, especially Deep Learning with LSTM for forecasting time series data.

---

[2]https://www.digitalocean.com

Fig. 1. System Architecture for Shrimp/Fish Pond Monitoring

B. The Proposed LSTM Architectures

LSTM (Long Short-Term Memory) is well-known deep learning method which widely-used in forecast models and can be used to avoid the long-term dependency problem [2], [15].

We propose the LSTM architecture as presented in Fig. 2. For testing purposes, the proposed model includes a single LSTM layer with four units followed by a Fully Connected layer using ReLU as the activation function (the leftside of Fig. 2) and another network containing two recurrent hidden states with 16 units for the first LSTM layer and 4 units for the second LSTM layer which we call stacked LSTM (the rightside of Fig. 2). The input to these models are a sequence values (based on timestamp) of each indicator (e.g., salinity, pH, DO, and Temperature) and the output are their forecast values. We implemented the LSTM with Adam optimizer [21] as the optimized function. The learning rate used is initiated at 0.001. In the experiments (will be presented in Section IV-C) we will validate to choose which LSTM model is appropriate for the kind of sensor data.

When we look inside the LSTM (shown in Fig. 3 as an example), its major and important part of a LSTM network is cells that provide a bit of memory to the LSTM so it can remember the past. LSTM usually contains three types of gates including Input Gate, Output Gate and Forget Gate. Gates in LSTM consist of sigmoid activation functions which produce output ranging from 0 to 1. In most of the cases, the value is either 0 or 1. The sigmoid function (Equation 1) is used commonly for gates because we want a gate to give only positive values and should be able to give us a clear cut answer whether, we need to keep a particular feature or we need to discard that feature.

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{2}$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{3}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{4}$$

Where:

Fig. 2. The Proposed Single LSTM (Left Side) and Stacked LSTM (Right Side) Models

Fig. 3. An Illustration of LSTM Cell [15].

- i_t denotes input gate.

- f_t reveals output gate.

- o_t represents forget gate.

- w_m shows weights for the respective $gate(m)$.

- h_{t-1} denotes the output of the preceded LSTM block at the timestamp t-1.

- x_t is the input at the current timestamp.

- b_m exhibits the biases of the respective $gate(m)$.

Equation 2 reveals what new information will be stored in the cell state while Equation 4 aims to provide the activation to the final output of the LSTM block at timestamp t. The forget gate at Equation 3 give us the information on which information to throw away from the cell state.

IV. EXPERIMENTAL RESULTS

A. Data Sets for the Experiments

In the proposed architecture in Fig. 1, the sensors' implementation is deployed by another group in our team. However, this system is just started and we do not have enough data for training the model. Luckily, there are several published data sets (presented in Table I) which have the structures similar to our data structure, so we have used them to test the proposed forecasting approach.

The first data set is published by Australian Government, which is the sensor data of Temperature, pH and salinity at New South Wales Estuary.

The second data set is published by the BCO-DMO, located at the Woods Hole Oceanographic Institution, US. This data set relates to water quality parameters (e.g., pH, salinity, temperature, and dissolved oxygen - DO) in Tomales Bay, CA (38.148N, 122.898W) which is a monthly time series through sampling every 2 km.

The third data set is published by the University of Melbourne, Australia. This data set relates to an IoT system which provides the Sensor data including of temperature, light level, humidity every 5 minutes at 8 locations from 2014 to 2015.

The fourth and fifth data sets are published by the Government of Canada. These data relate to the Surface temperature and salinity measured along the track of commercial ships, mostly between Montreal (Quebec) and St. John's (Newfoundland) from 1999 to 2018. We have just used the data of 2017 and 2018 for testing.

The sixth data set is published by the European Multidisciplinary Seafloor and water column Observatory. This dataset contains temperature and dissolved oxygen concentrations and the associated sensor raw data acquired between July 2017 and August 2018 on EMSO-Azores observatory by the EGIM.

After pre-processing by removing missing data, the remaining number of observations are presented in the column #Records in Table II.

TABLE I. PUBLISHED DATA SET RESOURCES

No.	Data set name	Link (accessed July, 2020)
1	NSW Estuary Temperature, pH and salinity data	https://data.gov.au/dataset/ds-nsw-c2041218-0496-453a-ac22-9e974fdebf89/distribution/dist-nsw-5154e497-647f-4e2a-a2ad-151fa293b0da/details?q=
2	Water quality - Tomales Bay, CA	https://www.bco-dmo.org/dataset-deployment/455476
3	Melbourne - Sensor readings	https://data.melbourne.vic.gov.au/Environment/Sensor-readings-with-temperature-light-humidity-ev/ez6b-syvw
4	Fisheries & Oceans Canada 2018	https://open.canada.ca/data/en/dataset/8a3dc9e5-f3af-4270-8c09-43fa2c25848b
5	Fisheries & Oceans Canada 2017	https://open.canada.ca/data/en/dataset/8a3dc9e5-f3af-4270-8c09-43fa2c25848b
6	Temperature and DO from the EGIM, EMSO-Azores observatory, 2017-2018	https://www.seanoe.org/data/00453/56501

B. Evaluation Metrics and Baseline

Root mean squared error (RMSE) is used to evaluate the models. The RMSE for n observations is calculated by equation (5).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2} \qquad (5)$$

where, p_i is the true value, and $\hat{p}_i$ is the forecasted value.

We have used regression version of the Support Vector Machines (SVM) as a baseline for comparison.

C. Experimental Results

Fig. 4. Relationship between the RMSE and the Number of Neurons (Salinity Indicator on Data Set 2) using LSTM. Other Data Sets are nearly Similar

In the first group of experiments, a single layer of the LSTM (model in the left side of Fig. 2) is performed to select the number of hidden neurons. Results show that, on these data sets, the RMSE is not much different when using the number of hidden neurons from 4 to 60. An example is presented in

Fig. 4. Thus, we have selected four hidden neurons for faster training.

Fig. 5. Relationship between the RMSE and the Number of Epochs (on Small Data Sets) using LSTM.

Fig. 6. Relationship between the RMSE and the Number of Epochs (on Larger Data Sets) using LSTM.

The second group of experiments is performed to choose the number of optimizing epochs. Results show that the RMSE is the best with 10 epochs for small data sets (data set 1 and data set 2) as presented in Fig. 5. However, the number of epochs should be increased for larger data sets (data sets 3,4,5, and 6), i.e., from 60 epochs to 80 epochs as an example in Fig. 6.

In the third group of experiments, we have revised the single LSTM model by adding more layers, which is called stacked LSTM (e.g., model in the right side of Fig. 2). Results show that the RMSE can be improved when using stacked LSTM as presented in Fig. 7. However, the RMSE is not significantly improved by adding more than 2 layers to the LSTM model, thus we have used the model with 2 layers as proposed in the right side of Fig. 2 for the rest experiments.

Fig. 7. RMSE Comparison between LSTM and Stacked-LSTM on Data Set 6. Other Data Sets are also Similar

The final results are presented in Table II. There are four considered indicators which include salinity, pH, temperature, and dissolved oxygen (DO). As shown from Table II, most of the cases, the LSTM has lower error (RMSE) than the baseline SVM, thus it can produce better forecasting results.

Fig. 8 and Fig. 9 visualize the true values and the forecast values of the salinity and the pH indicators on small data set while Fig. 10 and 11 visualize the forecast values of the temperature and the DO indicators on lager data set, respectively. The result shapes of other datasets are also similar when visualizing.

These results show that the forecast values of the LSTM model are nearly approached to the true values, thus the proposed model could be appropriate for sensor data with time series, and it could be applied to the real world. These results also consist with other works in this domain [5], [7], [9], [11].

V. CONCLUSION

In this work, we proposed an IoT system for monitoring water quality in aquaculture and fisheries, especially the model for forecasting the quality indicators such as salinity, pH, temperature, and DO. Experimental results on the considered data sets with four indicators show that the proposed approach can be applied for the real system. By monitoring these real-time indicators and getting early warning, the system can help the farmers to manage the quality of the water, thus collecting both quality and quantity in shrimp/fish raising.

In the future, we will collect the data provided in our system to retrain the model for implementing to the real world. In this work, we deployed a simple LSTM to perform the forecasting tasks. Some sophisticated models should be taken into account to improve the forecast performance in further researches. The LSTM model in this work uses one variable (the indicator) for forecasting, however, by utilizing other features/variables in the dataset to adapt the LSTM to Multi-Variate LSTM forecasting models could be an approach for improving the forecasting performance.

TABLE II. RMSE COMPARISON BETWEEN LSTM AND SVM. THE RESULTS FORMATTED IN BOLD AND RED ARE THE BEST RESULTS CORRESPONDING THE EACH INDICATOR.

No.	Data set	Indicator	#Records	RMSE	
				SVM	LSTM
1	NSW Estuary Temperature, pH and salinity data	Salinity	6,244	9.0773	**8.2611**
		pH	3,358	0.4709	**0.4396**
		Temperature	6,261	1.9212	**1.8062**
2	Water quality - Tomales Bay, CA	pH	246	0.1953	**0.1922**
		Salinity	229	3.0744	**2.8755**
		Temperature	229	1.6553	**1.6150**
		Percent DO (O2Sat)	246	11.7498	**11.7002**
		DO (O2Mg)	246	1.0571	**1.0482**
3	Melbourne - Sensor readings	Temperature	56,570	1.9700	**1.1922**
		Humidity	56,570	**7.0900**	7.5474
4	Fisheries & Oceans Canada 2018	Salinity	33,835	0.3786	**0.3640**
		Temperature	34,047	**0.3419**	0.3423
5	Fisheries & Oceans Canada 2017	Salinity	30,714	0.4350	**0.4274**
		Temperature	30,714	0.3112	**0.3104**
6	Temperature and DO from the EGIM, EMSO-Azores observatory, 2017-2018	DO	38,657	4.2445	**4.1971**
		Temperature	38,657	0.6095	**0.6079**

ACKNOWLEDGMENT

REFERENCES

[1] Shafique, Kinza & Khawaja, Bilal & Sabir, Farah & Qazi, Sameer & Mustaqim, Muhammed. (2020). Internet of Things (IoT) For Next-Generation Smart Systems: A Review of Current Challenges, Future Trends and Prospects for Emerging 5G-IoT Scenarios. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2970118.

[2] Lim, B., & Zohren, S. (2020). Time Series Forecasting With Deep Learning: A Survey. arXiv preprint arXiv:2004.13408.

[3] ZhanJie Wang and A. B. M. Mazharul Mujib. (2017). The Weather Forecast Using Data Mining Research Based on Cloud Computing. Phys.: Conf. Ser. 910 012020.

[4] Aditya Grover, Ashish Kapoor, and Eric Horvitz. (2015). A Deep Hybrid Model for Weather Forecasting. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). Association for Computing Machinery, New York, NY, USA, 379–386. DOI:https://doi.org/10.1145/2783258.2783275

[5] Ben Abdel Ouahab Ikram, Boudhir Anouar Abdelhakim, Astito Abdelali, Bassam Zafar, and Bouhorma Mohammed. (2019). Deep Learning architecture for temperature forecasting in an IoT LoRa based system. In Proceedings of the 2nd International Conference on Networking, Information Systems & Security (NISS19). Association for Computing Machinery, New York, NY, USA, Article 43, 1–6. DOI:https://doi.org/10.1145/3320326.3320375

[6] Harun, Zambri & Reda, Eslam & Hashim, Harris. (2018). Real time fish pond monitoring and automation using Arduino. IOP Conference Series: Materials Science and Engineering. 340. 012014. 10.1088/1757-899X/340/1/012014.

[7] Li D., Liu S. (2013). Remote Monitoring of Water Quality for Intensive Fish Culture. In: Mukhopadhyay S., Mason A. (eds) Smart Sensors for Real-Time Water Quality Monitoring. Smart Sensors, Measurement and Instrumentation, vol 4. Springer, Berlin, Heidelberg

[8] Dieu Tien Bui, Khabat Khosravi, John Tiefenbacher, Hoang Nguyen, Nerantzis Kazakis. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms, Science of The Total Environment, Volume 721, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2020.137612.

[9] Eiman Kanjo, Eman M.G. Younis, and Chee Siang Ang. (2019). Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. Information Fusion, Volume 49, Pages 46-56, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2018.09.001.

[10] Qiang Zhang, Fengchen Fu, and Ran Tian. (2020). A deep learning and image-based model for air quality estimation. Science of The Total Environment, Volume 724, 138178, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2020.138178.

[11] Lee D, Kang S, Shin J. (2017). Using Deep Learning Techniques to Forecast Environmental Consumption Level. Sustainability. 9(10):1894.

[12] Hino, M., Benami, E. & Brooks, N. (2018). Machine learning for environmental monitoring. Nat Sustain 1, 583–588. https://doi.org/10.1038/s41893-018-0142-9

[13] Preetham, K., Mallikarjun, B. C., Umesha, K., Mahesh, F. M., & Neethan, S. (2019). Aquaculture monitoring and control system: An IoT based approach. International Journal of Advance Research, Ideas and Innovations in Technology, 5(2).

[14] Darmalim, U., Darmalim, F., Darmalim, S., Hidayat, A. A., Budiarto, A., Mahesworo, B., & Pardamean, B. (2020). IoT Solution for Intelligent Pond Monitoring. E&ES, 426(1), 012145.

[15] Harini Suresh. The Vanishing Gradient Problem. http://harinisuresh.com/2016/10/09/lstms/, accessed on August 20, 2020.

[16] Wang, Z., Li, P., Li, L., Huang, C., & Liu, M. (2015). Modeling and forecasting average temperature for weather derivative pricing. Advances in Meteorology, 2015.

[17] Licata, R.J., Tobiska, W.K., & Mehta, P.M. (2020). Benchmarking Forecasting Models for Space Weather Drivers. arXiv: Space Physics.

[18] Dong, Jianhua & Wang, Guoyin & Yan, Huyong & Xu, Ji & Zhang, Xuerui. (2015). A survey of smart water quality monitoring system. Environmental Science and Pollution Research. 22. 10.1007/s11356-014-4026-x.

[19] Kachroud, Moez & Trolard, Fabienne & Kefi, Mohamed & Jebari, Sihem & Bourrié, Guilhem. (2019). Water Quality Indices: Challenges and Application Limits in the Literature. Water. 11. 1-26. 10.3390/w11020361.

[20] Siddiqui, Nihal & Tauseef, S.M. & Dobhal, Rajendra. (2020). Advances in Water Pollution Monitoring and Control Select Proceedings from HSFEA 2018: Select Proceedings from HSFEA 2018. 10.1007/978-981-32-9956-6.

[21] Diederik P. Kingma and Jimmy Lei Ba. Adam : A method for stochastic optimization. (2014). arXiv:1412.6980v9

Fig. 8. Forecasting Visualization for the pH in Dataset 2. The Blue Line is the True Values, the Orange Line is the Predicted Values on Training Set (67%), and the Green Line is the Predicted Values on Test Set (33%)

Fig. 9. Forecasting Visualization for the Salinity in Dataset 2. The Blue Line is the True Values, the Orange Line is the Predicted Values on Training Set (67%), and the Green Line is the Predicted Values on Test Set (33%)

Fig. 10. Forecasting Visualization for the Temperature in Dataset 1. The Blue Line is the True Values, the Orange Line is the Predicted Values on Training Set (67%), and the Green Line is the Predicted Values on Test Set (33%)

Fig. 11. Forecasting Visualization for the Dissolved Oxygen (DO) in Dataset 6. The Blue Line is the True Values, the Orange Line is the Predicted Values on Training Set (67%), and the Green Line is the Predicted Values on Test Set (33%)

A Complete Methodology for Kuzushiji Historical Character Recognition using Multiple Features Approach and Deep Learning Model

Aravinda C.V[1]
Department of C.S.E
N.M.A.M Institute of Technology
Nitte, Karnataka
INDIA-574110

Lin Meng[2] and ATSUMI Masahiko[3]
Department of E.C.E
Ritsumeikan University,
Kusatsu
Japan

Udaya Kumar Reddy K.R[4]
and Amar Prabhu G[5]
Department of C.S.E
N.M.A.M Institute of Technology
Nitte, Karnataka
INDIA-574110

Abstract—As per the studies during many decades, substantial research efforts have been devoting towards character recognition. This task is not so easy as it it appears; in fact humans' have error rate about more than 6%, while reading the handwritten characters and recognizing. To solve this problem an effort has been made by applying the multiple features for recognizing kuzushiji character, without any knowledge of the font family presented. At the outset a pre-processing step that includes image binarization, noise removal and enhancement was applied. Second step was segmenting the page-sample by applying contour technique along with convex hull method to detect individual character. Third step was feature extraction which included zonal features (ZF), structural features (SF) and invariant moments (IM). These feature vectors were passed for training and testing to the various machine learning and deep learning models to classify and recognize the given character image sample. The accuracy achieved was about 85-90% on the data-set which consisted of huge data samples round about 3929 classes followed by 392990 samples.

Keywords—*Kuzushiji character; zonal features; structural features; invariant moments*

I. Introduction

The huge quantity of data, either modern or historical, we have in our occupancy nowadays, because of expansions of digital libraries for reliable and accurate systems for processing. These documents are important because they are more significant part of our cultural heritage. Countless commercial products have been present and getting released to convert digitized documents into text files, either in the Unicode or ASCII format [1]. Unfortunately these products process only machine printed documents successfully, but if these machines fails when it comes to handwritten documents especially historical documents which results in poor performance [2] [3]. To solve this problems, recognition of historical documents is one of the most challenging tasks in recent days. The Japanese writing pattern is an arousing curiosity study of innovation and tradition. It combines a set of Chinese logo grams and two Chinese-derived syllabaries into a complex logo syllabic system. Scripting evolved to Japan from China during the 5th century. The first Japanese characters were written in Chinese characters (kanji), a system called kanbun [14]. The Japanese language, has incurred verbs and post positions, requiring concatenation of suffixes and particles to words and clauses in a sentence [9]. In order to overcome the grammatical units, the Japanese historical used certain Chinese characters for their sound values. This lead to the system which was ambiguous, and hard to tell whether a character [5].

In term of ancient Japanese documents, people used kuzushiji character for writing documents and publishing books. These ancient documents and books are found one by one currently, and waiting to be understood, which store a larger number of potential knowledge. However, few people know the kuzushiji character currently [7] [13]. And the kuzushiji characters have many variation, sometimes characters are connected one by one. Furthermore, aging causes the documents are uncleanness and worm-eaten happened. These problems increase the difficult of understanding these ancient documents. Especially, the traditional OCR(Optical Character Recognition) can not achieve better performance for the kuzushiji characters recognition [8] [12].

The problems in automatic recognition for totally unconstrained handwritten characters are greater than that of printed characters. Mistakes in reading the handwritten characters are more rates than of printed ones [4] [6].

A common OCR system will have several components, which illustrate the organization of usage. The input documents are scanned by external devices to produce gray-level image or binary bit-mapped image [10] [11]. This computational technique is knows as threshold.The flow of work is as shown in Fig. 1

A. Problem Statement

There are several problem related to handwritten historical kuzushiji character. The most important problems are notified and mentioned below

1) **Shape Discrimination:** A single character has variety of font style and a lot of free flow styles.
2) **Distorting of the character:** This is mainly because of the following reasons mentioned below
   - **Noise:** The reason is disconnected line segments, breaks in lines, isolated dots.
   - **Translation:** This is for the movement of full character or its elements.

- **Rotation:** This changes in orientation.

3) **Size variations:** The area of the character may be of 10, 15, 20 which specifies that are 10,15,20 characters per inch. 10 pitch are usually bigger than width and height than in 15. In continuation of these problems characters which are proportional spacing and variable line spacing.

II. SEGMENTATION PROCESS

At first the document was binarized a top-down segment approach. Next lines of the documents were detected followed by segmentation of individual characters is as shown in Fig. 3.

A. Pre-Processing

First the image was converted from 3d color image to gray-scale image is as shown in Fig. 4. Next the same image was binarized for black and white using Otsu threshold technique is as shown in Fig. 5.The binarized image was dialated using 20*20 kernel of 1s. The intention of this step for kuzushiji characters was applied mainly because of disconnected characters in document. For the better enhancement of the segmentation process the disconnected regions of the single character needs to be combined to make it as one character for recognition. The produced image was blurred by applying Gaussian Filter technique for noise removalis as shown in Fig. 2.

B. Canny Edge Detection

The Canny edge detector is an edge detection operator that uses a multi-stage algorithm to detect a edges in images. The filtered image was skeletonized by applying canny edge detection algorithm. The advantage of applying this algorithm for segmentation resulted to Noise reduction, Gradient calculation, Non-maximum suppression, Double threshold, Edge Tracking by Hysteresis for kuzushiji character is as shown in Fig. 6.

C. Contour Technique

The segmentation is one of the most challenging task in the tedious process for the separation or segregation of required information in the character. The contour is one of the best active models in segmentation process, for separation of region of interest.This active contour segmentation was used for the separation of pixels of interest for character segmentation is as shown in Fig. 7.

D. Convex Hull Technique

After applying the contour as mentioned in the Section II-C, next step was processed by applying the convex hull algorithm for better enhancement. The advantage of this algorithm was tend to be useful for the proposed problem. By using this convex hull algorithm a set of points was identified as the smallest convex polygon which enclosed all the points, since the character was curve in shape. The polygon was sketched around each individual character which enclosed all the points calculated in the previous Section II-C belonging to the individual character as shown in the Fig. 7 and 8. The draw back of convex hull found to be from the result was observed

Fig. 1. Flow of Block Diagram for OCR

Fig. 2. Input of the Image

that convex hull gives an approximate bounding polygon which enclosed the entire character. As per the problem statement this was not feasible,since the area of the character should get segmented in the form of bounding box. To achieve this a bounding box was constructed form the points obtained after applying the convex hull technique as mentioned. A bounding box was drawn using 4 quadrant properties which are X and Y coordinates of top left corner, width and height of the bounding

Fig. 3. Sub Section of Page-Labelled

Fig. 6. Canny Edge Detection

Fig. 4. Gray Scale

Fig. 7. Individual Character for Contour Technique

Fig. 5. Binary Image

Fig. 8. Contour and Convex Hull of Character

box form the point of X, Y is as shown in Fig. 9.

E. Threshold

As mentioned in the previous Section II-D by applying the bounding box to all the obtained convex hull, which crops

Fig. 9. Contour and Convex Hull for Sub-Section of Page

Fig. 11. Flow of Segmentation Technique

Fig. 10. Bounding Box

character was considered as as separate array and k-means clustering were applied on the array separately with 2 clusters under the assumption that all the characters, small radicals' and noises will be grouped as 1 cluster. Later all the combined characters and some of the individual character will be grouped in 2nd cluster. The minimum of the cluster center were taken as the height threshold and width threshold respectively.The complete flow of segmentation technique is as shown in the Fig. 11

$$T_h = min\left(cluster\_center\left(height\right)\right)$$
$$T_w = min\left(cluster\_center\left(width\right)\right)$$

for every bounding box do:
if height $> T_h$ and width $> T_w$
crop character
else
ignore.
Finally, all the characters whose height was shown greater than $1/4^{th}$ * page height were ignored. Hence the threshold condition can be modified as,
height of bounding box $> T_h$
width of bounding box $> T_w$
height of bounding box $< 1/4^{th}$ * page height.
Experimentation found that this method extracted at-least 90% of the area of interest (Individual character)

III. RECOGNITION PROCESS FLOW OF WORK CARRIED OUT

A. Data Set Description

The huge data-sets were considered which consists of about 3929 classes kuzushiji characters for the model selection experiment.The total number of training images was more than 600,000 images. The kuzushiji data-set are obtained from Center for Open Data in the Humanities (CODH). Out of this there were three types of data-sets: 1)Katakana, 2) Kanji, 3) Hirangana.
The datasets was initially 64*64 image sample size which consisted of white text written on black background. The data-set was inverted with respect to color to get black text on white background which was resized to 100 * 100 for experimental purposes. Morphological opening and closing were carried out to remove salt and pepper noise on the image.

B. Feature Extraction

To recognize the particular character four types of features were extracted from the training set image:

the area of interest of individual character along with minor distortion in the given page data sample and also the radicals' of the character is as shown in Fig. 11. However, the minor distortion and radicals' need to be ignored for recognition purpose. This critic point made to apply threshold for the said problem. In continuation of analysing the page samples it was found that mainly three types of regions from page were cropped. The representation of the experiment visual is as shown in the Fig. 10.

- The area of interest (Individual Character).

- The small distortion and noise along with radicals' whose width and height were smaller than the area of interest (Individual Character).

- The combination of characters written as 1 character whose width and height were much bigger than the area of interest.

F. K-means Clustering

To solve the threshold problem as mentioned in the previous Section II-E the width and height of the individual

1) Zonal feature: The given character image was divided into zones of equal width and height of the image. This image was divided into 25 zones where each

Fig. 12. Zonal Feature Projection Profile

Fig. 13. Distance Feature Vertical Projection Profile

Fig. 14. Distance Feature Horizontal Projection Profile

zone had width of 20 pixel and height of 20 pixel which makes 400 pixel in each zones.

The density of the black pixels in each zone were calculated in order to construct a feature vector histogram. From the zonal features a histogram of 25 bins were obtained which described the density in 25 zones of the image is as shown in Fig. 12.

2) Distance Feature: The distance feature was then extracted both from horizontal and vertical direction of the image is as shown in Fig. 13. **Vertical Profile:** A centroid was calculated in the vertical direction which was image width/2 for the mentioned problem statement 50 units was calculated is as shown in Fig. 14. The image was divided into two sections based on this centroid namely left section and right section. Each section was then subdivided into 10 subsections. In each of the subsection, Euclidean distance between the furthest pixel in the sub-section $(i.e, pixel - closest - to - the - outer - boundary)$ and the center point of the centroid for that section was found. $distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ This created a 20 bin histogram which was appended to the feature vector.

3) Invariant Moments: Similar to Vertical profile, a centroid was found in the horizontal direction with $c_h = imageheight/2$, which resulted in 50 units. The image was divided into 2 sections upper X and lower section Y, these sections were divided into 10 subsections to find the distance between the furthest black pixel and cnetroid of subsection. This created a 20 bin histogram which was appended to the feature vector array.

4) Hu Moments: By using this feature moments 7 bin histogram was created. The total size of the feature vector resulted in $25 + 20 + 20 + 7 = 72bins$.

IV. CLASSIFICATION

A. Support Vector Machine

SVM is generally useful for statistical learning and determining the point location of decision boundaries which results the optimal separation of classes. The SVC classifier was used in implementing the "one-against-one" approach for multi- class classification problem where the label's were drawn from finite set of several elements.The samples of kuzushiji characters, round about 3923 class was taken as the number of classes, then this $(3923 * 3923 - 1)/2$ classifiers are constructed and each one samples was trained data from two classes. The decision function shape option allows to transform the results of the "one-against-one" classifiers to a decision function of shape $(297300 samples, 3923 classes)$. Applying each classifier to the test data vectors gives one vote to the exact class. The results of a recent analysis of multi-class strategies are provided in Fig. 15.

NOTE Visualization of support vector classifier. The graph is plotted for 25 sample points which enclose only 2 features since the data samples is huge for reference is as shown in Fig. 15.

B. Neural Network Classifier

The usage of Neural nets was taken for the classification and recognition, since it consists of artificial network of

Fig. 15. Distance Surface of Radial Bias Function

TABLE II. RESULTS COMPARISON OF CLASSIFIERS

Layer (type)	Output Shape	Param
Sl.No	Model	Accuracy
1	Support Vector Machine	87.4%
2	Neural Network Classifier	90%

Fig. 17. Validation and Training Accuracy

Fig. 16. Simple Neural Network Architecture

Fig. 18. Validation Loss and Training Loss

TABLE I. ARCHITECTURE OF SIMPLE NEURAL NETWORK

Layer (type)	Output Shape	Param
dense1 (Dense)	(None, 72)	5256
dense2 (Dense)	(None, 820)	59860
dropout1 (Dropout	(None, 820)	0
dense3 (Dense)	(None, 1640)	1346440
dropout2 (Dropout)	(None, 1640)	0
dense4 (Dense)	(None, 3923)	6437643

functions called parameters which was able to learn all the feature of the images for analyzing the new data after receiving one or multiple inputs as shown in the Fig. 16 followed by the validation results is as shown in the Fig. 17 and the layers is as mentioned in the Table I.

C. Accuracy Comparison Table

The results as shown in the Table II

V. CONCLUSION

In this research work a complete methodology and multiple feature extraction technique was applied for historical documents for recognition purposes [15]. This methodology can be applied to either machine printed or handwritten documents. It is not necessary nor any prior knowledge of the fonts nor the existence of standard database because it can adjust depending on the type of documents that needed to process [16].

VI. FUTURE WORK

The next work is to focus on optimizing the recognition rate and finding out the new algorithm for feature extraction for better enhancement.

VII. ACKNOWLEDGMENT

REFERENCES

[1] T.M.Rath and R. Manmatha, "Word spotting for historical documents", International Journal on Document Analysis and Recognition (IJDAR), Vol.9, No 2 – 4, pp. 139 – 152 , 2006

[2] V. Lavrenko, T. M. Rath, R. Manmatha: "Holistic Word Recognition for Handwritten Historical Documents", Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04),pp 278287, 2004.

[3] T. Adamek, N. E. O'Connor, A. F. Smeaton, "Word Matching Using Single-Closed Contours for Indexing Handwritten Historical Documents", International Journal on Document Analysis and Recognition (IJDAR), special Issue on Analysis of Historical Documents, 2006.

[4] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis and S. J. Perantonis, " Keyword - Guided Word Spotting in Historical Printed Documents Using Synthetic Data and User feedback ", International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents, Vol. 9, No. 2-4, pp. 167-177, 2007.

[5] V.G.Gezerlis and S.Theodoridis, "Optical Character Recognition for the Orthodox Hellenic Byzantine music notation", Pattern Recognition, Vol.35, pp. 895 – 914, 2002.

[6] L. Laskov, "Classification and Recognition of Neume Note Notation in Historical Documents", International Conference of Computer Systems and Technologies (CompSysTech), 2006.

[7]] K. Ntzios, B. Gatos, I. Pratikakis, T. Konidaris and S.J. Perantonis, "An Old Greek Handwritten OCR System based on an Efficient Segmentation-free Approach", International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents, Vol. 9, No. 2-4, pp. 179-192, 2007.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. 2015 "Deep Residual Learning for Image Recognition." arXiv:1512.03385

[9] G.Huang, Z.Liu, L.van der Maaten, K.Q.Weinberger. 2016 "Densely Connected Convolutional Networks." 2016. IEEE Conference on Pattern Recognition and Computer Vision (PRCV2016)

[10] Szegedy, C., Liu, W., Jia, Y.Q., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. 2016 "Rethinking the Inception Architecture for Computer Vision." IEEE Conference on Pattern Recognition and Computer Vision(PRCV 2016)

[11] Francois Cholle. 2017 "Xception: Deep Learning with Depthwise Separable Convolution." IEEE Conference on Pattern Recognition and Computer Vision(PRCV 2017)

[12] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Wey, Marco Andreetto and Hartwig Adam. 2017. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." arXiv:1704.04861.

[13] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen. 2018. "Mobilenetv2: Inverted residuals and linear bottleneck." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2018).

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu and Alexander C. Berg, 2018. "SSD: Single Shot MultiBox Detector." arxiv:1512.02325

[15] http://codh.rois.ac.jp/char-shape/book/ (2020.2.26 accessed)

[16] Aravinda C.V, Meng Lin, and Amar Prabhu G. "Kuzashi recognition API" http://www.atait.se.ritsumei.ac.jp/KuzushijiMser/ (2020.2.26 accessed)

Impact Analysis of Network Layer Attacks in Real-Time Wireless Sensor Network Testbed

Navjot Sidhu[1]
Research Scholar
I. K. Gujral Punjab Technical University
Kapurthala, Punjab, India

Monika Sachdeva[2]
Associate Professor
I. K. Gujral Punjab Technical University
Kapurthala, Punjab, India

Abstract—With the rapid increase in the demand for Wireless Sensor Network (WSN) applications. The intrusive activities are also raised. To save these networks from the intruders it is required to understand the implications of any malicious act. Most of the researchers have utilized simulated software to understand the impact of such intrusions, however, real network conditions vary from the simulated environment. Therefore, the current work focuses on analyzing the impact of network layer attacks in real-time WSN testbed. The contributions of this work are threefold. Firstly, it presents the deployment of a real-time experimental testbed using standardized sensor devices in a multi-hop topological arrangement. Secondly, it provides the implementation details of seven network layer attacks: Blackhole (BH), Dropping Node (DN), Drop Route Request (DRREQ), Drop Route Reply (DRREP), Drop Route Error (DRERR), Grayhole (GH) and Sinkhole (SH) in a single testbed. Finally, the testbed performance with and without each attack is monitored and compared in terms of network performance metrics to understand the attacks' impact. This work will be helpful for the research community for proposing efficient attack detection and prevention solutions for these networks.

Keywords—Attack; impact; performance; real-time; Wireless Sensor Network (WSN)

I. INTRODUCTION

Wireless Sensor Network (WSNs) is a widely used technology in most of the monitoring applications nowadays. The common applications include security monitoring of homes, health monitoring in hospitals, traffic monitoring on roads, warehouse monitoring, weather monitoring, etc. The basic aim of all of its applications is to monitor the sensing environment and send the sensor readings to the base station so that the appropriate actions can be taken on time. Hence, data from the sensor nodes is quite important for mission-critical applications [1].

To achieve these applications' goals, the deployed sensor nodes should monitor the environment and communicate with other network devices timely. However, the intruders usually disrupt the normal network operations by launching intrusions in these networks. As a result of which, sensor nodes can not communicate with each other and the network performance degrades drastically. Hence, to prevent such networks from degradation, these attacks' impact should be monitored carefully.

However, most of the previous studies investigated the impact of a few intrusive activities by implementing the attacks in a simulated environment only. The simulation results vary drastically from the actual network performance. So, the simulated findings can not be directly applied to real life application scenarios.

For this work, to observe the network's actual performance with and without attack, the real-time WSN testbed is deployed for experimentation which presented a realistic performance measurement. For differentiating the behavior of the network under intrusive and legitimate activities, seven similar network layer attacks are implemented in the testbed. So, the majority of attacks at the network layer can be studied in a single study. The work successfully shows the impact of each attack in the testbed in terms of computed performance metrics.

This work can be of great value to understand such attacks in real-life applications. It can provide an insight for the developers to design secure sensor devices. Moreover, it can be used to introduce protective measures to prevent these networks from such critical attacks.

This paper is organized into the following sections: Section II describes the related literature work. Section III is an experimental methodology that defines the WSN testbed, the routing protocol used in the testbed, attack implementations, and performance metrics used. Section IV presents and discusses the results. Finally, Section V concludes the work.

II. RELATED WORK

This section presents the overview of existing attack implementations and experimentation analysis in WSN.

Tripathi et al. [2] simulated the blackhole and grayhole attack in LEACH protocol based upon the energy threshold in NS-2 and compared the attack impact. The authors suggest the detection of such attacks at the base station by observing the cluster head node and its data transmission. Dini and Tiloca [3] presented a stimulative approach for attack impact analysis and ranked the attacks according to severity. The paper also analyzed separate countermeasures for each attack type. However, it is an extremely costly solution for resource-constrained networks. Riecker et al. [4] measured the impact of Denial-of-Service Attacks: Jamming and Blackhole, using a testbed consisting of TelosB motes. The authors identified the performance metrics classes based on their capability to detect attacks. They used the packet delivery rate as a metric to differentiate between attack and normal network behavior. However, the detection of attacks only by observing performance metrics is not appropriate as sometimes the specific network parameters show significant variation in metrics in case of normal traffic

scenarios. Therefore, the fluctuation of legitimate traffic can be misunderstood as an attack. Chaudhary and Thanvi [5] analyzed the performance of modified AODV protocol under the DoS attack in NS-2. The paper suggested attack detection based on the RREP sequence number attribute of packets. But, such attack specific solutions are not feasible to save the real networks from attacks. Almomani and Al-Kasasbeh [6] presented the impact analysis of DoS attacks in LEACH: Blackhole, Grayhole, Flooding attack, and Scheduling attack using NS-2. The paper showed the attack impact with a major drop in packet delivery ratio. Nevertheless, the LEACH is not a standardized protocol used in WSN hardware. Rupayan Das et al. [7] compared the effect of network and physical layer attacks in the AODV routing protocol. The attacks were simulated in OPNET 14.5 and attack impact was analyzed with quality of service parameters.

Ioannou and Vassiliou [8] implemented routing layer packet drop attacks and investigated the attack impact from the sink node and victim node using the COOJA simulator. The authors considered the variations in the network topology to study attack impact. The paper showed that using some network parameters the presence of attacks can be identified however, it did not state the type of attack. Diaz and Sanchez [9] proposed a simulator for performance analysis of three attacks with attacker modeling and attack simulation, and suggested to be used by developers to understand attack behavior and develop secure systems. The paper did not provide any strategy to identify a specific type of attack. Govindasamy and Punniakody [10] analyzed the performance of AODV, OLSR, and ZRP routing protocols in Qualnet 5.0 under wormhole attack only using performance metrics. Tomin and McCann [11] analyzed the network layer attacks deployed against the routing protocol for low-power and lossy networks (RPL) using the COOJA network simulator. The impact of attacks was shown using the packet delivery ratio and end-to-end delay. But, specific attack types cannot be identified from the performance metrics of a network. Baskar et al [12] simulated and analyzed the network performance under the sinkhole attack in WSN in terms of energy consumption, throughput, and packet delivery ratio. The authors concluded that in a network with a large number of network nodes and few attacker nodes, attack impact was less. However, there is no practical evidence of this conclusion. Rana and Kumar [13] provided a detailed analysis of the AODV routing protocol with and without the presence of malicious node in the network using Qualnet 5.0 simulator and analyzed the throughput, average jitter, and packet drop ratio. Gomez et al [14] implemented the wormhole attack in a ZigBee experimental framework using XBee S2C nodes to find signatures for detecting this attack in real environments.

To sum up, much of the preceding research was performed in a simulated environment to understand the impact of attacks in WSNs. Moreover, the majority of the earlier studies focused on considering merely one or two attack types at a time. But, the real-time network behaves differently than the controlled simulated network.

Therefore, the multiple related attacks are implemented in real network scenarios to study the detailed behavior of malicious activities. And, the behavior of these attacks as compared to legitimate network performance is analyzed. Besides, the proposed study used the standardized network

Fig. 1. WSN Testbed

devices to deploy the testbed for experimentation to create a real network similar to real-life application scenarios.

III. Experimental Methodology

The experimental methodology comprises of the detail of the components of testbed, the routing protocol used, the definitions of attacks implemented, and performance metrics calculated.

A. WSN Testbed

Most of the research in the field of security of WSN is being done in a simulated environment. Some researchers preferred to use open-source network simulators and others proposed new WSN specific simulators. However, the actual network environment differs significantly from the simulated one. Therefore, the real WSN testbed is being deployed for experimentation.

To conduct the experiments, $SENSEnuts^{TM}$ wireless sensor kit is used. The kit consists of three main components. The wireless sensor module senses the light and temperature of the sensing environment. The radio module is responsible for creating a route between the source and the destination nodes. It connects the nodes at multi-hop distance. The Gateway module is connected with the laptop/desktop computer and responsible for sending the captured information to the computer using the SenseLive software. The SenseLive is a C based software that provides functionalities to the users like programming a sensor and radio node, displays the data sensed by sensor nodes in the form of user-defined tables, saves the data tables in database files, etc.

Fig. 1 shows the topological arrangement of Sensenuts nodes used to conduct experiments. The real network of 26 nodes includes 15 sensor nodes and 11 radio nodes. 1 radio node is programmed as the personal area network (PAN) Coordinator which manages the whole network, 10 as relay
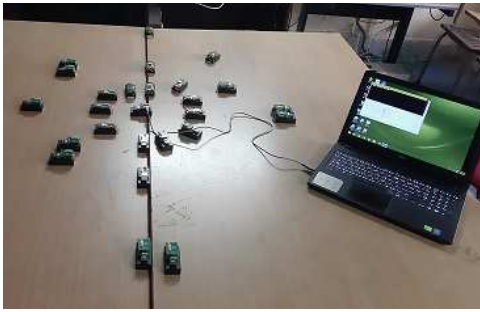
Fig. 2. WSN Testbed in LAB

nodes/ coordinators which have routing capabilities and 15 sensor nodes as sensing devices. Si represents the sensor nodes, Ri shows the relay nodes and P is a PAN coordinator. MAC id of each node is mentioned against each node. All the sensor nodes and relay nodes are battery-operated devices. PAN is connected with the computer system through the gateway module and uses the computer power to operate in the network. Hence, the PAN of the network has the highest operating power as compared to other devices.

Each sensor node searches a route to the PAN of the network either directly at a 1-hop distance or through relay nodes at 2-hop and 3-hop distance when it has the data to transmit. The route between the sensor node and PAN is established through a ZigBee based routing protocol. After the route is formed, the sensed information is sent by the sensor nodes to the PAN. All the performance-related calculations are performed at PAN as it has maximum operating capacity in comparison to other limiting power devices. After calculations, the required results are communicated to the computer system and stored in forms of tables. Fig. 2 shows the WSN testbed deployed in real-time.

B. Routing Protocol

The Sensenuts WSN platform is based on ZigBee wireless standard [15] which uses Adhoc On Demand Distance Vector (AODV) Routing protocol [16] at the network layer. AODV uses mainly three control packets i.e. Route Request (RREQ), Route Reply (RREP), Route Error (RERR) to search, establish, and repair the routes between the devices. When a sensor node has data to send but does not have a route to the destination node, it broadcasts the RREQ packet in the network. The RREP packet is unicasted either by the destination node or by an intermediate node if it knows an active route to the destination node. In case of a link failure or a route error, a RERR packet is either unicasted or broadcasted by an intermediate node. The implementation of AODV protocol under normal network conditions as well as under each attack, is shown in algorithm 1.

C. Attack Implementation

In the current work seven network layer attacks are implemented: Black hole, Dropping Node, Drop RREQ, Drop RREP, Drop RERR, Gray hole, Sinkhole. To implement these attacks, the attack actions to launch each attack type and goals achieved by the attackers are identified [17]. The analysis shows that all

of these routing layer attacks degrade the network performance and deplete the network resources. The implementation details of each attack type in real WSN test-bed is defined follows and implementation steps are described in algorithm 1:

1) Black hole Attack: For the implementation of the black hole attack [2], [18] when the attack node receives the RREQ message, it generates a false RREP message using a very high destination sequence number value to prove that it has a fresh route to the base station/destination. On the reception of the RREP message by the source node, it establishes a route through it based upon its highest value of destination sequence number field. When the sensed data is transferred through this route, the black hole node simply drops all the data packets. Therefore, when the route is formed through the black hole node, the information from the source node cannot reach to the destination.

2) Dropping Node Attack: The dropping node [18] does not launch any attack during route establishment like the black hole node. It simply waits for a route to be establish through it. When a route between a source and a destination node is established through the dropping node, it drops all the data packets.

3) Drop RREQ Attack: The drop RREQ attack [19] is implemented in a relay node as when the compromised node receives the RREQ message from its neighboring nodes, it simply drops the RREQ messages. This attack node either debars the route formation for neighboring source nodes or delays the route formation period.

4) Drop RREP Attack: The drop RREP node [19] drops the RREP message when received from any destination node or intermediate node. Therefore, the route is not established between two nodes.

5) Drop RERR Attack: For the implementation of drop RERR [19], when an attack node receives a RERR message, instead of forwarding, drops it. So that in case of any link failure and route error the corresponding processes cannot be initiated by the intended nodes.

6) Gray hole Attack: The Gray hole attack [2] is implemented as a variation of dropping node attack. It does not forcefully form any route through it. However, when a route for data transfer is formed through a Gray hole attack node, it launches a selective packet drop attack. The random packet drop is implemented for dropping data packets.

7) Sinkhole Attack: The Sinkhole attack [12], [20] is implemented as an extension to the Black hole attack. This attack node forces the destination node to send the RREP message through it. So, it changes the destination sequence number field of RREQ with a very high value and hop-count with a minimum value and broadcast RREQ towards the destination node. At the same time, it generates the RREP using a very high destination sequence number. Hence, the sinkhole node wins the route, and the detection of this node becomes difficult. After the route establishment, the attacker node randomly drops some of the data packets received from the sensor nodes.

The real-time experimentation is performed to capture the data from WSN testbed which is further used to analyze the malicious behavior during network communication. For analyzing the performance of WSN test-bed under attacks all

Algorithm 1 AODV operation under normal and attack conditions

$S \rightarrow SourceNode$
$D \rightarrow DestinationNode$
$I \rightarrow IntermediateNode$
$RT \rightarrow RouteTable$
$RREQ \rightarrow RouteRequestMessage$
$RREP \rightarrow RouteReplyMessage$
$RERR \rightarrow RouteErrorMessage$

Step 1: S checks RT entry for D.
Step 2: If (route exists)
  {
    Forward packets to next-hop.
  }
  Else
  {
    Initiates Route Discovery Process.
  }

Route Discovery:
Step 1: S creates and broadcasts RREQ.
Step 2: I receives RREQ.
Step 3: If (I=Drop RREQ Node)
  {
    I drops the RREQ.
    Exit
  }
  Else
  {
  If (D's sequence no in latest S'RREQ>D's sequence no in previous S'RREQ)
    {
      I sets up a Reverse Route Entry for S in its RT.
      If (Route exists from I to D in I's RT)
        {
        If (D's sequence no in I's RT>=D's sequence no in RREQ)
          {
          I Creates RREP.
          If (I= Sinkhole Node or Blackhole Node)
           {
           Set RREP.dst_seq_no= Higher value.
           }
          I unicasts RREP towards S.
          }
        }
      Else
        {
        I increments hop count in RREQ.
        If (I= Sinkhole Node)
          {
          Set RREQ.dst_seq_no= Higher value.
          }
        I broadcasts RREQ to its neighbors.
        }
    }
    Else

    {
      Discard RREQ.
    }
  }

Step 4: RREQ reaches D, provided D is reachable from S.
Step 5: D creates RREP.
Step 6: D unicasts RREP towards S.
Step 7: I receives a RREP.
Step 8: If (I=Drop RREP Node)
  {
    I drops RREP.
    Exit
  }
  Else
  {
    Sets up a Forward Route Entry to D in its RT.
    I forwards the RREP towards S.
  }

Data Transmission:
Step 1: S receives RREP.
Step 2: S starts packet transmission on route to D.
Step 3: I receives the data packets.
Step 4: If (I=Dropping Node or Blackhole Node)
  {
    Drop all the data packets.
  }
  Else If (I=Grayhole Node or Sinkhole Node)
  {
    Drop some of the data packets randomly.
  }
  Else
  {
    I forwards the data packets towards D.
  }
  End If

Route Maintenance:
Case I: S disconnects/moves from route established. S initiates a new route discovery.
Case-II: Either I or D disconnects/moves
Step 1: I creates RERR.
Step 2: I forwards RERR towards S.
Step 3: If (I=Drop RERR Node,receives RERR)
  {
    I drops RERR.
  }
  Else If (S receives RERR)
  {
    Delete route to D.
    Initiates a new route discovery.
  }
  Else If (Another I receives RERR)
  {
    Delete route to D.
    Forwards RERR towards S.
  }
  End If

TABLE I. NETWORK PARAMETERS

Parameters	Value
PAN Node	1
Relay Nodes	10
Sensor Nodes	15
Routing Protocol	AODV
Topology	Multi-hop (up to 3-hop distance)
Packet Size	6 Bytes
Packet Interval	1 sec
No of Packets Send	100
No of Attacker	01
No of Attacks Launched	07

network layer attacks are being implemented on the testbed. In the real network of 26 nodes, firstly the legitimate performance of the test-bed has been captured so that the comparison can be made in case of any intrusion in the network. After that, for each scenario, one relay node is programmed as an attack node and the behavior of the network is captured in terms of performance metrics. For calculating the results, each experiment is run five times so that real network efficiency can be observed. Table I summarizes the network parameters used for conducting experiments.

D. Performance Metrics

The following performance metrics for each sensor node are measured at PAN to compare the network performance with and without any attack.

1) Average Throughput (in bps): It is the average of the number of bits received from a sensor node to its data transmission duration.

$$Avg.Throughput = (No.of PacketsReceived * PacketSize)/Time \quad (1)$$

2) Packet Delivery Ratio (PDR): It is the ratio of the number of packets received at the base station from a sensor node to the number of packets sent by that sensor node.

$$PDR = (No.of PacketsReceived)/(No.of PacketsSent) \quad (2)$$

3) Number of Packets Received: It is the sum of the number of packets received at the base station from a sensor node.

4) Average Inter-arrival Time (IAT) (in secs): It is the average time difference between two consecutive packets when reached the base station from a sensor node.

$$Avg.IAT = (CurrentPacketTime - Previous PacketTime)/(No.of PacketsReceived) \quad (3)$$

All of the calculations are performed at the base station only as it has the highest battery power. The impact of each attack in terms of calculated performance metrics is presented in the form of graphs in the next section.

IV. RESULTS AND DISCUSSION

The results show the impact of each attack with respect to the legitimate network performance in terms of identified performance metrics.

1) Black hole Attack Analysis: Fig. 3(A) shows the impact of Black hole attack in the testbed in terms of average throughput. The average throughput lies between 32 to 49 bps in case of the legitimate network. However, when one relay node acted as black hole attacker, it affected majority of sensor nodes and resulted in zero bps average throughput of these sensor nodes. Because such sensor nodes could not connect with the base station during network formation and all of their data packets are dropped. Similarly, the PDR of these nodes as shown in Fig. 3(B), is also zero as all the data packets from these sensor nodes are dropped by the black hole attack node. It can be further verified from Fig. 3(C), which shows the number of packets from each sensor node reached at the base station. It is zero for the attack affected sensor nodes. Fig. 3(D) shows the average IAT, which is zero for the sensor nodes which could not be connect with the base station during the experimentation.

2) Dropping Node Attack Analysis: As shown in Fig. 4 sensor nodes 0B87, D035, and D326 are under the influence of the dropping node attack. The average throughput of these sensor nodes is zero bps as shown in Fig. 4(A). Also Fig. 4(B) and (C) are showing the PDR and number of packets received, respectively for these affected nodes that are also zero. The average IAT for those sensor nodes that could not communicate with the base station is also zero (see Fig. 4(D)).

3) Drop RREQ Attack Analysis: Fig. 5 shows the impact of Drop RREQ attack. Due to the presence of DRREQ attacker, two sensor nodes could not connect with the base station because the RREQ packets sent by these sensor nodes dropped by the attacker node. Therefore, the average throughput of these nodes is zero bps and average IAT is zero seconds. Also the route establishment of few nodes is delayed by DRREQ attack as a result of which less number of packets from those sensor nodes reached at the base station resulted in less or zero PDR and low packet received count.

4) Drop RREP Attack Analysis: As shown in Fig. 6, the impact of Drop RREP attack is either delayed route formation or no route between a sensor node and the base station. Average throughput of the affected sensor nodes is increased because such nodes are able to send less number of packets to the base station and most of their data packets are dropped while waiting for the route establishment. In Fig. 6(B) and (C), the PDR and number of packets received from the sensor nodes clearly shows the influence of DRREP attacker in the testbed.

5) Drop RERR Attack Analysis: Fig. 7 shows the impact of Drop RERR attack in the testbed. The DRERR attacker delayed the route repair in the network in case of a link failure. As a result of which less number of packets from the affected sensor nodes reach the base station and hence showing less PDR. The average throughput and average IAT for such sensor nodes are dropped due to delayed route repair and packet dropped during this period.

6) Gray Hole Attack Analysis: The Gray Hole attacker drops the data packets selectively therefore results in decreased average throughput, PDR, packet received count and increased average IAT of the affected sensor nodes. As shown in Fig. 8(A) the average throughput of two sensor nodes is decreased. Fig. 8(B) and (C) shows the decrease in PDR and
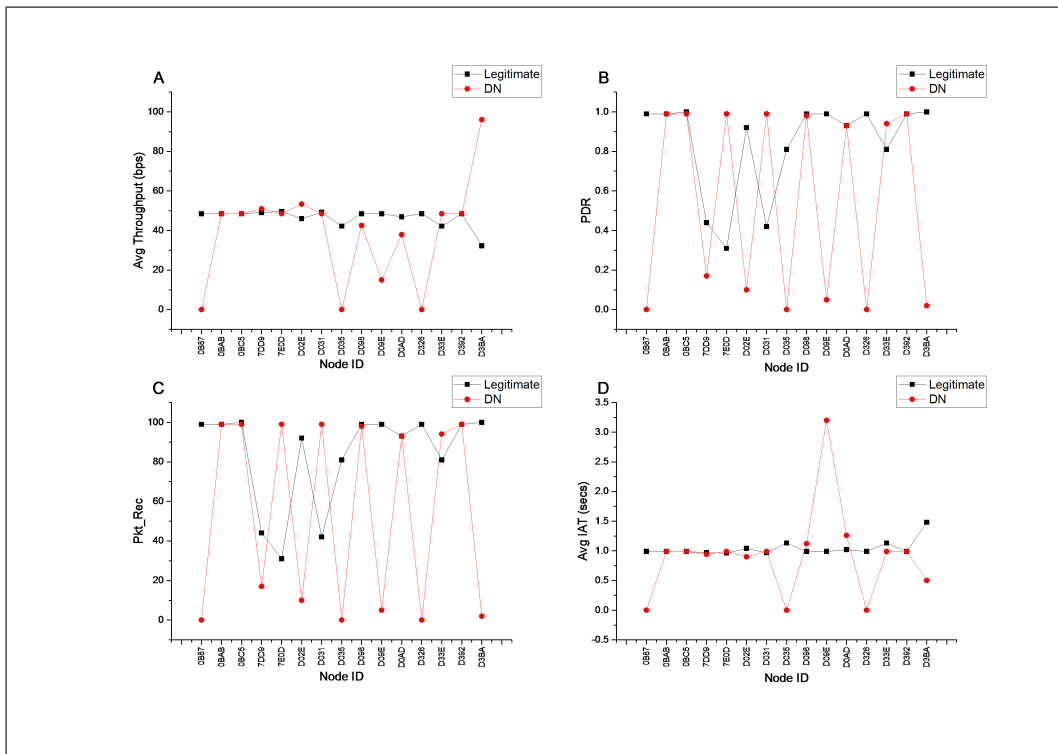
Fig. 3. Impact Analysis-Black Hole Attack

Fig. 4. Impact Analysis-Dropping Node Attack

packet received count. Fig. 8(D) shows the increased average IAT for such sensor nodes.

7) Sink Hole Attack Analysis: Like a Gray hole attacker, the Sink hole attacker also drops the data packets selectively. Besides, it influences more number of sensor nodes in the
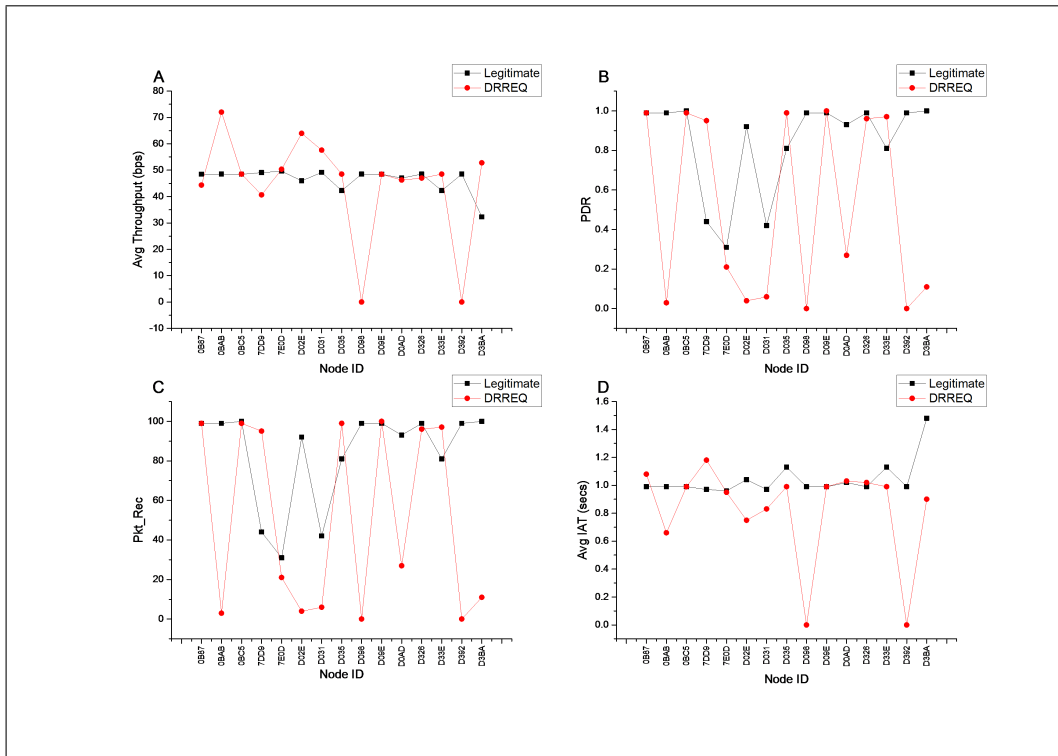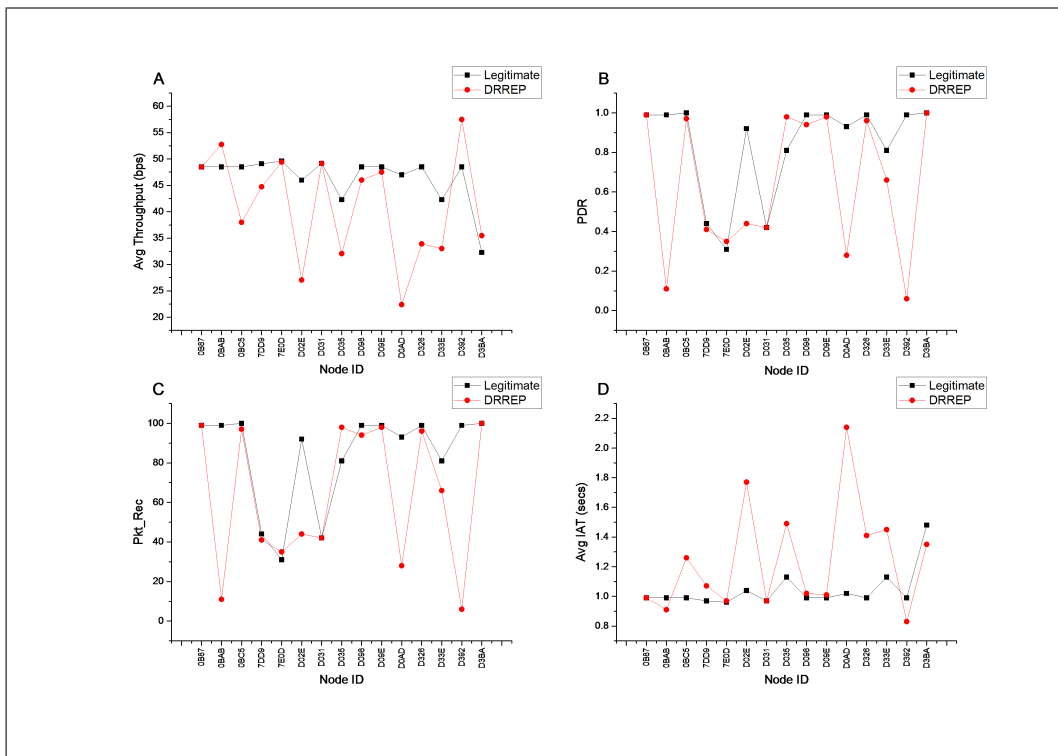
Fig. 5. Impact Analysis-Drop RREQ Attack

Fig. 6. Impact Analysis-Drop RREP Attack

network because it forces the sensor nodes to form their routes to the base station through it. The impact of the Sink hole attack is clearly visible in Fig. 9 as the average throughput, PDR and packet received count are decreased and the average IAT of the affected sensor nodes is increased.

Fig. 7. Impact Analysis-Drop RERR Attack

Fig. 8. Impact Analysis-Gray Hole Attack

V. CONCLUSION

The study presents the implementation of seven network layer attacks on a single testbed. The attacks are chosen to understand the difference between their implications on the network. The attacks are launched one at a time so that the impact can be clearly shown in terms of performance metrics.

Fig. 9. Impact Analysis-Sink Hole Attack

The experiments are created to capture the network performance with and without attack. The results clearly show the impact of each attack in the testbed in terms of performance metrics. In case of an attack, the average throughput is either zero when a node could not send any data to the base station or it is reduced significantly due to selective packet drop and delayed route establishment. A similar impact is found on the PDR and the number of packets received at the base station. The average inter-arrival packet time increased in case of selective forwarding attacks. Therefore, the results depict that the impact of each attack type is visible in the testbed as per the attack definitions, in terms of network performance metrics, which validates the accurate implementation of all network layer attacks. A clear understanding of this behavior is helpful for the research community to differentiate between legitimate and attacked network.

For the future work, this experimentation is used to construct a real-time dataset for the WSN, which will further be used to propose and validate attack detection techniques in these networks.

Acknowledgment

References

[1] P. Rawat, K. D. Singh, H. Chaouchi, and J. M. Bonnin, "Wireless sensor networks: A survey on recent developments and potential synergies," *The Journal of supercomputing*, vol. 68, no. 1, pp. 1–48, 2014.

[2] M. Tripathi, M. S. Gaur, and V. Laxmi, "Comparing the impact of black hole and gray hole attack on leach in wsn," *Procedia Computer Science*, vol. 19, pp. 1101–1107, 2013.

[3] G. Dini and M. Tiloca, "On simulative analysis of attack impact in wireless sensor networks," in *IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA)*. IEEE, 2013, pp. 1–8.

[4] M. Riecker, D. Thies, and M. Hollick, "Measuring the impact of denial-of-service attacks on wireless sensor networks," in *39th annual IEEE conference on Local Computer Networks*. IEEE, 2014, pp. 296–304.

[5] S. Chaudhary and P. Thanvi, "Performance analysis of modified aodv protocol in context of denial of service (dos) attack in wireless sensor networks," *International Journal of Engineering Research and General Science*, vol. 3, no. 3, pp. 486–491, 2015. [Online]. Available: http://pnrsolution.org/Datacenter/Vol3/Issue3/65.pdf

[6] I. Almomani and B. Al-Kasasbeh, "Performance analysis of leach protocol under denial of service attacks," in *6th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2015, pp. 292–297.

[7] R. Das, S. Bal, S. Das, M. K. Sarkar, D. Majumder, A. Chakraborty, and K. Majumder, "Performance analysis of various attacks under aodv in wsn & manet using opnet 14.5," in *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2016, pp. 1–9.

[8] C. Ioannou and V. Vassiliou, "The impact of network layer attacks in wireless sensor networks," in *International Workshop on Secure Internet of Things (SIoT)*. IEEE, 2016, pp. 20–28.

[9] A. Diaz and P. Sanchez, "Simulation of attacks for security in wireless sensor network," *Sensors*, vol. 16, no. 11, p. 1932, 2016.

[10] J. Govindasamy and S. Punniakody, "A comparative study of reactive, proactive and hybrid routing protocol in wireless sensor network under wormhole attack," *Journal of Electrical Systems and Information Technology*, vol. 5, no. 3, pp. 735–744, 2018.

[11] I. Tomic and J. A. McCann, "A survey of potential security issues in existing wireless sensor network protocols," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1910–1923, 2017.

[12] R. Baskar, P. Raja, C. Joseph, and M. Reji, "Sinkhole attack in wireless sensor networks-performance analysis and detection methods," *Indian Journal of Science and Technology*, vol. 10, no. 12, 2017.

[13] R. Rana and R. Kumar, "Performance analysis of aodv in presence of malicious node," *Acta Electronica Malaysia (AEM)*, vol. 3, no. 1, pp. 1–5, 2019.

[14] J. R. Gomez, H. F. V. Montoya, and A. L. Henao, "Implementing a wormhole attack on wireless sensor networks with xbee s2c devices," *Revista Colombiana de Computacion*, vol. 20, no. 1, pp. 41–58, 2019.

[15] W. C. Craig, "Zigbee: Wireless control that simply works," *Zigbee Alliance ZigBee Alliance*, 2004. [Online]. Available: https://eclass.uop.gr/modules/document/file.php/TST220/bibliography/Zigbee-tutorialZMDAmerica.pdf

[16] C. Perkins, E. Belding-Royer, and S. Das, "Rfc3561: Ad hoc on-demand distance vector (aodv) routing," 2003. [Online]. Available: https://dl.acm.org/doi/pdf/10.17487/RFC3561

[17] N. Sidhu and M. Sachdeva, "A comprehensive study of routing layer intrusions in zigbee based wireless sensor networks," *International Journal of Advanced Science and Technology*, vol. 29, no. 3, pp. 514–524, 2020. [Online]. Available: http://sersc.org/journals/index.php/IJAST/article/view/3951

[18] H. Ehsan and F. A. Khan, "Malicious aodv: Implementation and analysis of routing attacks in manets," in *11th International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE, 2012, pp. 1181–1187.

[19] A. A. Korba, S. Ghanemi, and M. Nafaa, "Analysis of security attacks in aodv," in *International Conference on Multimedia Computing and Systems (ICMCS)*. IEEE, 2014, pp. 752–756.

[20] H. Moudni, M. Er-rouidi, H. Mouncif, and B. El Hadadi, "Performance analysis of aodv routing protocol in manet under the influence of routing attacks," in *International Conference on Electrical and Information Technologies (ICEIT)*. IEEE, 2016, pp. 536–542.

Deep Learning with Data Transformation and Factor Analysis for Student Performance Prediction

Tran Thanh Dien[1], Sang Hoai Luu[2], Nguyen Thanh-Hai[3], Nguyen Thai-Nghe[*4]

College of Information and Communication Technology
Can Tho University, Can Tho, Vietnam
[*]Corresponding Author: Can Tho University, 3/2 Street, Ninh Kieu District, Can Tho city, Vietnam

Abstract—Student performance prediction is one of the most concerning issues in the field of education and training, especially educational data mining. The prediction supports students to select courses and design appropriate study plans for themselves. Moreover, student performance prediction enables lecturers as well as educational managers to indicate what students should be monitored and supported to complete their programs with the best results. These supports can reduce formal warnings and expulsions from universities due to students' poor performance. This study proposes a method to predict student performance using various deep learning techniques. Also, we analyze and present several techniques for data pre-processing (e.g., Quantile Transforms and MinMax Scaler) before fetching them into well-known deep learning models such as Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) to do prediction tasks. Experiments are built on 16 datasets related to numerous different majors with appropriately four million samples collected from the student information system of a Vietnamese multidisciplinary university. Results show that the proposed method provides good prediction results, especially when using data transformation. The results are feasible for applying to practical cases.

Keywords—*Deep learning; student performance; mark prediction; Long Short Term Memory (LSTM); Convolutional Neural Networks (CNN); data pre-processing; multidisciplinary university*

I. INTRODUCTION

Recently, the number of students who have been warned and forced to leave school tends to increase. One of the reasons can be that students are not able to evaluate and predict correctly his or her ability to select appropriate courses. Student performance is an important task of higher educational institutions because it is a criteria for high quality universities that are based on excellent profile of their academic achievements. There are several definitions on student performance. According to [1], student performance can be obtained by measuring the learning assessment and curriculum. However, most of the studies mentioned about graduation being the measure of students' success [2], [3].

In recent years, the situation of students in the institutions have been academically warned tended to accelerate. For example, at Can Tho University[1], in the first semester of the school year 2018-2019, the number of students who academic warned in one semester were 886 and the two semesters were 125, these number in the first semester of the academic year

2019-2020 were 986 and 196 respectively. One of the main reasons for the students' poor performance is that they have not selected appropriate courses to their competencies. These results in extension of learning term and increase of cost for their families, higher educational institutions and society as well. Therefore, predicting students' performance is an important research topic in exploiting educational data, which is of interest to many researchers [4]–[8].

Currently, there are a lot of proposed approaches to predict student performance, in there data mining is one of the most popular approaches to be widely applied in educational area. One of the most popular techniques to predict student performance is classification. There are several algorithms used for classification task such as Decision Tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machines [3]. However, the existing researches are primarily based on learning results of previous semesters to predict student performance of next semester or Current Grade Point Average (GPA), but do not analyze additional factors such as English entrance testing grades, activity incentive grades, etc. that affect their performance. Moreover, the researchers have not sufficiently compared among techniques, especially deep learning techniques with other traditionally machine learning techniques.

This study presents an approach of deep learning techniques [9] using the convolutional neural network on 1D data (CN1D) and the Long Short Term Memory (LSTM) to build a student's performance prediction model for predicting student performance in next semesters based on the course's achievement results of the previous semesters.

We analyze and introduce some techniques for data pre-processing (including Quantile Transforms and MinMax Scaler) before fetching them into well-known deep learning algorithms such as LSTM and Convolutional Neural Networks to do prediction tasks. In addition, in order to improve the predictive results, we also consider other additional factors such as entrance English testing grades, activity incentive grades etc. for the proposed model. Moreover, a comparison between deep learning techniques and traditionally machine learning ones is also conducted. Experimental results show that the proposed model provides rather accurate prediction and it can be applied in practical other cases.

In the remainder of this study, we present a literature review on studies performing on student performance prediction. We introduce the considered 17 datasets in Section III. The works

---

[1]www.ctu.edu.vn

with the proposed techniques and algorithms are presented in Section IV. The experimental results include the model settings and performance metrics, data transformations, optimizer functions, results' comparison in Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), and factor analysis which are presented in Section V. We conduct and summarize some remarks in Section VI.

II. A LITERATURE REVIEW ON STUDENT PERFORMANCE PREDICTION

Predicting student's academic performance becomes critical demands that institutions can support students to achieve more performance in their learning. In [2], almost 70 papers were analyzed to show different modern techniques widely applied for predicting students' performance. These techniques belongs to several areas such as Artificial Intelligence, Machine Learning, Collaborative Filtering, Recommender Systems, and Artificial Neural Networks.

According to [8], predicting student performance is an important task in exploiting educational data; student's knowledge can be improved and accumulated over time. From this idea, the authors proposed an approach that uses Tensor Factorization (TF) to predict student performance. With this approach, the authors can personalize the prediction for specific student. Experiential results on two large datasets showed that incorporating prediction matrix factorization techniques is an effective and promising approach.

The authors in [10] investigated the effectiveness of transfer learning from deep neural networks for the task of students' performance prediction in higher education. Experiments were conducted based on data originating from five compulsory courses of two undergraduate programs. The experimental results demonstrate that the prognosis of students at risk of failure can be achieved with satisfactory accuracy in most cases.

The authors in [11] developed a student performance prediction system using the open source recommendation system called *MyMediaLite*. For the grade databases collected from the academic management system of a university, the authors proposed using Biased Matrix Factorization (*BMF*) technique to predict the learning results. This results can help students choose more appropriate courses.

The ability to combine the prediction techniques is also used by researchers. [12] developed a model to predict the student learning outcomes based on the combination of Taylor approximation method and Grey models to obtain the most optimal predicted values by multitimes approximate calculation to improve the predicted accuracy of two grey models. Research results can help teachers and educational managers have appropriate solutions to improve the academic results of students who have unstable learning process. In addition, [13] used Collaborative Filtering, Matrix Factorization and Restricted Boltzmann Machines techniques to systematically analyze data collected from a university. The results showed that Restricted Boltzmann Machines technique predicts students' academic results better than the remaining techniques.

In fact, collaborative filtering algorithms are commonly used in recommendation systems due to their simplicity and effectiveness. However, the sparsity of the data limits the effectiveness of these algorithms and it is difficult to further improve the prediction results. Therefore, the models that combine collaborative filtering algorithms with deep learning techniques are more interested. [14] proposed a model based on quadratic polynomial regression model to obtain more accurate latent features by improving the traditional matrix factorization algorithm. Then, the latent features are the input data of the deep neural network model. The experiments on three datasets showed that the proposed model improves the prediction efficiency very well compared to traditional ones. Some other approaches combining collaborative filtering model with deep learning are also proposed by [15]. With this approach, during the prediction period, a feed-forward neural network is used to simulate the interaction between the are user and the item, in which the feature vectors in pre-process used as input to the neural network. The experiments based on two datasets of *MovieLens* with one million samples and MovieLens ten million features to verify the effectiveness of this method and gave very feasible results. [16] also presented a review on machine learning based approaches for predicting student performance. Other approaches can be found in [17]–[20].

The problem of student performance prediction has been taken into account in numerous previous research using machine learning theory but factor analysis for student performance prediction based on explanation models and data transformation techniques are still the gap for improvements. This research aims to create a new approach that leverages Deep learning. Especially, the Long Short Term Memory (LSTM) can be used with time-based features. This study includes several contributions as the following:

- Deep learning models (LSTM, CNN) with a shallow architecture are leveraged to do student performance prediction tasks. As shown from the results, deep learning techniques can produce feasible prediction scores.

- Various optimizer functions are tested to choose an appropriate one for the considered regression problem.

- Data transformation techniques are also considered to enhance the performance of deep learning models. The feature values which are greater than 1 cause poor performance for deep learning model. By using and testing various data pre-processing techniques, we found that regression tasks with deep learning can converge sooner and also archive a better result.

- We investigate and consider various 17 datasets related to a vast of majors and study fields for the comparison in a multidisciplinary university. Based on the time to divide the training and testing data, we obtain the various ratio between the training set and test set to evaluate the difference in the prediction performance.

- A variety of model explanations are brought to analyze factors which can influence on student performance. From analysis results, educational managers can propose appropriate policies and strategies to support their students.

TABLE I. INFORMATION ON THE NUMBER OF SAMPLES OF EACH STAGE AND PERCENTAGE OF TRAINING DATA OF THE CONSIDERED EDUCATIONAL UNITS AND DEPARTMENTS

Dataset	#Train Instances	#Test Instances	%Training ratio	Total
Education	292,297	78,987	78.73%	371,284
Mekong Delta Development Research Institute	27,795	10,206	73.14%	38,001
Agriculture	294,694	179,042	62.21%	473,736
Biotechnology Research & Development Institute	46,556	30,075	60.75%	76,631
Physical Education	26,318	6427	80.37%	32,745
Engineering Technology	418,835	21,4710	66.11%	633,545
Information and Communications Technology	132,907	86,901	60.47%	219,808
Natural Sciences	79,368	42,121	65.33%	121,489
Rural Development	101,039	102,994	49.52%	204,033
Environment & Natural Resources	125,659	83,687	60.02%	209,346
Economics	518,392	171,538	75.14%	689,930
Foreign Languages	125,882	56,559	69.00%	182,441
Social Sciences and Humanities	96,491	47,469	67.03%	143,960
Aquaculture & Fisheries	109,637	49,498	68.90%	159,135
Law	155,099	54,194	74.11%	209,293
Political Science	33,493	30,009	52.74%	63,502
Total (full dataset)	2,584,462	1,244,417	67.50%	3,828,879

Fig. 1. Percentages of Educational Units and Departments Comparing to the Full Dataset

III. DATA DESCRIPTION

In order to evaluate the proposed model, we have collected real data at a multidisciplinary university, a case of Can Tho University, Vietnam. However, the model can be applied to other universities, schools, colleges as well. The collected data relates to students, courses, marks, and other information from the year 2007 to 2019 with 3,828,879 records, 4,699 courses (subjects), and 83,993 students. Data distributions are described in Table I with information on samples and the ratio for training of educational units/department/institutes at the considered multidisciplinary university.

The set of datasets consists of student performance from 16 academic units (faculties/colleges/schools) that belong to Can Tho University. For each unit, we separate the data into two parts, one of them for the training stage and the remaining for the test stage. Because of data division based on periods (from 2007 to 2016 for training, and from 2017 to 2019 for testing), the size of data for training and testing of each unit is different. Adding to these 16 academic units, we also evaluate our proposed method on the full dataset which includes all

Fig. 2. Distribution of Mark Levels on Training Set of the Full Dataset.

samples from 16 academic units. Fig. 1 reveals the percentages of each units comparing to the full dataset. Economics dataset occupies the highest value with 18% and the lowest belongs to Physical Education with 32,745 samples constituting 0.9%.

Fig. 3. Distribution of Mark Levels on Test Set of the Full Dataset.

The distribution of mark levels of the full dataset for training and testing described in Fig. 2 and Fig. 3, respectively. For these distributions, most of the marks are greater than or equal medium level as 2 (89.7% for full training dataset and 88.6% for full testing dataset). The distribution is similar to most units in the university, for instance, the mark level distribution of Engineering Technology dataset described in Fig. 4 and Fig. 5, respectively.

Fig. 4. Distribution of Mark Levels on Training Set of Engineering Technology Dataset.

Fig. 5. Distribution of Mark Levels on Test Set of Engineering Technology Dataset.

IV. A Proposed Approach based on Deep Learning for Student Performance Prediction

First, we collect real datasets at the Student Management System of a university, then data is pre-processed to remove noise, redundant attributes, etc. Next, we divide the data for training and testing in term of "time order" which means that we use the "studied courses" to predict the "to be studied courses". For example, we have used data from 2007 to 2016 for training, and data range from 2017 to 2019 for testing. The purpose of this division is to use courses results in the past (history data) to predict results of course in the future. In order to evaluate the efficiency of the prediction model, "the future" in this context is referred to data of the year from 2017.

A. Data Pre-Processing and Transformation

Since the dataset collected from the Student Management System have a lot of information, we have pre-processed them as described in the following steps:

- **Step 1**: Remove redundant attributes such as Student Name, Course Name, Lecturer Name, etc.

- **Step 2**: Remove redundant/noise records such as the courses which are registered by the student but have not been taken examination (i.e., the null marks), exemption courses, etc.

- **Step 3**: Remove the courses which have not enough registration (in some universities, if the courses are registered by less than 15 students, they will be removed).

- **Step 4**: Transform the text values to numeric values and other formats.

After carefully analyzing the data, we have selected the input attributes for learning model as described in Table II. This selection based on pre-experimental results and previous analysis in predicting student performance [11], [21].

With various data distribution of obtained different attributes, we suggest using Quantile Transformation (QTF) and MinMaxScaler (MMS) [25] for generating and convert all values to the value range where deep learning algorithms can converge.

QTF, a non-linear transformation, is considered a strong preprocessing technique because it reduces the effect of outliers. Values in new/unseen data (for example, a test/validation set) that are lower or higher than the fitted range are set to the bounds of the output distribution. As shown from Fig. 6 with an example from samples of Rural Development, before data is transformed, data range and distributions of each feature have great differences. Data transformed with QTF with all features range from 0 to 1 (Fig. 6b). Fig. 6a exhibits the result of scaler for each feature also enables its distribution to become more normal distribution.

MMS is also used for creating bins for images. This algorithm scales each feature to a given range with formulas 1 and 2:

$$X_{std} = \frac{X - X.min}{X.max - X.min} \quad (1)$$

TABLE II. INPUT ATTRIBUTES. EACH SAMPLE HAS THE LABEL WHICH IS A MAGNITUDE OF MARK/SCORE AS OUTPUT CORRESPONDING TO THAT SAMPLE. THE MAGNITUDE OF MARK/SCORE IS THE TARGET FOR THE PREDICTION TASKS.

Feature names	Description
CGPA	Average marks of the courses credited (passed) in all previous semesters
CGPA-PreSemester	Average marks of the courses credited (passed) in the preceded semester
CourseID	ID of the course
Credits_earned	The number of credits which the student earned from previous semesters
EnglishMark_l1	Basic English Course- Level 1
EnglishMark_l2	Basic English Course - Level 2
EnglishMark_l3	Basic English Course - Level 3
EntranceMark_s1	Entrance mark for Course 1
EntranceMark_s2	Entrance mark for Course 2
EntranceMark_s3	Entrance mark for Course 3
EntranceYear	Course of the student, e.g. 2007, 2009, etc.
Faculty	Student belongs to a faculty (e.g. Economics, etc.)
FieldOfStudy	Student's Field of study
Gender	Student gender
GPA_Semester	Average marks of the courses in the preceded semesters
HighSchoolPlace	High school where the student graduated
LecturerID	Lecturer ID who taught the course
Mark_RecordedTime	The time when the mark was recorded in the system
No._Semester	Student's Semester order (eg. 1st term, 2nd term, etc.)
NumberOfCredits	The number of credit of the course
StudentID	Student ID

Fig. 6. Data Distributions before and after using QTF on the Training Set of Rural Development Dataset. (a) The Results of before and after Transforming a Feature (CGPA-PreSemester) as an example. (b) An Illustration of Data Distribution with Various Data Ranges from some Features before applying the Transformation and Data Distribution of all Features Transformed using QTF with the Scale from 0 to 1.

$$X_{scaled} = X_{std} * (max - min) + min \qquad (2)$$

These algorithms are proven as an efficient method in classification tasks in [22]. In this study, the experiments also reveal promising results comparing to original data with regression tasks. The scaler is learned from the training set and applying to the test set.

B. Proposed Models

Two deep learning and a robust regression (Linear Regression) algorithms are carried out to run prediction tasks.

```
OPERATION           DATA DIMENSIONS    WEIGHTS(N)   WEIGHTS(%)

  Input    #####      21     1
 Conv1D    \|/     -------------------     256        17.4%
   relu    #####      19    64
Flatten    |||||   -------------------      0          0.0%
           #####           1216
  Dense    XXXXX   -------------------     1217        82.6%
sigmoid    #####            1
```

Fig. 7. The Proposed CN1D Architecture

The convolutional neural network (namely, CN1D) receives 1D data with 21 features, then passing through a stack of one convolutional layer with 64 kernels of 3 (stride 1), followed by a ReLU activation function used after each convolution (shown in Fig. 7).

```
OPERATION           DATA DIMENSIONS    WEIGHTS(N)   WEIGHTS(%)

  Input    #####       1    21
   LSTM    LLLLL   -------------------    22016       99.7%
   tanh    #####           64
  Dense    XXXXX   -------------------      65         0.3%
sigmoid    #####            1
```

Fig. 8. The Proposed LSTM Architecture

The Long Short Term Memory (LSTM) includes 64 Tanh units and one time step (Fig. 8). Both the CN1D and the LSTM produce output by a sigmoid function (Equation 3). The output of the sigmoid function ranges from 0 to 1, so this output then multiplied by 4 to corresponding the grades scale ranging from 0.0 to 4.0 for the mark prediction.

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

V. EXPERIMENTAL RESULTS

A. Learning Settings and Performance Metrics

All networks with deep learning models deploy either Adam algorithm or Root Mean Square Propagation- RMSprop algorithm [23] as the optimization functions with a learning rate of 0.0001, a batch size of 16000 running to 500 epochs. In order to reduce overfitting, we used early stopping with the epoch patience of 5. If the loss cannot be reduced after consecutive epochs, the learning will be stopped. The scaler algorithm learns from the training set and transforms both training and test sets.

The regression performance is measured using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) averaged 5 times of run on the test set. The root mean squared error and mean absolute error (MAE) are calculated by equations (4) and (5), respectively.

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{4}$$

$$\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{5}$$

where, y_i is the true value (Grades: 0.0-4.0 scale), and $\hat{y}_i$ is the predicted value.

The experimental results are presented as follows. The results with various scalers are illustrated in Section V-B where we show QTF as an appropriate solution for data preprocessing for regression tasks. Then, the selected scaler is run with two optimizer functions including Adam and RMSprop for the comparison in Section V-C. We run the experiments using deep learning and linear regression with the best scaler and optimizer function in Section V-D on all 16 datasets including department and institutes of the considered university and also carried out the prediction on the full dataset which merged from 16 datasets of educational units.

B. Scalers Enhance the Performance of Deep Learning

Various scalers provide the results shown in Table III. It is clear to see that the scalers are able to improve the performance of deep learning algorithms. QTF which can be the best choice among the considered scalers reveals the highest performance on 15 out of 16 datasets for CN1D and all datasets for LSTM. Fig. 9 and Fig. 10 also exhibit a clear view of the improvement. CN1D benefits from the scaler with a significant improvement.

Fig. 9. MAE Comparison using Various Scalers with LSTM

C. Root Mean Square Propagation (RMS) Optimizer with Deep Learning Regression

Table IV presents the result comparison between RMSprop and Adam optimizer function. We obtain the improved performances by using RMS on 14 out of 16 considered datasets while the difference in the performance is trivial for other two datasets. The average improvement on all datasets is estimated at 3.3%. Engineering Technology dataset achieves the most improvement with reducing 0.0660 in MAE while some of the other datasets such as Environment & Natural Resources, Aquaculture & Fisheries and Information and Communications Technology can reach better results by over 0.04 in MAE.

TABLE III. DETAILS OF STUDENT PERFORMANCE PREDICTION RESULTS (IN MAE) WITH LSTM AND CN1D USING VARIOUS SCALERS. THE RESULTS IN BOLD TEXT ARE THE BEST PERFORMANCE CORRESPONDING TO EACH DATASET AND EACH CLASSIFIER

Dataset	CN1D			LSTM		
	None	MMS	QTF	None	MMS	QTF
Education	0.8595	0.6231	**0.5847**	0.6637	0.6341	**0.6220**
Environment & Natural Resources	1.0308	0.6588	**0.6130**	0.6716	0.6531	**0.6461**
Economics	1.1129	0.6122	**0.6098**	0.6375	0.6130	**0.6092**
Foreign Languages	3.2603	0.5382	**0.4961**	0.5736	0.5336	**0.5031**
Social Sciences and Humanities	3.0476	**0.5711**	0.5793	0.5827	0.5653	**0.5647**
Aquaculture & Fisheries	1.1499	0.6502	**0.6471**	0.6835	0.6393	**0.6377**
Law	1.0574	0.6017	**0.5675**	0.6112	0.5924	**0.5878**
Political Science	1.0017	0.5565	**0.5547**	0.5868	0.5581	**0.5558**
Mekong Delta Development Research Institute	3.1145	0.5803	**0.5684**	0.7377	0.5746	**0.5587**
Agriculture	0.9733	0.5995	**0.5828**	0.6658	0.6042	**0.5814**
Biotechnology Research & Development Institute	0.8133	0.6016	**0.5980**	0.6308	0.5928	**0.5855**
Physical Education	3.0487	0.6942	**0.6853**	0.8159	0.6795	**0.6720**
Engineering Technology	1.3427	0.7761	**0.7487**	0.8491	0.8267	**0.8217**
Information and Communications Technology	2.7080	0.7728	**0.7285**	0.8306	0.8108	**0.7637**
Natural Sciences	1.2244	0.8016	**0.7989**	0.8210	0.8008	**0.7972**
Rural Development	1.2134	0.7302	**0.6936**	0.7606	0.7220	**0.7004**

Fig. 10. MAE Comparison using Various Scalers with CN1D

TABLE IV. PERFORMANCE COMPARISON (IN MAE) BETWEEN RMSPROP AND ADAM OPTIMIZER. RMSPROP GIVES BETTER PERFORMANCES WITH THE RESULTS FORMATTED BOLD. WE REPORT THE AVERAGE PERFORMANCE ON ALL DATASETS AT THE LAST ROW (AVG).

Dataset	RMS	Adam
Education	**0.5882**	0.6220
Environment & Natural Resources	**0.5979**	0.6461
Economics	**0.5753**	0.6092
Foreign Languages	**0.4910**	0.5031
Social Sciences and Humanities	0.5672	0.5647
Aquaculture & Fisheries	**0.5934**	0.6377
Law	**0.5566**	0.5878
Political Science	**0.5556**	0.5558
Mekong Delta Development Research Institute	**0.5541**	0.5587
Agriculture	**0.5804**	0.5814
Biotechnology Research & Development Institute	**0.5763**	0.5855
Physical Education	0.6837	0.6720
Engineering Technology	**0.7556**	0.8217
Information and Communications Technology	**0.7152**	0.7637
Natural Sciences	**0.7740**	0.7972
Rural Development	**0.6989**	0.7004
AVG	0.6165	0.6379

D. Student Performance Prediction with Shallow Deep Learning Architectures

The performance of three learning models with the RMSprop optimizer and QTF scaler is shown in Table V. We almost achieve the best performance with deep learning (13 datasets in MAE and 12 datasets in RMSE out of 16 datasets, respectively).

CN1D holds the first place in both MAE and RMSE with achieving the best results on 9 datasets in MAE and 10 datasets in RMSE. The best MEA is obtained on Foreign Languages dataset while the worst in the prediction results are for Engineering Technology dataset. Using the CN1D model, 11 datasets have MAEs which are lower than 0.6 while there are two datasets which get high MAEs being greater than 0.7. On the other hand, the results are rather similar to the metric of RMSE. The best RMSE (0.64607) is achieved when marks come from Foreign Languages. The poor prediction results of

Engineering Technology samples can be explained by label distribution as shown in Fig. 4 and Fig. 5 where we can observe that the distribution of training set and test set exist some differences at the mark level of 3.5 while the low results of Rural Development dataset seem to be that the number of samples in the training set is even less than the number of samples in the test set. The models may not obtain enough data in the training set to capture the characteristics in the test set. Predicting marks from Physical Education students can be challenging because of special characteristics from this department where each student can own special talents. However, it can be that many of them usually have to focus and concentrate on various Sports competitions, and hence, he or she cannot spend more time to performs well on other subjects.

TABLE V. THE PREDICTION PERFORMANCE WITH VARIOUS LEARNING ALGORITHMS IN RMSE AND MAE USING QTF SCALER AND RMSPROP (RMS) OPTIMIZER FUNCTION. THE BEST RESULTS OF EACH DATASET ARE FORMATTED WITH BOLD TEXT

Dataset	MAE			RMSE		
	CN1D-RMS	LSTM-RMS	Linear Regression	CN1D-RMS	LSTM-RMS	Linear Regression
Education	**0.57331**	0.58822	0.58366	**0.73217**	0.74917	0.76199
Environment & Natural Resources	0.59888	**0.59792**	0.61096	0.75927	**0.75896**	0.76621
Economics	0.59218	**0.57528**	0.58187	0.74045	**0.72979**	0.73904
Foreign Languages	**0.48534**	0.49103	0.50141	**0.64607**	0.64981	0.65925
Social Sciences and Humanities	0.59204	**0.56724**	0.57305	0.75332	0.73727	**0.73185**
Aquaculture & Fisheries	**0.59180**	0.59338	0.59882	**0.75729**	0.75875	0.76099
Law	**0.55460**	0.55659	0.56638	**0.69150**	0.69367	0.70134
Political Science	0.57651	**0.55560**	0.56474	0.72644	0.74287	**0.70574**
Mekong Delta Development Research Institute	0.56779	0.55407	**0.54695**	0.70561	0.69634	**0.69463**
Agriculture	0.58055	0.58043	**0.57992**	**0.73776**	0.74044	0.74419
Biotechnology Research & Development Institute	**0.53296**	0.57634	0.54973	**0.69808**	0.77467	0.71557
Physical Education	**0.67617**	0.68367	0.68269	**0.82931**	0.83464	0.85973
Engineering Technology	**0.74536**	0.75564	0.75749	**0.90762**	0.91895	0.92585
Information and Communications Technology	**0.69027**	0.71518	0.73927	**0.84843**	0.87515	0.89998
Natural Sciences	**0.67251**	0.77397	0.71112	**0.84147**	0.99010	0.87977
Rural Development	0.71336	0.69887	**0.69280**	0.85659	0.84676	**0.84369**

Fig. 11. The Performance in Training Phase (train_mae) and Test Phase (test_mae) of the Proposed Models on the Full Dataset in MAE

Fig. 12. The Performance in Training Phase (train_mae) and Test Phase (test_mae) of the Proposed Models on the Full Dataset in RMSE

With the scaler of QTF and RMSprop optimizer, we run the prediction tasks on the full dataset which contains 3,828,879 samples of all departments and institutes of the considered university. The training set includes 2,584,462 samples (67.5%)

recorded from 2007 to 2016 while the test set consists of 1,244,417 samples (32.5%) from 2017 to 2019. The results from the considered classifiers are exhibited in MAE and RMSE in Fig. 11 and Fig. 12, respectively. The LSTM holds the best in both RMSE and MAE with values of 0.79522 and 0.63115, respectively. Another deep learning algorithm also obtains better performance than Linear Regression with the results of 0.64107 and 0.79918 in MAE and RMSE, respectively. The performance in the training phase is lower than the validation phase but the difference is trivial. Comparing by MAE, the difference between training and test phases with LSTM is about 0.02846 while this value is 0.03031 for CN1D.

From these experimental results, we can observe that the proposed models could produce acceptable prediction results, thus, the system could support the students to select appropriate courses and to design suitable study plans. Moreover, student performance prediction enables lecturers as well as educational managers to indicate what students should be monitored and supported to complete their programs with the best results. These supports can reduce formal warnings and expulsions from universities due to students' poor performance.

E. Influence Factor Analysis for Student Performance Prediction

In order to know which features are important for the model to learn, we calculate the Pearson product-moment correlation which indicates the covariance of the two variables. This correlation is particularly helpful for regression problem. Next, we compute each variable's standard deviation. The correlation coefficient is indicated by dividing the covariance by the product of the two variables' standard deviations (Equation 6).

$$p_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \qquad (6)$$

where

Fig. 13. Factor Analysis with Pearson Correlation Coefficient Raking. Features that their Values are Higher than Zero have Rich Information for the Model to Learn

Fig. 14. Influence Factor Explanation from Local Interpretable Model-Agnostic Explanations (LIME) [24] for the Prediction Results. (a) The Prediction is 3.04 while the Real One is 3.0 (with a Low Error). (b) The Model Predicted a Mark of 3.67 while the Real Value is 2.0 (a High Error).

- p_{xy} denotes Pearson product-moment correlation coefficient.

- $Cov(x, y)$ is covariance of variable x and y.

- σ_x exhibits standard deviation of x.

- σ_y denotes standard deviation of y.

The results of factor analysis with Pearson correlation coefficient show that the CGPA and the CourseID have the most correlation to the target attribute (the mark) while the StudentID has negative effect. Other features are presented in Fig. 13.

Taking one bad prediction and one good prediction for the influence factor analysis as shown in Fig. 14, we can see that CGPA (see Table II to get details of features) contributes a positive effect on the mark to produce a good prediction (Fig. 14a) while the result in Fig. 14b considers CGPA as a negative factor so it reveals a prediction with a higher error. An observation from Fig. 14a, we noted that CGPA and the number of semesters (No._Semester) the student studied as well as the course which student was taking also contribute positive effects on the mark.

VI. Conclusion

In this study, we proposed deep learning models (Long Short Term Memory and Convolutional Neural Networks) to predict the student performance prediction problem in educational data mining. We analyze and propose using some techniques for data pre-processing (e.g., Quantile Transforms, MinMax Scaler) before fetching them into deep learning models and robust machine learning such as Linear Regression to do prediction tasks. Moreover, we adapt the models by using different optimizer functions including Adam and RMSprop for improving the prediction performance. Experimental results on the dataset collected from a Vietnamese multidisciplinary university's information system show that the proposed methods provide good prediction results and is expected to apply in practical cases.

Using these results, we can help both the educational managers and the students to know early warning results so that the students can have a better plan for studying. Moreover, evaluating various training courses to help the managers to propose appropriate policies.

In the future, we continue to perform experiments on other published datasets and to change the model setting for better performance as well as to compare with other approaches. Moreover, instead of using one model to predict all of the students, future studies can investigate on separated various groups of students depending on different levels of marks to create group of models for enhancing the prediction performance. Further researches should also take into account sophisticated models which can be potential to improve the performance.

References

[1] U. b. Mat, N. Buniyamin, P. M. Arsad, and R. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention," in 2013 IEEE 5th Conference on Engineering Education (ICEED), 2013, pp. 126-130.

[2] Rastrollo-Guerrero, Juan & Gomez-Pulido, Juan A. & Domínguez, Arturo. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. Applied Sciences. 10. 1042. 10.3390/app10031042.

[3] A. M. Shahiri, W. Husain, and N. a. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," Procedia Computer Science, vol. 72, pp. 414-422, 2015/01/01/ 2015.

[4] Lorenz Kemper, Gerrit Vorhoff & Berthold U. Wigger. Predicting student dropout: A machine learning approach, European Journal of Higher Education, 10:1, 28-47, 2020. DOI: 10.1080/21568235.2020.1718520

[5] Ang, Kenneth Li-Minn, Feng Lu Ge, and Kah Phooi Seng. "Big Educational Data & Analytics: Survey, Architecture and Challenges." IEEE Access 8 (2020): 116392-116414.

[6] Kim, Byung-Hak, Ethan Vizitei and Varun Ganapathi. "GritNet: Student Performance Prediction with Deep Learning." ArXiv abs/1804.07405 (2018): n. pag.

[7] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting Students Performance in Educational Data Mining," in 2015 International Symposium on Educational Technology (ISET), 2015, pp. 125-128.

[8] N. Thai-Nghe, T. Horvath, and L. Schmidt-Thieme. Factorization Models for Forecasting Student Performance, in Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero , C., and Stamper, J. (Eds.) Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011).

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015/05/01 2015.

[10] Tsiakmaki, Maria & Kostopoulos, Georgios & Kotsiantis, Sotiris & Ragos, Omiros. (2020). Transfer Learning from Deep Neural Networks for Predicting Student Performance. Applied Sciences. 10. 2145. 10.3390/app10062145.

[11] T.-N. Huynh-Ly and N. Thai-Nghe, "A system for predicting students's course result using a free recommender system library of MyMediaLite (in Vietnamese)," in Information technology conference 2013, Can Tho University, 2013.

[12] N. P. Hai, T.-W. Sheu, and M. Nagai, "Predicting the Student Learning Outcomes Based on the Combination of Taylor Approximation Method and Grey Models (in Vietnamese)," VNU Journal of Science: Education Research, vol. 31, no. 2, pp. 70-83, 2015.

[13] Z. Iqbal, J. Qadir, A. Mian, and F. Kamiran, "Machine Learning Based Student Grade Prediction: A Case Study," 08/17 2017.

[14] L. Zhang, T. Luo, F. Zhang, and Y. Wu, "A Recommendation Model Based on Deep Neural Network," IEEE Access, vol. 6, pp. 9454-9463, 2018.

[15] M. Fu, H. Qu, Z. Yi, L. Lu, and Y. Liu, "A Novel Deep Learning-Based Collaborative Filtering Model for Recommendation System," IEEE Transactions on Cybernetics, vol. 49, no. 3, pp. 1084-1096, 2019.

[16] J. Rastrollo-Guerrero, J. A. Gomez-Pulido, and A. Domínguez, "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review," Applied Sciences, vol. 10, p. 1042, 02/04 2020.

[17] B. Sekeroglu, K. Dimililer, and K. Tuncal, Student Performance Prediction and Classification Using Machine Learning Algorithms. 2019, pp. 7-11.

[18] Arto Hellas, Petri Ihantola, Andrew Petersen, Vangel V. Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, and Soohyun Nam Liao. Predicting academic performance: a systematic literature review. In Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion). Association for Computing Machinery, New York, NY, USA, 175–199. DOI:https://doi.org/10.1145/3293881.3295783

[19] 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, p. 3, 2020/02/10 2020.

[20] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Transfer Learning from Deep Neural Networks for Predicting Student Performance," Applied Sciences, vol. 10, p. 2145, 03/21 2020.

[21] N. Nguyen Thai, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in 2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007, pp. T2G-7-T2G-12.

[22] T. H. Nguyen and J. Zucker, "Enhancing Metagenome-based Disease Prediction by Unsupervised Binning Approaches," in 2019 11th International Conference on Knowledge and Systems Engineering (KSE), 2019, pp. 1-5.

[23] S. Ruder, "An overview of gradient descent optimization algorithms."

arXiv preprint arXiv:1609.04747, 2016.

[24] Marco Tulio Ribeiro et al. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. https://arxiv.org/pdf/1602.04938.pdf. 2017.

[25] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, 01/02 2012.

Scalability Validation of the Posting Access Method through UPPAAL-SMC Model-Checker

Bethaina Touijer[1]
LRIT, Rabat IT Center, Faculty of
Sciences, Mohammed V University
in Rabat, Morocco

Yann Ben Maissa[2]
Telecommunication Systems Networks and
Services Laboratory, National Institute
of Posts and Telecommunications,
Rabat, Morocco

Salma Mouline[3]
LRIT, Rabat IT Center, Faculty of
Sciences, Mohammed V University
in Rabat, Morocco

Abstract—The standard IEEE 802.15.6 provides a new phys-ical layer (PHY) and medium access control sublayer (MAC) specifications that support several challenges of wireless body area networks (WBANs). The posting is the access method of the IEEE 802.15.6 MAC protocol that is used by the hub to send data to the nodes. In this paper, we use a formal method to evaluate the posting access method under the WBANs stochastic environment. Based on the statistical model checking (SMC) toolset UPPAAL-SMC, we model and evaluate the behavior of the posting access method in terms of scalability. The evaluation results showed that according to the allocated time intervals, the energy consumption, and the throughput the scalability was validated.

Keywords—*WBANs; IEEE 802.15.6 MAC protocol; posting access method; UPPAAL-SMC; energy consumption; throughput*

I. Introduction

Wireless body area network (WBAN) [1] [2] [3] [4] is composed of biomedical sensors nodes that can be worn on or placed in the human body to measure several physiological parameters of the human body, such as temperature and pressure. These sensors nodes must wirelessly send their data to a control and monitoring device carried on the body. This device then delivers its data via a cellular or Internet network to an emergency center or a doctor room.

The WBANs [1] [2] [3] [4] [5] are designed to support a wide range of medical applications, such as asthma, cardio-vascular disease, and cancer detection. They have enormous potential to revolutionize the future of healthcare by diagnosing many deadly diseases and providing remote and real-time monitoring of patients' health status without any restrictions, which improve their quality of life and reduce their costs of hospitalization. On the other hand, they impose several challenges related to the medium access control (MAC) pro-tocols design in terms of energy efficiency, quality of service, reliability, priority, security, and scalability.

In November 2007, the IEEE 802 created a standard called IEEE 802.15.6 for WBANs, its final version [6] is published in February 2012. The main idea behind this standard is to define two new layers that are the physical (PHY) layer and the MAC sublayer dedicated to WBANs.

The IEEE 802.15.6 standard [6] [7] organizes the nodes into one- or two-hop star topology. A single control and monitoring device controls the entire operation of each WBAN.

The WBAN must have one control and monitoring device (i.e., the hub) and a number of sensors nodes, ranging from 0 to 64. The node(s) will refer, in the rest of this paper, to the sensor(s) node(s).

Following our earlier study [8] regarding the existing MAC protocols for the WBAN, we found that the MAC protocol of the IEEE 802.15.6 standard takes into consideration all the WBAN challenges. With this MAC protocol, the hub and the nodes can employ one or more access methods to transmit their data frames. These access methods are the carrier sense multiple access with collision avoidance (CSMA/CA), the slotted aloha, the type-I polling, the type-II polling, and the posting.

The posting is the most important access method for the hub. This latter uses it to grant itself a posted allocation for initiating one or more frame transactions. A posted allocation is a downlink allocation time interval, during which the hub can service and transmit unexpected or extra management and data traffic to the node. For example, it is used in the case of network management needs, data rate variations, and channel impairments.

Due to the important role of the hub in the WBAN, as a controller and monitor device, evaluating the scalability of its posting access method is important. Scalability represents the ability of the network to continue to operate with the same performance despite the addition of other nodes [1]. Validating this property allows the validation of the posting access method and, therefore, the performance of the WBAN.

The problem here is that the WBAN is considered as a stochastic environment, where the prediction of when the phys-iological parameters change their values is non-deterministic. This stochastic nature makes it difficult for the prediction of who it should allocate, between the hub and the nodes, the time interval to transmit their frames. Consequently, it makes it difficult to model and evaluate the behavior of the posting access method.

In this paper, we propose to use the statistical model check-ing (SMC) toolset UPPAAL-SMC [9] to investigate the posting access method under the WBANs stochastic environment, as depicted in Fig. 1. The UPPAAL-SMC has the ability to provide a stochastic interpretation of the stochastic behavior of complex and real-time systems, such as WBANs, and it is based on the statistical model-checking (SMC) [10]. The model-checking [11] has been used primarily in the verification

Fig. 1. The WBAN Illustration and the Used Approach for the Communication Method Modeling and Evaluation. EEG: ElectroEncephaloGraph. ECG: ElectroCardioGraph. SpO2: Blood Oxygen Saturation Level.

of the synchronization protocols, and it is considered as a useful method for the analysis and the evaluation of the communication protocols [11] [12].

Based on UPPAAL-SMC, we model and evaluate the behavior of the posting access method in terms of scalability. This is the first study of the posting access method through UPPAAL-SMC in our best knowledge. We first use the stochastic timed automata (STA) formalism provided by UPPAAL-SMC to construct a detailed model of this behavior. Then we use the metric interval temporal logic (MITL) specifications adopted by UPPAAL-SMC to evaluate the scalability of this behavior. In addition to evaluate the scalability according to the energy consumed by the hub and the throughput, which are, respectively, the dominant problem and the key performance properties that we should validate for WBAN. We should, also, evaluate the number of the posted allocation time intervals of the hub. This property shows the ability of the hub to still communicating with the nodes despite the growth of the network density.

The rest of the paper is organized as follows: the next section describes the behavior of the posting access method. Section 3 provides the posting access method behavior modeling. Section 4 presents the posting access method behavior evaluation. Section 5 presents the paper's conclusion.

II. POSTING ACCESS METHOD: BEHAVIOR DESCRIPTION

The IEEE 802.15.6 MAC protocol, as defined in [6], supports three access modes: the beacon mode with superframes, the non-beacon mode with superframes, and the non-beacon mode without superframes, as depicted in Fig. 2a, Fig. 2b, and Fig. 2c, respectively. The time axis of the latter access mode is divided into time intervals, where the node and the hub can employ one or more access methods. However,

the time axis of the other two access modes is divided into superframes of equal length. The superframe can be divided into one or more access phases, where the node and the hub can employ one or more access methods. The access methods of the IEEE 802.15.6 MAC protocol are divided into five classes that are scheduled, scheduled-polling, unscheduled, improvised, and contention, as depicted in Fig. 2a, Fig. 2b, and Fig. 2c. Accordingly, the posting is the only access method that is used by the hub to transmit its data frames in scheduled-polling, unscheduled, and improvised access methods classes. In this section, we describe the behavior of the posting access method before, within, and after the posted allocation time interval.

1) Before the Posted Allocation Time Interval: Based on the posting access method, to grant to itself a posted allocation time interval, the hub sends to the node a poll frame, as depicted in Fig. 3. This latter is a control frame addressed to the node to inform it about a future post. A post is a management or data frame sent by the hub to the node within a posted allocation time interval. While granting the posted allocation time interval, the hub can start sending posts after a pre-determined time.

2) Within the posted allocation time interval: When the posted allocation time interval starts, the hub can transmit one or more new frames and it can retransmit one or more old frames. These frames are separated by a short inter-frame spacing (*pSIFS*) time. The hub transmits the frame with a required immediate or block acknowledgment frame and with the more data (M) and the last frame (L) fields of the MAC frame header. The values of these fields can be:

1) Case 1, with $M = 0$ and $L = 0$: it means that no frames are waiting for transmission. Other than

(a) Beacon Mode with Superframes

(b) Non-Beacon Mode with Superframes

(c) Non-Beacon Mode without Superframes

Fig. 2. Access Modes of the IEEE 802.15.6 MAC Protocol [6].

Fig. 3. Posting Access Method Illustration: The Future Posts Transactions [6].

probable retransmission of the last frame in the currently posted allocation time interval. That one is most likely due to the no reception of the required acknowledgment frame. Otherwise, the hub should relinquish and reclaim the posted allocation time interval if it has received the required acknowledgment frame from the node.

2) Case 2, with $M = 1$ and $L = 0$: it means that one or more frames are waiting for transmission or retransmission after a *pSIFS* time in the currently posted allocation time interval.

3) Case 3, with $M = 0$ and $L = 1$: it means that no frames are waiting for transmission. Other than probable retransmission of the last frame in the next posted allocation time interval. That one is most likely due to the no reception of the required ac-

knowledgment frame. As well, is due to not enough time remaining in the currently posted allocation time interval for completing another frame transaction plus an appropriate guard time. Thus, the hub should relinquish and reclaim the currently posted allocation time interval.

4) Case 4, with $M = 1$ and $L = 1$: it means that one or more frames are waiting for transmission or retransmission in the next posted allocation time interval. That one is most likely due to not enough time remaining in the currently posted allocation time interval for completing another frame transaction plus an appropriate guard time. In this case, the hub should relinquish and reclaim the currently posted allocation time interval.

After sending the required acknowledgment frame to the hub, the node behaves according to the cases defined above. For the first case, it should be ready to receive the retransmission of the last frame after a *pSIFS* time. As well, it should relinquish the currently posted allocation time interval after a time out (*mTimeOut*) if at this time it has not received it. As for the second case, it should be ready to receive the transmission or the retransmission of one or more frames in the currently posted allocation time interval. While for the third and fourth cases, it should relinquish the currently posted allocation time interval.

3) After the posted allocation time interval: After the end of the posted allocation time interval or after reclaiming it, the hub can send to the node a poll frame conveying an immediate

or a future new posted allocation time interval, extending the remaining or the existing one if there are other frames to transmit. Moreover, it can cancel the future posted allocation time interval by sending a poll frame to the node before the start of it.

Thereafter, we will model the detailed communication between the hub and the nodes through the posting access method and within the stochastic environment of WBAN.

III. POSTING ACCESS METHOD: BEHAVIOR MODELING

The UPPAAL-SMC, as defined in [9], is an alternative formalism to the timed automata limits. The timed automata formalism, as defined in [13] [14], is not flexible and expressive enough to model the stochastic behavior of systems. The UPPAAL-SMC formalism is based on the timed automata formalism and a stochastic interpretation, thus it generates stochastic timed automata (STAs). A model in UPPAAL-SMC consists of a network of STAs interacting components. The STAs components communicate via broadcast channels and share variables to generate networks of stochastic timed automata (NSTAs).

In this section and based on the STA formalism of the UPPAAL-SMC model-checker, we model the whole behavior of the posting access method before, within, and after the posted allocation time interval. As well, we model the parts of the node behavior that is associated with it. The resulting model is a network of stochastic timed automata (NSTA) that is composed of a couple of two templates, as depicted in Fig. 4. Moreover, we mention in Table I and Fig. 5 the intervals of the random functions, the values of the parameters used in this model, and the illustration of these parameters value determination.

1) Behavior modeling of the hub and the nodes before the posted allocation: Consider the Hub template, as depicted in Fig. 4(a). It starts with the stochastic interpretation of its non-deterministic choice between allocating time intervals for itself or to the node. Indeed, according to the data sensed by the node or the data received by itself, the hub determines its choice. The node allocated time interval starts at the location *ImmNodeAlloc*. The hub allocated time interval (i.e., the posted allocation time interval) starts after an (X) time-units. The X is selected randomly through the random function *rand* $(e : rand)$. The posted allocation is considered as a future allocation compared to the allocation of the node. This latter is considered as an immediate or future allocation. In this section, we model the allocation of the node as an immediate allocation.

To inform the node about its allocated time interval, the hub sends to it a poll frame. In this case, it sends the signals $(IP[id]!)$ and $(EP[id]!)$. These signals indicating, respectively, the start and the end of the poll frame. While in the case of the hub allocated time interval, it sends to the node the signals $(FP[id]!)$ and $(EP[id]!)$. The same as the first signals, these signals indicating, respectively, the start and the end of the poll frame.

We model the communication between the hub and the nodes of the network by a random selection of a node (i)

from the network to communicate with the hub. We use the function *randid* $(eid : randid)$ for this random selection. Then, we put in it the variable $(id)(id = eid)$. This latter represents the identifier of the selected node. As well, we define the time interval (M) of the node allocation randomly through the function *randm* $(em : randm)$.

We define the posted allocation time interval and the required time of the frame transmission by a random selection: $(ep:randp)$ represents the random selection of the posted allocation time interval, and $(ef:randf)$ represents the random selection of the frame transmission time. Then, we put them in the variables $(hp = ep)$ and $(hf = ef)$, respectively. The variable (MD) used in the Hub template, indicates to the Node template if the hub has more data to transmit $(MD == 1)$ or not $(MD == 0)$. This variable is determined randomly by the function *randmd* $(emd : randmd, MD = emd)$. The clock (c) is used in the Hub template to compute the current time of the posted allocation time interval. It starts after the locations *PostedAllocationStart* or *HpSIFS2* and resets at the end of the posted allocation time interval, in the location *PostedAllocationEnd*. The clock (h) is used locally to compute the time in the locations.

2) Behavior modeling of the hub and the nodes within the posted allocation: Once the posted allocation time interval starts (i.e. immediately after the locations *PostedAllocation-Start* or *HpSIFS2*), the Hub starts the transmission of its frame by sending to the Node the synchronization signal $(IS[id]!)$. After the transmission time of the frame $(h == hf)$, the Hub sends to the Node the synchronization signal $(ES[id]!)$ indicating the end of the frame transmission. Then, it moves to the location *HWAck* waiting for the reception of the acknowledgment frame:

1) If the Hub receives from the Node the synchronization signals $(Ak[id]?)$ and $(EAk[id]?)$ that are indicating, respectively, the start and the end of the immediate acknowledgment frame transmission, it moves to:

    a) The transmission of a new frame with new frame transmission time $(ef:randf)$, if it has more data to transmit $(MD == 1)$ and if there is enough time in its posted allocation time interval $(c \leq hp-(ef+tg))$. This transmission will start after staying in the location *HpSIFS4* for the pSIFS time $h == 1$. The guard $(c \leq hp - (ef + tg))$ represents the L state (in this case $L = 0$) and its value is determined as depicted in Fig. 5.

    b) The location *PostedAllocationEnd*, if it has more data to transmit $(MD == 1)$, but it has no more time to complete its frame transaction $(c > hp - (hf + tg)$, in this case $L = 1)$.

    c) The location *PostedAllocationEnd* if it has no more data to transmit $(MD == 0)$. The Hub resets the clock c before moving to the location *PostedAllocationEnd*, which means the relinquishment of the posted allocation time interval.

2) If the Hub has not received during $(h == at)$ time-units the synchronization signal $(Ak[id]?)$, it should

(a) Hub Template

(b) Node Template

Fig. 4. NSTA Model of the Posting Access Method.

retransmit, after staying in the location HpSIFS for ($h == 1$), its last frame if it has enough time in its current posted allocation ($c \leq hp - (hf + tg)$). Otherwise, it resets the clock c and moves to the location *PostedAllocationEnd*.

Consider the Node template, as depicted in Fig. 4(b). After receiving from the Hub the synchronization signals ($IS[id]?$) and ($ES[id]?$), the Node waits for a *pSIFS* time ($n == 1$) in the location *NpSIFS1*. Then, it moves to the transmission of the acknowledgment frame, which starts and ends after sending to the Hub the synchronization signals ($Ak[id]!$) and ($EAk[id]!$), respectively. After that, the Node moves to the location *NTimeOut*, where it waits for (tg) time-units the reception of the signals ($IS[id]?$), ($ST[id]?$), or ($Fin?$). The

value of tg is determined as depicted in Fig. 5. If at this time it has not received any signal from the Hub (e.g., the hub is broken down), the Node relinquishes the current allocation time interval by returning to the location *Nstart*. The signal ($ST[id]?$) indicates the transmission of a new frame or the retransmission of the last frame. The signal ($IS[id]?$) indicates the start of a newly posted allocation time interval and the start of the first Post. While, the signal ($Fin[id]?$) indicates the end of the allocation time interval of the hub.

The acknowledgment frame: missing situation: we consider that the hub and the nodes are connected. As mentioned in Fig. 3, the hub sends to the node the poll frame without a required acknowledgment frame. The problem is that in the case of the transmission loss of the poll frame, the hub

can not synchronize with the node when the posted allocation starts. Therefore, this situation allows the data frame transmission loss. We suppose that the posting access method treats this situation through the requirement of an acknowledgment frame transmitted with the data frame, as explained before in Section II. However, the posting access method has not treated the case of the transmission loss of this acknowledgment frame. To prevent that the hub stays blocked until the end of the posted allocation time interval waiting for the acknowledgment frame, we proposed in our model a bounded time (at) during with the Hub automaton can stay in the location *HWAck*. Its value is determined as depicted in Fig. 5.

3) Behavior modeling of the hub and the nodes after the posted allocation: The blue lines represent the Hub decision after the end of the allocation time interval of itself or for the node. After the end of the allocation time interval, we use another stochastic interpretation, with the same probability weight (i.e., the probability weight as called in UPPAAL-SMC $Pw = 1$), to model the non-deterministic choice of the hub between extending the existing allocation and granting a probably immediate or a future new allocation time interval.

Based on this accurate behavior model of the posting access method and according to the values of the parameters used in this model, we will evaluate in the next section, the performance of this access method.

Fig. 5. Illustration of the Parameters Value Determination. GT: Guard Time. Ack: Acknowledgment.

TABLE I. PARAMETERS VALUES AND RANDOM FUNCTIONS INTERVALS.

$rand$	$[5, 10]$	$randp$	$[10, 25]$
$randf$	$[3, 5]$	$randm$	$[10, 25]$
$randmd$	$[0, 1]$	a	3
tg	7	at	5
$pSIFS$	1	GT	1

IV. POSTING ACCESS METHOD: BEHAVIOR EVALUATION

To specify properties over NSTAs, UPPAAL-SMC uses a weighted extension of the MITL [15]. Besides, to analyze these properties, UPPAAL-SMC uses the SMC [10]. This latter uses the Monte Carlo simulation to respond to the quantitative questions (i.e., probability estimation), and it uses sequential hypothesis testing to respond to the qualitative questions (i.e., hypothesis testing and probability comparison). Additionally, UPPAAL-SMC provides the simulation of the system behavior and the evaluation of the expected values of the min or the max expression. In this section, to evaluate the scalability of the posting access method, we use the following property:

- The evaluation of the expected values of max:

$$E[bound; N](max : expr) \qquad (1)$$

Where $bound$ is a time-bound in the evaluation, N is the number of runs, and $expr$ is the expression to evaluate.

In a network of nodes ranging from 4 to 64 and through $N = 10000$ runs of stochastic scenarios generated by UPPAAL-SMC, we evaluate the scalability. This latter, we evaluate it in terms of the number of the allocated time intervals of the hub and nodes. As well, the energy consumed and the successful transmitted frames by the hub during a determined period of time.

In the experiments, we use three networks and three periods of time ($T1$, $T2$, and $T3$) to evaluate and visualize the scalability of the posting access method. The first network is composed of one hub and 4 nodes. The second network is composed of on hub and 16 nodes. As for the third network is composed of one hub and 64 nodes. Moreover furthermore, the first period of time $T1 = 3600$ time-units, the second period of time $T2 = 7200$ time-units, and the third period of time $T3 = 10800$ time-units.

A. Allocated Time Intervals

Using Equation 1, we evaluate the average of the maximum number of allocated time intervals of the hub and the nodes within the three networks and during the three periods of time:

$$E[<= T; N](max : Hub.HA) \qquad (2)$$

$$E[<= T; N](max : Hub.NA) \qquad (3)$$

These formulae compute, in the interval of time T (i.e., the T can be $T1$, $T2$, or $T3$) and using N runs, the average of the maximum value of the counters (HA) and (NA).

In our NSTA model, as depicted in Fig. 4, the Hub template uses the HA and NA counters to compute the number of allocated time intervals of the hub and the nodes, respectively. The Hub automaton increments the counter ($HA + +$) when the posted allocation time interval starts. As well, it increments the counter ($NA + +$) when the node allocated time interval starts.

UPPAAL-SMC estimates the averages of Equations 2 and 3 to be in the confidence intervals, as depicted in Tables II and III, respectively. As well, Fig. 6 and 7 present the visualization of the results for the hub and the nodes, respectively. The results show that within the three networks, the hub and the nodes retain the same number of the allocated time intervals during each period of time. Along with this, we remark that the hub allocates the double of the time intervals number compared to the nodes.

TABLE II. CONFIDENCE INTERVALS OF THE ESTIMATED AVERAGES OF THE MAXIMUM NUMBERS OF ALLOCATED TIME INTERVALS FOR THE HUB.

Parameter evaluated: HA		
T	4 Nodes	16 Nodes	64 Nodes
$T1$	161.860 ± 0.248666	161.857 ± 0.249261	161.824 ± 0.246545
$T2$	323.919 ± 0.347292	323.716 ± 0.347743	323.389 ± 0.351441
$T3$	485.783 ± 0.431237	485.471 ± 0.433604	485.899 ± 0.430501

Fig. 6. Averages of the Maximum Numbers of Allocated Time Intervals for the Hub within Three Networks and During Three Periods of Time.

TABLE III. CONFIDENCE INTERVALS OF THE ESTIMATED AVERAGES OF THE MAXIMUM NUMBERS OF ALLOCATED TIME INTERVALS FOR THE NODES.

Parameter evaluated: NA		
T	4 Nodes	16 Nodes	64 Nodes
$T1$	81.2289 ± 0.163605	81.3295 ± 0.163182	81.4821 ± 0.162648
$T2$	162.254 ± 0.23184	162.237 ± 0.231811	162.305 ± 0.229111
$T3$	243.392 ± 0.285667	243.014 ± 0.285128	243.268 ± 0.281951

B. Energy Consumption

Using Equation 1, we evaluate the average of the maximum value of the energy consumed by the hub within the three networks and during the three periods of time:

$$E[<= T; N](max : Hub.E) \qquad (4)$$

This formula computes, in the interval of time T (i.e., the T can be $T1$, $T2$, or $T3$) and using N runs, the average of the maximum value of the variable (E).

In our NSTA model, as depicted in Fig. 4, the Hub template uses the variable E. This latter computes the energy consumed

Fig. 7. Averages of the Maximum Numbers of Allocated Time Intervals for the Nodes within Three Networks and During Three Periods of Time.

proportionally to the time spent by the hub when it passes by the locations that have ($E == 1$).

UPPAAL-SMC estimates the averages of Equation 4 to be in the confidence intervals, as depicted in Table IV. As well, Fig. 8 presents the visualization of the results. These latter show that within the three networks, the hub retains the same value of the energy consumed during each period of time.

TABLE IV. CONFIDENCE INTERVALS OF THE ESTIMATED AVERAGES OF THE MAXIMUM VALUES OF THE ENERGY CONSUMED BY THE HUB.

Parameter evaluated: E		
T	4 Nodes	16 Nodes	64 Nodes
$T1$	3595.17 ± 0.107094	3595.17 ± 0.105837	3595.10 ± 0.107957
$T2$	7195.17 ± 0.106899	7195.15 ± 0.107162	7195.21 ± 0.105604
$T3$	10795.2 ± 0.106891	10795.2 ± 0.106597	10795.2 ± 0.105469

Fig. 8. Averages of the Maximum Values of the Energy Consumed by the Hub within Three Networks and During Three Periods of Time.

C. Throughput

Using Equation 1, we evaluate the average of the maximum number of the successful transmitted frames by the hub within the three networks and during the three periods of time:

$$E[<= T; N](max : Hub.SucTx) \tag{5}$$

This formula computes, in the interval of time T (i.e., the T can be $T1$, $T2$, or $T3$) and using N runs, the average of the maximum number of the counter ($SucTx$).

In our NSTA model, as depicted in Fig. 4, the Hub template uses the counter $SucTx$ to compute the number of successful transmitted frames by the hub to the nodes. The Hub automaton increments the counter ($SucTx + +$) when it receives the acknowledgment frame from the node.

UPPAAL-SMC estimates the average of Equation 5 to be in the confidence intervals, as depicted in Table V. As well, Fig. 9 presents the visualization of the results. These latter show that within the three networks, the hub almost retains the same number of throughput during each period of time.

TABLE V. CONFIDENCE INTERVALS OF THE ESTIMATED AVERAGES OF THE MAXIMUM NUMBERS OF SUCCESSFUL TRANSMITTED FRAMES BY THE HUB.

Parameter evaluated: $SucTx$		
T	4 Nodes	16 Nodes	64 Nodes
$T1$	203.968 ± 0.31096	203.873 ± 0.312695	203.880 ± 0.312903
$T2$	408.322 ± 0.452698	408.563 ± 0.442229	408.371 ± 0.445739
$T3$	613.109 ± 0.554558	612.956 ± 0.545412	612.775 ± 0.540818

Fig. 9. Averages of the Maximum Numbers of Successful Transmitted Frames within Three Networks and During Three Periods of Time.

D. Results

From these evaluation results, we interpret the constant values obtained as that during each period of time T the hub allocates a fixed number of time intervals for itself and for the nodes. Accordingly, the values of the energy consumed and the throughput remain constant, even with the increase of the network density. Hence, extending the period of time T, we can get more time intervals, as well as more throughput rate and more energy consumption. As we interpret the double number of the allocated time intervals of the hub compared to the nodes as that the posting access method, as modeled in Section III, provides the priority to the hub to allocate more time intervals to still communicating with the nodes despite the growth of the network density. This is due to its important role in the network as a controller and a monitor device.

As a result, with the posting access method, the hub works with the same performance, even with the increase of the network density. Thus, validating the scalability of the posting access method, and therefore, validating the scalability of the WBAN.

V. CONCLUSION

In this paper, we used the statistical model checking toolset UPPAAL-SMC to investigate the posting access method under the WBANs stochastic environment. Based on UPPAAL-SMC, we modeled and evaluated the behavior of the posting access method in terms of scalability. We first used the stochastic timed automata formalism to construct a model of this behavior. Then we used the metric interval temporal logic specifications to evaluate it. The evaluation results showed that, with the posting access method, the hub works with the same performance in terms of the allocated time intervals, the energy consumed, and the throughput, even with the increase of the network density. The thing that validated the scalability of the posting access method and, therefore, the scalability of the WBAN.

Supplement to the first conclusion, we came to learn from this case study that with UPPAAL-SMC, we can model and evaluate accurately the behavior of the MAC protocols. Nevertheless, modeling and evaluating the behavior of MAC protocols through the stochastic timed automata and the metric interval temporal logic specifications, adopted by UPPAAL-SMC, necessitate a certain level of expertise in this side. The thing that is not available for many MAC protocol designers.

As future work, we will combine the posting access method with other access methods to construct a new MAC protocol for WBANs. Furthermore, to facilitate the use of Uppaal-SMC powerful modeling and analyzing algorithms, we propose to define a model-driven engineering approach that uses a domain-specific modeling language (DSML) as a start and the UPPAAL-SMC as a target and back. This DSML should be dedicated to the MAC protocols of WBANs.

REFERENCES

[1] S. Movassaghi et al: Wireless body area networks: A survey. IEEE Communications Surveys & Tutorials, 1658–1686 (2014)

[2]   B. Latré et al: A survey on wireless body area networks. Wireless Networks, 1–18 (2011)

[3]   Y. Ben Maissa et al: Modeling and analyzing wireless sensor networks with verisensor: An integrated workflow. Transactions on Petri Nets and Other Models of Concurrency VIII, Springer, 24–47 (2013)

[4]   Z. Mohammadi et al: New high-rate UWB scheme for WBAN-based healthcare systems. Progress In Electromagnetics Research, EMW Publishing, 125–139 (2014)

[5]   M. Chen et al: Body area networks: A survey. Mobile Networks and Applications, 171–193 (2011)

[6]   IEEE Standards Association and others: IEEE Standard for Local and Metropolitan Area Networks—Part 15.6: Wireless Body Area Networks. IEEE Standard for Information Technology, IEEE, 1–271 (2012)

[7]   S. Ullah et al: A review of IEEE 802.15. 6 MAC, PHY, and security specifications. International Journal of Distributed Sensor Networks, 950–704 (2013)

[8]   B. Touijer et al: MAC protocols for Wireless Body Area Networks: An overview. Wireless Communications and Mobile Computing Conference (IWCMC), 1227–1232 (2017)

[9]   A. David et al: Time for statistical model checking of real-time systems. International Conference on Computer Aided Verification, Springer, 349–355 (2011)

[10]   E.M. Clarke et al: Statistical model checking for cyber-physical systems. International Symposium on Automated Technology for Verification and Analysis, Springer, 1–12 (2011)

[11]   E.A. Emerson et al: The beginning of model checking: A personal perspective. 25 Years of Model Checking, Springer, 27–45 (2008)

[12]   E.M. Clarke et al: Formal methods: State of the art and future directions. ACM Computing Surveys (CSUR), 626–643 (1996)

[13]   R. Alur et al: A theory of timed automata. Theoretical Computer Science, Elsevier, 183–235 (1994)

[14]   A. Alur et al: Decision problems for timed automata: A survey. Formal Methods for the Design of Real-Time Systems, Springer, 1–24 (2004)

[15]   R. Alur: The benefits of relaxing punctuality. Journal of the ACM (JACM), 116–146 (1996)

A Prototype of an Automatic Irrigation System for Peruvian Crop Fields

Luis Nuñez-Tapia
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Peru

Abstract—**Water is an important factor to sustain life and for such a reason it is necessary to take care of it since this is a limited resource. In Peruvian agriculture; however, there is a high percentage of water wasted, as this activity consumes 92% of fresh water; thus, making Peru the 37th country worldwide in misusing water. Due to the aforementioned and considering that the agricultural sector is an important factor for the Peruvian economy, the current study aims to implement a system for automatic irrigation of crop fields in Peru, with the goal of optimizing the use of water and not to waste it as it usually happens. After the implementation of the first prototype of the irrigation system using an Arduino microcontroller and low-cost electronic components, it could be observed that during the tests, 75 and 76.5% of the water that is normally used for irrigation was saved for a dry rainless and dry rainy patch of crop field, respectively. The monitoring of the humidity of the soil was possible due to bluetooth communication. The presented results show the viability of the system and in a follow-up study, large-scale tests are expected.**

Keywords—*Automatic irrigation; crop fields; Arduino; bluetooth*

I. Introduction

In Peru, the agricultural sector is quite important since great part of the Peruvian population works on this activity. According the Ministry of Agriculture and Irrigation (MINAGRI), there is more than 3 million people who work in family farming from a universe of 3.8 million Peruvian farmers [1]. Therefore, it is necessary to support agriculture since the country has several beneficial factors that allow it to develop. In addition, it is important to highlight that this activity depends on water, which is an essential factor to sustain life and the agricultural production. Unfortunately, in Peru it is not being given the proper use in the agricultural sector and that is not only a national concern but also a global one. It has been indicated that Peru is ranked in the 37th place worldwide for mismanagement of water [2].

The traditional form of irrigation that is widely spread in Peru consist of basically flooding the earth with excessive water till this is visible to the farmers' eyes; bringing with it many negative consequences. According to a Peruvian ex-minister of agriculture, it has been indicated that the excess of water in Peruvian agriculture has caused the loss of crop fields in the last 20 years and fighting this has resulted very expensive [3]. Hence, poor water management in Peruvian agriculture is causing a lot of wasted water and a slow-paced damage to the national economy; and all this is extremely worrying as it reflects a lack of knowledge of an adequate irrigation process. The head of the National Water Authority in Peru has indicated that agriculture consumes the largest amount of water (80%) that is distributed in the country, unlike consumption in homes (10%), which still is very large when compared with international standards [4].

Automatic irrigation systems allow to use water more efficiently, unlike a non-technical system. For instance, a system developed by [5], consists of sensors that measure the water level and a sensor that gauge environmental quantities. Additionally, the system has a transceiver module that establishes a wireless communication to receive the data measured and together with a SIM card, the data are sent to a cloud server, and from there it is monitored by an Android application. The system was tested in the cultivation of strawberries and an optimally use of water was observed. In another study [6], which consists of an internet of things (IoT) system based on Raspberry Pi, it was possible to control the temperature sensors and soil moisture in a crop field. The data was stored in the cloud, and this was monitored by mobile or/and PC. The system allowed to observe constantly weather conditions of the crop field, thus optimizing the use of water. Another work [7] has used an Arduino microcontroller and different sensors to deal with different environmental factors. The authors in the study also implemented a website to display in real time values such as water flow, temperature and soil moisture. Further, the system had irrigation pumps and sprinkles which could be controlled by the implemented website. Overall, this system was very efficient for the water management and also of a very low cost.

In view of Peru's problematic with the misuse of water in the agricultural sector, this project shows the implementation of a prototype of an automatic irrigation system for crop fields. In Section II, the methodology will show the low-cost electronic components, the design of the irrigation system, a flowchart of this, the Android application and the prototype. In Section III, the results and discussion regarding the automatic irrigation system will be presented. Finally, in Section IV the conclusions are indicated.

II. Methodology

A. Electronic Components

Because a low-cost automatic irrigation system is desired, the following most commercial electronic components for the design and implementation were selected:

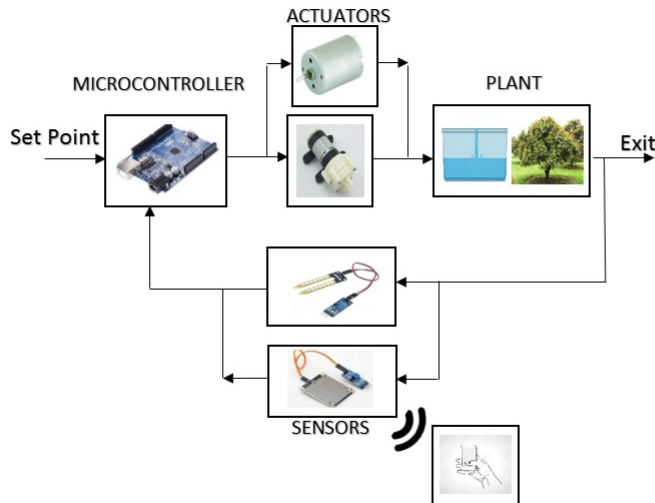- Arduino Uno: This is a microcontroller to receive

Fig. 1. Diagram of the Automatic Irrigation Crop System.

Fig. 2. Android Application to Monitor the Humidity in the Crop Field.

the data the sensors are reading and to control the actuators that will be used.

- YL-69: This is a humidity sensor that will capture the value of moisture of the crop field. Its operation is that while the humidity increases, so does the current.

- Rain sensor: It is a sensor that when detecting rain, a motor is activated to close the roof of the crop field and to avoid excess of water.

- HC-05: This bluetooth module will allow the wireless communication between the Arduino microcontroller and the mobile phone to display the values of the sensors by using an Android application. The bluetooth module allows to connect to a mobile phone without worrying about cables or the nearby position of the devices. It has been indicated that the penetration of smartphones has been increasing in the Peruvian society in the past years [8]. Hence, these could help a lot in the new irrigation systems used in this nation.

- Mini water pump: This is an actuator that, by pumping water, will serve to keep the humidity at a precise and necessary level in the crop field.

- H Bridge: This is necessary to control the roof of the crop field since it allows us to control the DC electric motor direction. The motor turns on if rain is detected by the rain sensor, otherwise it will be off keeping the roof open.

- LCD screen.

B. Design

The automatic irrigation system is an alternative presented in this project to optimize the use of water and thus displace the traditional irrigation system that is used in almost all of the fields of cultivation across Peru. All this will be possible through a reliable and economical design, as shown in Fig. 1.

When the input signal or set point enters the Arduino microcontroller, this is compared to the reference value. This

reference value depends on the type of crop, in this case for example, if the system needs to be tested in a crop field for strong avocado, according to the MINAGRI, it would have a reference value of 65–70% of humidity [9]. If the values are different, the microcontroller sends that signal to the actuator so that both values ares equal through means of irrigation. Being a closed loop system, the humidity sensor feeds back the system and according to the current humidity that is presented, it is compared again by the microcontroller. When the values are equal the microcontroller deactivates the actuator and the irrigation system stops working.

In Algorithm 1, the pseudocode of the main part of the program that operates the automatic irrigation system is shown. In this case, for the present project an ideal humidity range was considered, taking into account that the soil from the coast region in Peru is humid. Peru has three main natural regions, the coast, the highlands and the jungle [10].

Fig. 3. Prototype of the Automatic Irrigation System.

Algoritmo 1 Pseudocode of the Automatic Irrigation System

```
 1: if humidity<=1% & humidity>= 0%  then
 2:     Pump = OFF
 3:     Lcd = disconnected sensor
 4:     Led indicator = Blue
 5: else if humidity<=28% & humidity>= 2%  then
 6:     Pump = ON
 7:     Lcd = irrigation activated
 8:     Led indicator = Green
 9: else if humidity<=50% & humidity>= 29%  then
10:     Pump = OFF
11:     Lcd = Humid soil
12:     Led indicator = Yellow
13: else if humidity<=100% & humidity>= 51%  then
14:     Pump = OFF
15:     Lcd = Excess of humidity
16:     Led indicator = Red
17:     Roof = ON
18: end if
```


Fig. 4. Humidity Levels of References.

Fig. 5. Amount of Water used versus the Time Needed for its Uses.

The application in the smartphone is shown in Fig. 2. This app, which was developed using App Inventor, displays the humidity sensor values. App Inventor (https://appinventor.mit.edu/) is a platform that allows you to develop applications using internal block diagrams. The developed app allows visualizing the values that the sensors present. The connection between the app and the Arduino is possible by using bluetooh, which establishes a connection between the system and the smartphone.

III. RESULTS AND DISCUSSION

For the implementation of the prototype of the automatic irrigation system, main materials such as wood were used to build the sliding roof of the small-scale crop field (29 x 16 cm); the dimensions of the sliding roof are of a 15 x 14 cm. The whole circuit plus the crop field with the hovering roof can be seen in Fig. 3.

For the present project, a control test strategy was carried out in which four references (ref) were considered for the levels of humidity detected by the humidity sensor. Ref1 (between 0% and 1%) will indicate when the sensor is disconnected or does not detect soil moisture (between 0% and 1%). Ref2 (between 2% and 28%) indicates that it is necessary to carry out the irrigation by sending a signal to the actuator (the pump). Ref3 (between 29% and 50%) will indicate that the ground is in the humidity range optimum for the crop, so it will deactivate the actuator (the pump). Ref4 (between 51% and 100%) indicates if it exists an excess of water in the crop; in case it rains on the crop field, the sliding roof will be activated as shown in Fig. 4.

To obtain results that show the efficiency of the automatic irrigation system, four tests were carried out with the prototype. The amount of water used in all the tests was 2 Liters (L) considering a range of soil moisture between 29% and 50%, because this is the ideal humidity so that crops can grow in an optimal environment. Each test was performed every 4 hours. In addition, it must be taken into account that the soil of the coast is humid, so this humidity range was also accounted for. The first test was carried out with the automatic irrigation

system having dry and rainless land, the amount of water that was used was of 500 mL and the amount of water left (i.e., not used by the system) was 1.5 L, which represents 75% water saving. This whole irrigation process lasted an approximate time of 3 min and 5 sec. The second test used the automatic irrigation system with dry and rainy land, the water used by the system was 470 mL and the amount of water left was 1.53 L, which represents a 76.5% water saving taking this second test an approximate time of 3 min and 1 sec. The third test considered traditional irrigation with dry and rainless land, the amount of water used was 900 mL and the amount of water left was 1.1 L, which represents 55% of water saving, all this in an approximate duration time of 1 min and 9 sec. Finally, the fourth test was carried out considering a traditional irrigation system with dry land and with rain, the water used was 845 mL and the amount of remaining water was 1.16 L, which represents 57.8% of water saving in an approximate duration time of 1 min and 3 sec. The time needed for each test versus the amount of water used in each of these are shown Fig. 5.

The automatic irrigation system is effective and reliable, since it allows the irrigation of a crop field to work without the need of a person to intervene. This is a key point because as it happens with traditional irrigation in Peru, the water is mismanaged, not only causing unnecessary losses of this resource but also damaging the soil where the crop will grow. What would have been something additional to this project

is the placement of a temperature sensor like the DHT11, to know what temperature has the crop field and in this way to obtain both values such as temperature and humidity of the environment. Our results also go in agreement with other works of automatic irrigation systems [5], [6], [7], [11], [12], [13].

IV. Conclusions

The results of prototype of the automatic irrigation system were better than expected, since before starting the project an effectiveness in water saving of approximately 50% was sought. With the system, an effectiveness of water saving of 75 and 76.5% could be observed. It was seen that the humidity sensor had a good performance and also the wireless communication via bluetooth was possible between that sensors and the application in the smartphone. Hence, the proposed system is quite suitable for crop fields and in addition, also, of a very low cost.

As a future work, it is being considered the implementation of system with the transmission of the data through the internet, i.e. an IoT system, with the purpose that the user can know the humidity of the crop field, even when the farmer is far away, since the bluetooh module is short range which limits the range of movement of the farmer for monitoring. Moreover, the system is scalable and additional sensors to measure the environmental factors analyzed in this study can be added to the system in order that this system can be used in much larger crop fields

References

[1] Food and Agriculture Organization of the United Nations, Family Farming Knowledge Platform. [Online]. Available: http://www.fao.org/family-farming/detail/en/c/297933/.

[2] Correo, Perú ocupa puesto 37 por mal uso de agua. [Online]. Available: https://diariocorreo.pe/peru/peru-ocupa-puesto-37-por-mal-uso-de-agua-390988/.

[3] Food and Agriculture Organization of the United Nations, Agronoticias: Agriculture News from Latin America and the Caribbean. [Online]. Available: http://www.fao.org/in-action/agronoticias/detail/en/c/494657/.

[4] Agraria, 80% del agua en Perú se destina a la agricultura; urge un cambio cultural para dejar de desperdiciarla. [Online]. Available: https://agraria.pe/noticias/80-del-agua-en-peru-se-destina-a-la-agricultura-urge-13448.

[5] E. Avşar, K. Buluş, M. A. Saridaş and B. Kapur, "Development of a cloud-based automatic irrigation system: A case study on strawberry cultivation," 2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST), Thessaloniki, 2018, pp. 1-4.

[6] R. N. Rao and B. Sridhar, "IoT based smart crop-field monitoring and automation irrigation system," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018, pp. 478-483.

[7] P. Singh and S. Saikia, "Arduino-based smart irrigation using water flow sensor, soil moisture sensor, temperature sensor and ESP8266 WiFi module," 2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Agra, 2016, pp. 1-4.

[8] C. Sotomayor-Beltran and L. Andrade-Arenas, "A spatial assessment on internet access in Peru between 2007 and 2016 and its implications in education and innovation," 2019 IEEE 1st Sustainable Cities Latin America Conference (SCLA), Arequipa, Peru, 2019, pp. 1-4.

[9] MINAGRI. Requerimiento agroclimático del cultivo de palto. [Online]. Available: https://www.minagri.gob.pe/portal/download/pdf/ais-2015/ficha12-palto.pdf.

[10] C. Sotomayor-Beltran, G. W. Zarate Segura and D. Tarazona, "Anemia During Pregnancy in Peru in 2017: A Geographic Information System Study," 2018 IEEE 38th Central America and Panama Convention (CONCAPAN XXXVIII), San Salvador, 2018, pp. 1-5.

[11] N. Agrawal and S. Singhal, "Smart drip irrigation system using raspberry pi and arduino," International Conference on Computing, Communication & Automation, Noida, 2015, pp. 928-932.

[12] K. Taneja and S. Bhatia, "Automatic irrigation system using Arduino UNO," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017, pp. 132-135.

[13] S. Vaishali, S. Suraj, G. Vignesh, S. Dhivya and S. Udhayakumar, "Mobile integrated smart irrigation management and monitoring system using IOT," 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2017, pp. 2164-2167.

Non-invasive Device to Lessen Tremors in the Hands due to Parkinson's Disease

Juan Hinostroza-Quiñones[1]
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Peru

Manuel Vasquez-Cunia[2]
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades
Lima, Peru

Abstract—One of the severe neurological disorders that affects the central nervous system is Parkinson's disease, which causes that patients can not perform routine tasks such as eating and writing. According to statistical data, there are more than 10 million people in the world who suffer from this disease and the Latin American nation of Peru is no stranger to this, since approximately 30 thousand people suffer from it. Until today there is not a cure for this disease; however, there are different chemical, biological and electronic methods that help to improve the quality of life of patients with this disease. This research aims to design a low-cost device that is able to diminish tremors in patients with Parkinson. The non-invasive device presented and developed in this study will work with the help of 5 vibratory motors and a microcontroller. The vibrations generated by the motors in the patient's wrist will distract the brain and as result the tremors of the hand due to Parkinson's disease will be reduced.

Keywords—*Parkinson; non-invasive device; vibrations; Arduino*

I. Introduction

In recent years, one of the silent diseases that has been increasing, and that most of the people do not know about, is Parkinson's disease (PD), which is the second most frequent neurodegenerative condition after Alzheimer. Parkinson's disease is a neurodegenerative disorder that affects approximately more than 10 million people around the world [1]. These cases particularly affect people aged 60 and over, with certain cases involving young people and adults. In Peru, based on the statistical data of the National Institute of Neurological Sciences (INCN), it has been indicated that 30,000 people suffer from this disease and each year approximately 3,000 cases are reported [2].

Parkinson's symptoms include increased quivering movements, tremors, muscle stiffness, slow movement, and motion irregularity. All of these symptoms cause the patient to harm their daily routines, for instance, in holding certain objects that are necessary to lead a comfortable quality of life.

It is well known that each of the motor functions of the human body is produced thanks to the electrical stimuli that the brain emits, and these stimuli are transmitted by nerve cells (neurons) [3]. Many assistive devices are now available for people with PD; most of them are portable and effective in suppressing tremors. Blended portable robots and portable orthopedic exoskeletons are prominent in the recovery of patients with PD [4], [5], [6], [7]. An Imperial University of London research team has designed an exoskeleton that uses DC motors and a mechanical arrangement of springs to keep the hand steady [8]. The orthosis of the semi-active arm tremor is one of the new developments in the area of the tremor suppression [9], [10]. This device uses an intelligent fluid and its properties can be varied according to the intensity of the tremor, thus damping the hand tremor. Listenme is a device that helps PD patients who present difficulty in walking. Through small auditory signals that reach the brain and the subconsciousness, this device returns the patient the capacity to rotate the body without falling, proving that it eliminates the frozen gait [11]. On the other hand, gyro gloves are devices that suppress tremors by using the principle of a gyroscope, that is, a set of rotating discs which serves to counteract the intensity of the hand tremor [12], [13]. The vast majority of portable cataclysmic suppression devices are costly or, at worst, unavailable to patients with PD living in developing countries such as Peru, where technology lags far behind developed countries.

The present research proposes the design of a low-cost device to reduce wrist tremors of PD patients. This portable device will be able to generate vibrations by means of small vibratory motors. In this way, it will be possible to achieve comfort and calm in patients suffering from this disease so that they can carry out their daily routines without the need of others' external help.

II. Implementation of the Prototype

To obtain the expected results, and to reduce the tremors in the hands, the vibratory motors need to be placed near the ulnar and the median nerves, since they are the ones that will control the hand and forearm in greater proportion. Once the device fulfills its function of reducing the tremors, it will automatically shut down as there will be a vibratory sensor next to it that manages the vibration range, so that the patients feel in optimal conditions to perform their daily routines.

A. Design

Once the biological parameters to be analyzed are identified, we carried out the design of the system; the diagram of this is shown in Fig. 1. The component used for this device are relatively inexpensive and can be obtainable easily in the market.

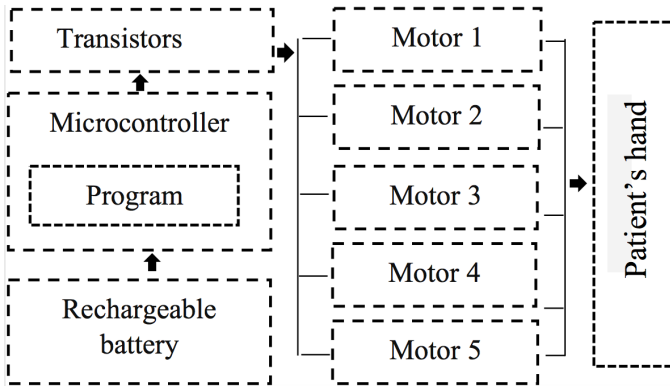These are the following components to be used:

- Vibratory motors (5 units)

Fig. 1. Diagram of the Device to Lessen Tremors due to PD.

Fig. 3. DC Vibration Mini Motor

- Two-pole transistor

- Microcontroller

- Rechargeable Battery (6-9 V)

All these components are integrated by an Arduino Nano ATmega 328p microcontroller, which will process the program previously uploaded, and will send signals to the motors to cause the vibrations needed. Since five vibration motors are being used, and there is a risk that the microcontroller will be damaged by the increase of the current supplied to these, a 2N2222 transistor is put into use.

B. Electronic Components

Arduino Nano: A small card with a microcontroller used to store the developed algorithm to control the system.

2N2222 Transistor (Fig. 2): This component is used to account for the current needed by the vibratory motors.

Vibration mini motor (Fig. 3): It is a DC motor that receives electrical pulses to convert them into vibrations. A total of 5 are used.

The development of the program was mainly based on three random vibration ranges: if the first value is equal to the second and the third ones as well, the five motors are activated; otherwise, we need to perform a second evaluation. If the second random value is equal to the third one, the five motors must be deactivated, otherwise, one last block of code is executed: the engines of equal values of random variables 1 and 2 must be activated. To understand this algorithm better, the flow chart is shown in Fig. 4. The algorithm consists of random ranges that are defined through the X, Y and Z variables.

C. Simulation of the Prototype

In Fig. 5, the simulation of the prototype done in the software Proteus is shown. This simulation is accompanied by an Arduino Nano, which function to send signals to each vibratory motor. For this purpose, the program developed was compiled at the beginning before conducting the corresponding tests. Additionally, it can be noted that in the simulation each of the vibratory motors are connected to the transistor 2N2222. The transistor base is connected to one of the ports assigned in the program so that the Arduino can recognize it to send the

Fig. 2. 2N2222 Transistor.

Fig. 4. Flowchart of the Program.

Fig. 5. Simulation of the Device in Proteus.

signals for the motors to work. Finally, in the simulation can be seen the vibratory motors that are in charge of receiving the signals sent from the Arduino to perform the necessary vibrations.

III. RESULTS AND DISCUSSION

Research shows that Parkinson's disease affects adults aged 35 years and older and it has not a definitive cure until today, reason why the proposed low-cost device in this work.

Fig. 6 shows the implemented prototype, with all components and a battery to have it functioning. The vibration motors are attached to a small elastic band, which is of 5 cm wide and has a length of 30 cm. The device is put on the wrist and attached with the help of a strap secured with velcro.

The arm nerves can feel the vibrations that are generated by the motors, which is why the brain focuses more on random vibration patterns and ignores the hand tremors generated by the Parkinson's disease. The arms nerves are the receptors that work with the brain to regulate voluntary movements. The

Fig. 7. The Prototype in Test Mode.

effect of the vibrations will depend on the seriousness of the tremors. The choice of vibration pattern will depend from one patient to another. A random vibration may be more effective than the autonomous vibration produced by the motors.

Fig. 7 shows the tests of the device performed on a person. A random and rhythmic vibration pattern has been generated and applied to the wrist area.

The device was tested on patients with arm tremors, for which there were evaluations from time to time, as Table I shows. Each test was performed for a total period of 20 minutes for each patient, and every 5 minutes they were asked about their sensations. The sensation were classified as: A) no reduction of tremors, B) minimal reduction and C) significant reduction.

On completion of the tests, positive results were obtained by the device. For patients 1 and 3 after the 20 minutes of evaluation there was a significant reduction in the tremors

Fig. 6. Implemented Prototype.

TABLE I. RESULTS OF THE TEST OF THE PROTOTYPE.

	5 min	10 min	15 min	20 min
Patient 1	A	B	C	C
Patient 2	A	A	B	B
Patient 3	A	A	B	C

sensations. In the case of patient 2 the reduction was minimal; thus, not considering this as a positive result. One reason for the aforementioned can be that not all patients react in the same manner to a similar intensity of vibrations. Overall, the initial prototype test showed satisfying results, and test with a big number of patients with Parkinson's disease would be performed in the coming months to improve this device.

To obtain a high effectiveness from the device at the moment of being exposed to tests in a patient with Parkinson's disease, it is proposed to raise the vibration intensity of the motors. However, when the vibrations are raised, it will generate a greater energy consumption, thus it is suggested the use of lithium battery that presents the technical characteristic of lasting longer and of being rechargeable when needed.

The proposed device is intended to be portable, enduring and affordable for the patient with limited financial resources. On the other hand to further lower costs, it is recommended the use of PICs (Programmable Interface Controllers), as these are cheaper than Arduinos and also would lower the dimensions of the device. Additionally, the device presented in this work can be complemented with the use of a smartphone, devices which are nowadays ubiquitous in Peru [14], in order to monitor better the patients.

IV. Conclusions

The device presented in this work uses light and portable components so that it can be easily handled to abate tremors from the patient with Parkinson's disease. Moreover, the components used are of very low cost allowing the accessibility to many patients with scarce resources, as happens in many developed nations like Peru.

The tests in three patients have shown positives results, however, follow-up studies will be undertanken with a much bigger sample. Also, the identification of the main nerves of the arm that are interconnected with the brain will allow us to determine accurately the position of where the vibratory motors will be attached to the patient's wrist.

References

[1] N. Ball, W.P. Teo, S. Chandra, and J.Chapman, "Parkinson's Disease and the Environment," *Frontiers in Neurology*, vol. 10, 2019.

[2] El Peruano, Parkinson afecta a 30,000 peruanos. [Online]. Available: https://www.elperuano.pe/noticia-parkinson-afecta-a-30000-peruanos-77467.aspx

[3] BBC, Nervous system - Nerve Cells and Nerves. [Online]. Available: https://www.bbc.co.uk/science/humanbody/body/factfiles/nervecellsand–nerves/nerve_cells_and_nerves.shtml

[4] D. Wright, K. Nakamura, T. Maeda, K. Kutsuzawa, K. Miyawaki and K. Nagata, "Research and development of a portable device to quantify muscle tone in patients with Parkinsons disease," 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, 2008, pp. 2825-2827.

[5] R.T. Meyer, and Y. Sergeeva, "Mixed-reality assistive robotic power chair simulator for Parkinson's tremor testing," *Medical Engineering & Physics*, 2020.

[6] Y. Miyake, "Interpersonal Synchronization of Body Motion and the Walk-Mate Walking Support Robot," in IEEE Transactions on Robotics, vol. 25, no. 3, pp. 638-644, June 2009.

[7] R.W. Horst, "A bio-robotic leg orthosis for rehabilitation and mobility enhancement," 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, 2009, pp. 5030-5033.

[8] A. Burton, "Expecting exoskeletons for more than spinal cord injury," *The Lancet Neurology*, vol. 17, pp. 302-303, 2018.

[9] S. Costa, J. Bourget, G. Jablonski, L. Maire, A. Rabelo, M.I. Okereke, L. Chagas, A. Pereira, and A. Andrade, "Ergonomic Evaluation of an Active Wrist Orthosis for the Treatment of Muscular Rigidity in Individuals with Parkinson's Disease," in XXVI Brazilian Congress on Biomedical Engineering, vol. 70, 2019.

[10] S.M. Hashemi, M.F. Golnaraghi, and A.E. Patla,"Tuned vibration absorber for suppression of rest tremor in Parkinson's disease," *Medical and Biological Engineering and Computing*, vol. 42, pp. 61–70, 2004.

[11] El Tiempo, Colombianos crean gafas para el Parkinson. [Online]. Available: https://www.eltiempo.com/archivo/documento/CMS-14582977

[12] H. Dai and L. T. D'Angelo, "Quantitative assessment of tremor during deep brain stimulation using a wearable glove system," 2013 IEEE International Workshop of Internet-of-Things Networking and Control (IoT-NC), New Orleans, LA, 2013, pp. 53-57.

[13] E. Rovini, C. Maremmani and F. Cavallo, "How Wearable Sensors Can Support Parkinson's Disease Diagnosis and Treatment: A Systematic Review," *Frontiers in Neuroscience*, vol. 11, 2017.

[14] C. Sotomayor-Beltran and L. Andrade-Arenas, "A spatial assessment on internet access in Peru between 2007 and 2016 and its implications in education and innovation," 2019 IEEE 1st Sustainable Cities Latin America Conference (SCLA), Arequipa, Peru, 2019, pp. 1-4.

Evaluating the Quality of a Person's Calligraphy using Image Recognition

Avila Cordova[1], Aaron Walter[2], Flores Choque[3], Armando[4], Paucar Nuñez[5], Joseph Clinthon[6]

Escuela Profesional de Ingeniería
de Sistemas
Universidad Nacional de San Agustin
Arequipa, Perú

Abstract—The problem of not developing good handwriting as a child has serious consequences for learning, these range from training human memory to the capacity for innovation. To assess the quality of a person's handwriting, it is necessary to process large amounts of images and with current improvements in machine learning this process is increasingly precise, but the development of these algorithms is complicated. For this reason, this article presents the proposal developed to evaluate the quality of a person's handwriting through image recognition in order to assist in its improvement, but performing image processing in a practical way. This evaluation will be carried out on a group of university students from the Arequipa region, in Peru, using an image processing system that allows character recognition. According to the degree of proximity, the level of handwriting will be determined in a quantified way in percentage degrees. The tests carried out show that the quality of calligraphy in university students in the Arequipa region varies between low and medium.

Keywords—*Calligraphy; image processing; character recognition*

I. Introduction

The poor quality of calligraphy is a latent problem due to its repercussions in the field of academic learning. In recent years with the advancement of the digital age the art of handwriting has been put aside by virtual writing. The use of voice recognition has even begun to replace writing with an electronic keyboard, thus leaving conventional calligraphy to drift. In Peru, schools still provide the art of calligraphy, but this soon disappears in the student due to his constant relationship with computer keyboards.

Despite the benefits obtained with digital writing, such as the hypertextuality where the texts refer to other texts and other information in multimedia format, the interactivity that allows the texts to be pluridirectional, simplified language that are essential elements and suppose an increase in expressive potentialities [1], it is necessary for students to develop good calligraphy as it allows them to better develop their learning skills.

Due to what has been described and the great advancement of image processing techniques, such as character recognition systems, a tool was developed to perform the evaluation of the quality of a person's handwriting through image recognition.

For the developed system, Script typography is considered, this takes as input a template provided to the student, evaluates the characters found in the image and compares them with the characters of an ideal or desired script. Giving as a result the percentage of similarity of the input text with the desired text.

The development of image recognition software for handwritten characters will help improve the quality of people's calligraphy, thus promoting the exploitation of the advantages it offers. The use of image recognition techniques (algorithms), which will be reviewed in the state of the art, will allow students and researchers to know the area itself. Also serving as the basis for future research projects.

The rest of the article is organized as follows: Section II presents the State of the Art of the investigation. Section III presents the Theoretical Framework necessary for the development of the proposal. In Section IV the Work Methodology is presented, followed in Section V of the Tests and Results. Finally, our Conclusions and Recommendations are presented.

II. State of the Art

In the field of research it is not a secret that there are systems that recognize alphanumeric characters. It should be noted that most only have the ability and limitation to recognize the characters written by a computer. For example, this feature can be observed in scanners. Therefore, various technologies have emerged with their respective algorithms that recognize handwritten alphanumeric characters. In the publication [2] explains the operation of the system based on taking a photo with a webcam to a completed data form manually, recognize them and save them in the database, in their corresponding boxes. Likewise, The Project [3] carries out a system based on the functioning of neural networks for the recognition of hand-drawn characters. The project is divided into 2 phases. The first is that of training using the algorithm "resilient backpropagation". For this, it is work with a training data, which is a string of drawings of handmade characters. The next phase is the testing phase. This phase seeks to know how effective the training process of the neural network system has been. For this, the system is tested by entering new information which has never been seen by the system. At the end of this phase, the degree of evaluation of the system is obtained in correctly recognizing each character entered into the system. In the publication [4] also use gradients to extract the properties of characters. A Sobel mask is used to obtain the gradients of each pixel. From these results, gradients are grouped in 12 directions. Since the character is made up of a 32x32 matrix, you would have 1024 address values to be trained by the neural network. 97% yield is obtained. It should be noted that the characters analyzed are three in Hindi, one

in English and one special character. In addition, the work [5] investigates the impact of detecting text using global features for the benefit of Image Based Recovery. in Content and a strategy is proposed. From the experimentation carried out, it is observed that using our strategy, greater precision (15%) is achieved in the recovery of digital images. In the thesis [6] an algorithm is implemented using using image processing techniques in order to recognize numerical characters located in a token device. These processing techniques allowed a total of 80 photographs to reach 99.5% efficiency.

In the article, [10] they use Deep Convolutional Neural Network (DCNN) based classifiers has become a triumph over the state of art machine learning techniques with ccuracy of 96.40% which outperforms some prominent techniques in existence.

In the article, [11] they created a thybrid model is created by evaluating the results of each CNN and revealing the best value. As a result of the experiment carried out on the test data set, it is observed that a performance increase of 1.1% is achieved with the created model.

III. THEORETICAL FRAMEWORK

In this section the main concepts to be used in the work methodology are detailed, these are divided into two parts: Basic Concepts in Image Recognition and Basic Concepts in Image Processing.

A. Basic Concepts in Image Recognition

1) Size or Dimension: It is the space occupied by the letters in the words, or the words in a line. It is based on the areas of the writing, on the size of the ovals of the letters. Classifying them as: small print, medium print or large print and one or more measurements can be taken [7].

2) Proportionality: This element can be considered as a subset of the measurements obtained in the "Dimension" characteristic. However, this property can be based on the proportions of the upper and lower zones of the deed with respect to the central zone [7].

3) Shape: It refers to the type of writing. There are several subdivisions that are grouped according to the features that the letters present, they can be curved, angular, complicated [7][8].

4) Inclination: It is the angle at which writing in a word, line or paragraph tends. Generally 2 types are considered: on the right, on the left.

B. Basic Concepts in Image Processing

1) Capture: This first step is to obtain the image with which we will work, this must be scanned. To obtain an appropriate image, some conditions must be taken into account such as: light, resolution [7].

2) Preprocessing: The received image must be converted to a black and white image through a binarization process, which will reduce it to values of 0 and 1. Subsequently, the colors must be inverted, that is, have a completely black background and our objects to work in white [7].

3) Binarization: Converts the received image into a binary image, thus separating the background from the objects to be analyzed [9] [12] [13].

Global methods try to find a threshold which apply to the entire image among these are threshold methods. The local methods obtain the threshold for each pixel in the image using the values of its neighbors, among which the methods of this category include the method of Niblack and Saovola.

Local methods generally produce a better result by binarizing the image even in situations where the illumination in the document is variable. However the processing and memory limitations make it difficult to implement on mobile devices.

4) Feature Extraction: It allows to know the characteristics such as size, perimeter, area, etc. As well as handwritten features such as segment orientation.

C. Correlation Coefficient

It is a measure of the linear relationship between two quantitative random variables. In a less formal way, we define it as an index that can be used to measure the degree of relationship of two variables as long as they are both quantitative.

D. Training

Preparation to perfect the development of an activity (character recognition). A relative minimum and maximum of inputs to training should be established so that it is as efficient as possible.

IV. WORKING METHODOLOGY

A. Proposal Outline

For the proposal of this work, two entities were identified: "User" who will be the person who uses the application and "System" which is the application that has been implemented. The user starts the system and accesses the "load an image" operation so that the system works with it. The application performs image processing, which consists of capturing, cropping region of interest, segmentation of objects, work material in the clipping to then apply the respective character recognition algorithms (with the appropriate prior training). Subsequently, the calligraphy quality evaluation is carried out for each letter and for all the recognized text. Finally, a plain text file with the detailed result from the system will be generated as output.The outline is shown in Fig. 1.
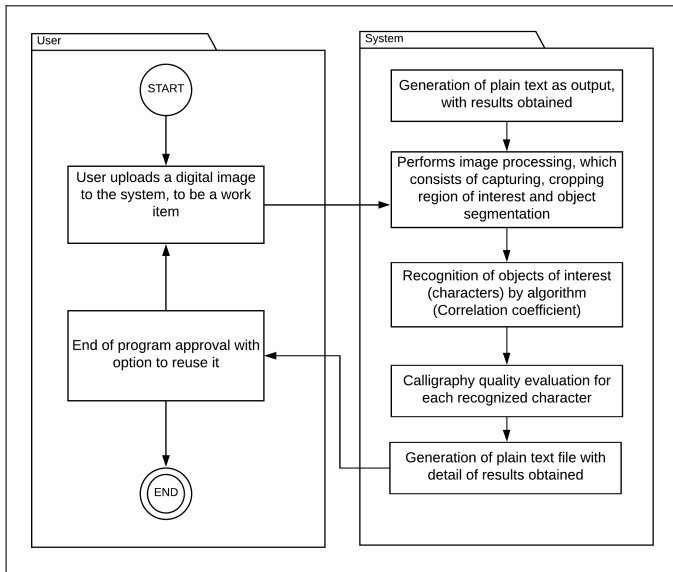
Fig. 1. Proposal Outline

1) Training: This process must be carried out before starting the system, which consists of building a database with all the characters (26-letter lowercase alphabet, the letter ñ is not considered for this process) in images. And also count for each character at least 10 copies to increase the algorithm efficiency by recognizing each character and giving an evaluation of the quality of calligraphy. For the present project, more than 25 copies were taken per letter, obtaining an efficiency greater than 90% (see Fig. 2).

Fig. 2. Character Training

2) Upload digital image to the System: In this process, the user will upload a digital image to the system, which must necessarily contain a handwritten text, which will be the work of the system.

- **Picture Format:** The unique image format (see Fig. 3) is established for the image loading process in the system, which presents the corresponding indications, highlighting the text to write and the font (script). The image that will be uploaded to the system must be formatted with the extension ".jpg".

Fig. 3. Input Image Format

- **Format Content** The text written in the format provided by the user must be as clear as possible, made with a black "fine pen" pen for the purpose of system efficiency.

3) Images Processing: Once the image loaded in the system, we will proceed with the image processing which consists of two parts: a) A capture of the entire image will be made at a size defined by the system in order to normalize the work object (image), b) An automatic cut will be made by the system delimiting only the region that will be the subject of work and c) It will be derived in the segmentation for each letter identified by the system within the delimited region, which also has a size defined by the system.

- **Chars Segmentation:** The system will perform the segmentation of the work objects (characters) based on the discontinuities of the gray level that consists of segmenting the image from the large changes in the gray levels between the pixels.

4) Recognition of Character in Image: In this process, the system will apply the character recognition algorithm (correlation coefficient) to each object identified in the segmentation process, and then present an evaluation of the quality of calligraphy for each recognized letter.

5) Chars Calligraphy Assessment: In this process, the calligraphy quality classification exercise will be carried out in each recognized character, having as reference the script letter of a computer, having intervals from 0% (unrecognized letter relative to the format text) to 100% (total equality). between a recognized letter and a letter generated by a computer).

- **Calligraphy Quality Rating:** Different ranges are established for evaluating the quality of the user's partial (each recognized character) and total (complete phrase) calligraphy, which are the following in Table I:

TABLE I. CALLIGRAPHY QUALITY RATING

QUALIFICATION	RANGE
Very bad calligraphy	< 30%
Bad calligraphy	30% to 40%
Regular Calligraphy	41% to 60%
Good calligraphy	61% to 70%
Very good calligraphy	≥ 71%

The ranges presented have been taken from an experimentation with people in the city of Arequipa, Perú.

6) Plain Text Generation as Output: As a last process, the system will generate a plain text file which will have detailed information on the identified and recognized letters, classifying them partially and totally. Highlighting the weighted quality of the calligraphy evaluated in the image.

V. Tests and Results

A. Training Results

A template (see Fig. 4) was made to write all the letters, objects of interest to the system and thus train it, these were made by anonymous people ready to collaborate with the system. This time more than 25 samples were taken per letter.

Fig. 4. Training Template

B. Test Results with Automatic Format

Once the system was trained, it could be started up. Calligraphy was evaluated anonymously and randomly to a person, in this part the input image format was used (Fig. 3).

A lowercase text string is established, which was random and must also contain all the letters for the user to write and put to evaluation, which is "david exige plazo fijo o embarque truchas a new york", it is shown in Fig. 5.

Fig. 5. Automatic Format Test Image

The system in execution requests the entrance of image to be work matter; Given the input format, the system automatically crops over the area of interest (user writing) and begins to perform letter-by-letter evaluation. Fig. 6 shows the main system interface.

Fig. 6. Main System Interface

When the system runs, it begins to segment letter by letter, showing them in real time, in Fig. 7 we see the results obtained.

Fig. 7. General Self Test Results

In the same way, the quality results are obtained for each letter (see Fig. 8).

```
Detalles:
para letra=  d tiene calidad de:  21.00 %
para letra=  a tiene calidad de:  21.00 %
para letra=  v tiene calidad de:  74.00 %
para letra=  i tiene calidad de:  44.00 %
para letra=  d tiene calidad de:  28.00 %
para letra=  e tiene calidad de:  67.00 %
para letra=  x tiene calidad de:  69.00 %
para letra=  i tiene calidad de:  41.00 %
para letra=  g tiene calidad de:  43.00 %
para letra=  e tiene calidad de:  65.00 %
para letra=  p tiene calidad de:  45.00 %
para letra=  l tiene calidad de:  24.00 %
para letra=  a tiene calidad de:  26.00 %
para letra=  z tiene calidad de:  36.00 %
para letra=  o tiene calidad de:  46.00 %
para letra=  f tiene calidad de:  41.00 %
para letra=  i tiene calidad de:  45.00 %
para letra=  j tiene calidad de:  44.00 %
para letra=  o tiene calidad de:  43.00 %
para letra=  o tiene calidad de:  60.00 %
para letra=  e tiene calidad de:  37.00 %
para letra=  m tiene calidad de:  31.00 %
para letra=  b tiene calidad de:  23.00 %
para letra=  a tiene calidad de:  21.00 %
para letra=  r tiene calidad de:  21.00 %
para letra=  q tiene calidad de:  37.00 %
para letra=  u tiene calidad de:  29.00 %
para letra=  e tiene calidad de:  59.00 %
para letra=  h tiene calidad de:  54.00 %
para letra=  o tiene calidad de:  43.00 %
para letra=  y tiene calidad de:  56.00 %
```

Fig. 8. Detailed Automatic Test Results

The most useful information is presented at the end of the details, which presents the fashion obtained from all the evaluated letters and by which the final result of the calligraphy quality will be established. In the present example (see Table II), the user got a Mode of 30.67%, that is, a bad calligraphy.

TABLE II. RESULTS OF FRECUENCY TABLE

N	Lower Limit	Upper Limit	Frequency
1	[14	24 >	8
2	[24	34 >	12
3	[34	44 >	10
4	[44	54 >	6
5	[54	64 >	6
6	[64	74 >	3

C. Results of Manual Tests

For this test (see Fig. 9), a manual cut was made on the image format delimiting the area of interest (user writing),

this input being for the system and at the same time it performs the evaluation letter by letter. In these results, the same image used in the previous section is taken as input (Results of automatic tests).

```
texto real=  davidexigeplazofijooembarquehoytruchasanewyork
texto reco=  davidexigeplazofijooembarquehoytruchasanewyork
numero de aciertos =  46  de  46
Ud obtiene un total de calidad al= 48.72 % de un total de  46  letras
Ud obtiene un total de calidad al= 48.72 % de un total de  46  aciertos
```

Fig. 9. General Manual Test Results

In the same way, the quality results are obtained for each letter (see Fig. 10).

```
Detalles:
para letra=  d tiene calidad de:  31.00 %
para letra=  a tiene calidad de:  31.00 %
para letra=  v tiene calidad de:  84.00 %
para letra=  i tiene calidad de:  54.00 %
para letra=  d tiene calidad de:  38.00 %
para letra=  e tiene calidad de:  77.00 %
para letra=  x tiene calidad de:  79.00 %
para letra=  i tiene calidad de:  51.00 %
para letra=  g tiene calidad de:  53.00 %
para letra=  e tiene calidad de:  75.00 %
para letra=  p tiene calidad de:  55.00 %
para letra=  l tiene calidad de:  34.00 %
para letra=  a tiene calidad de:  36.00 %
para letra=  z tiene calidad de:  46.00 %
para letra=  o tiene calidad de:  56.00 %
para letra=  f tiene calidad de:  51.00 %
para letra=  i tiene calidad de:  55.00 %
para letra=  j tiene calidad de:  54.00 %
para letra=  o tiene calidad de:  53.00 %
para letra=  o tiene calidad de:  70.00 %
para letra=  e tiene calidad de:  47.00 %
para letra=  m tiene calidad de:  41.00 %
para letra=  b tiene calidad de:  33.00 %
para letra=  a tiene calidad de:  31.00 %
para letra=  r tiene calidad de:  31.00 %
para letra=  q tiene calidad de:  47.00 %
para letra=  u tiene calidad de:  39.00 %
para letra=  e tiene calidad de:  69.00 %
```

Fig. 10. Detailed Manual Test Results

In the same way, the most useful information is presented at the end of the details, which presents the fashion obtained. In the present example, the user had a fashion of 40.67%, that is, a regular calligraphy.

System tests have been carried out on a certain number of people, with the following table describing the results obtained,

this in order to establish ranges for the calligraphy quality classification.

VI. Conclusions

An image recognition system that can assess the quality of a person's calligraphy has been proposed and presented. The system establishes the calligraphy quality classification arbitrarily with the aim of monitoring it and encouraging the improvement of calligraphy by end users.

It has been detected in the training part that, when the system has too many input samples, an "overtraining" originates, which leads the system to lower its performance considerably.

Based on the tests carried out, it is collected that, in an automatic test, lower results are obtained compared to an automatic test, this due to the automated trimming of the area of interest in the system, seen in the results section.

Today, in the field of character recognition in images we are very close to 100% efficiency

In summary, a working methodology has been developed to evaluate the quality of a person's calligraphy with good results, based on real the experimentation.

VII. Recommendations

Due to the lower performance in an automatic test compared to a manual test, it is recommended to implement an algorithm that does the crop in an ideal way in order not to reduce the image quality and thus obtain the expected results.

It is recommended to carry out a deep monitoring in the training part in order to establish an optimal range of inputs for it.

If you want to evaluate many characteristics of a handwritten text, it is recommended to use maching learning or other algorithms, which are more difficult to elaborate but are more efficient.

Based on the results obtained in the tests carried out in this project, it is proposed to carry out as future work a new

methodology for the improvement of calligraphy of students in the Arequipa region.

The developed proposal can be taken as the basis for future projects in which the typography options in the input text are increased.

References

[1] R. Cremades, E. Maqueda-Cuenca and J. L. Onieva, *Posibilidades didácticas de la escritura digital ubicua en la aplicación WhatsApp Messenger*, 2016.

[2] C. Espejo, *Sistema de reconocimiento de caracteres numéricos para actualización de base de datos*, 2017.

[3] C. Hernández, and S. Nahín, *Implementación de un sistema de información para el reconocimiento de caracteres basado en la red neuronal Perceptron*, 2015.

[4] D. Singh, S. Singh and M. Dutta, *Hand written character recognition using twelve directional feature input and neural network*, International journal of computer applications, 1(3) 94-98 2017.

[5] M. Mejía-Lavalle, M. Lux, C. Pérez and A. Martínez, *Detección de texto en imágenes digitales como estrategia para mejorar la recuperación de imágenes por contenido*, Research in Computing Science, 114, 151-160. 2016.

[6] M. Saldaña *Reconocimiento de Caracteres Numéricos en dispositivo token de entidad bancaria, utilizando procesamiento de Imágenes* 2019.

[7] I. Arévalo, *El futuro de la enseñanza de la escritura manual en la era tecnológica.* 2018.

[8] O. Hernandez and M. Rodríguez, *Procesamiento Digital de Texto Manuscrito para la Generación de CORPUS de Reconocimiento Automátio de Escritura.* Jóvenes en la ciencia, 2015.

[9] M. Forero, M. Merchan, C. Ruiz,*Análisis de técnicas de binarizaión basadas en histogramas 2D* 2016.

[10] C. Saha, R. Faisal and M. Rahman *Bangla Handwritten Basic Character Recognition Using Deep Convolutional Neural Network*, Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Spokane, WA, USA, 2019, pp. 190-195. 2019.

[11] F. Can and A. Yilmaz *Hybrid Handwriting Character Recognition with Transfer Deep Learning*,27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 2019, pp. 1-4 2019.

[12] V. Romera, *Nerea, et al. Procesamiento de texto manuscrito I Segmentación a nivel de palabras, indexación y clustering.* 2017.

[13] Sabeenian R S, Paramasivam M E and Dinesh P M, *Appraisal of localized binarization methods on Tamil palm-leaf manuscripts*, 2016.

Nitrogen Fertilizer Recommendation for Paddies through Automating the Leaf Color Chart (LCC)

Torikul Islam[1], Rafsan Uddin Beg Rizan[2], Yeasir Arefin Tusher[3], Md Shafiuzzaman[4],
Md. Alam Hossain[5] and Syed Galib[6]*

Department of Computer Science and Engineering, Jashore University of Science and Technology, Jashore, Bangladesh

Abstract—Nitrogen fertilizer is inevitable for rice production to ensure that the crop's nitrogen need is adequately supplied, during the growing season. International Rice Research Institute (IRRI) has proposed Leaf Color Chart (LCC) to detect the exact nitrogen need of paddy. Farmers generally monitor the plant's growth (which is also an indicator of the nitrogen concentration of leaves) by comparing the leaf color with the corresponding color of the LCC. Currently, in most cases, LCC is used manually to determine the fertilizer need and thus, there is a chance of either overestimating or underestimating the amount of fertilizer. To avoid this problem, a smart fertilizer recommendation system is proposed in this paper. The proposed method is able to automate the manual acquisition and interpretation of leaf color for classification through LCC. The experimentation considers a sample of 6000 Aman paddy leaf images. The data acquisition process was performed according to IRRI's guidance of taking the paddy leaf images within the body shade by our developed application. The data/images have already been made public in Kaggle - a well-known dataset website. The semantic segmentation of the dataset was performed by a powerful Convolutional Neural Network (CNN) backbone architecture - DeepLabV3+. Color classification into 4 categories of the LCC was performed by CNN architecture which consists of seven layers. Information gain based evaluation was performed in the Decision Tree (DT) approach to select features and with the selected features DT classified images into 4 categories. Color classification by our two proposed methods achieved 94.22% accuracy in CNN Model and 91.22% accuracy in the DT classifier.

Keywords—*Leaf Color Chart (LCC); Convolutional Neural Network (CNN); fertilizer recommendation system; color classification; Decision Tree (DT)*

I. INTRODUCTION

Paddy (*Oryza sativa L.*) is one of the principal food crops and also is consumed by one third of the world population residing in developing countries [1]. Nitrogen is a principle nutrient for paddy. The crop yields in the world are improved significantly due to the use of Nitrogen fertilizer [2]. However, farmers around the world are facing a problem to detect the nitrogen need of paddies for proper cultivation. Farmers attempt to estimate the required amount of nitrogen fertilizers through simply looking at the color of the crop's leaves. They generally apply nitrogen fertilizer too much (little Phosphorus (P) and Potassium (K) and other nutrients) that results in high pest and disease incidence [3]. Therefore, overestimation can be resolved by carefully matching leaf color with the LCC [4]. The optimum use of nitrogen fertilizer can be achieved by matching nitrogen supply with crop demand.

International Rice Research Institute (IRRI) has proposed Leaf Color Chart (LCC) to detect the exact nitrogen level of paddy [5]. LCC is used in the agricultural areas for recommending accurate amount of nitrogen fertilizer. In practice, leaf color is compared with its corresponding color in LCC inside body shade with proper lighting conditions. An exact color calibration process is necessary in the digital dimension for interpreting leaf colors. The calibration process evaluates the performance with the operational lighting conditions and determines whether the crops need fertilizers.

The LCC is being made of the plastic body having four green color levels with the standard suggestion of the amount of nitrogen fertilizers for different species such as Aman and Boro. It is arranged with such a shape where the panels are shown horizontally from yellowish green to dark green. IRRI's standard version ensures that the colors of the paddy leaves can be matched with corresponding LCC colors. The standard version having four green color variations as represented from two for yellowish green to five for the dark green is shown in Fig. 1.

Fig. 1. Standard Leaf Color Chart [5]

Currently, the leaf color comparison process with LCC is manual i.e. farmers have to place paddy leaves on the shades of LCC and match it by simply visualizing. Thus, accurate measurement process is difficult for them. The farmers are also required to compare 6 to 10 leaves with the LCC level and find the average values. If the average value of Aman paddy leaf color level is less than or equal to 3.5 LCC level, nitrogen fertilizer is required of 7.5 kg per 0.133 hectares of land [6]. Usually, this process lacks the ease of detecting the accurate amount of fertilizers due to assumption. With the use of LCC for paddy, there is a possibility of saving fertilizers which may cause a positive environmental effect. The estimated annual saving of urea is 261.6 tons for Bangladesh if 50 per cent of the farmers use LCC in the irrigated rice area of 3488 million hectares of land [4].

Similar research has been proposed for nitrogen fertilizers of Soybean crops [7] by using the Fuzzy Logic method. The nitrogen fertilizer demand for Broccoli Plants was estimated

by Simone Graeff et al. [8]. The authors used the L*a*b* color system, also known as CIELAB color space (more on L*a*b* color system can be found in Malacara, 2003 [9]). The parameter, b*, was used to determine the nitrogen fertilizer demand. Mercado-Luna et al. [10] analysed color images to determine the nitrogen need for tomato seedlings. Wang et al. [11] measured the leaf nitrogen concentration as well as rice chlorophyll content where they took pictures under natural lights using a digital still color camera. Need based nitrogen fertilizer management was performed for different maize (Zea mays L.) genotypes by calibrating the leaf color chart [12]. Sing et al. estimated the need based nitrogen management for Rice and Wheat by using Chlorophyll Meter along with the Leaf Color Chart [13]. However, an automated need based nitrogen fertilizer recommendation system for paddies is required based on standard LCC rules.

Classification of paddy leaves into four categories of LCC color is the principal step of this research. The exact color classification method can bring a solution to this problem. Color classification can be performed by Convolutional Neural Network as well as other Machine Learning (ML) algorithms. These ML techniques have been applied to color classification for crops, fruits or even other objects. Son, Kim and Kim [14] compared the performances of a Gaussian mixture model, an Artificial Neural Network (ANN) model, and a Support Vector Machine (SVM) with two color spaces to detect concrete structural component and found that SVM with HSI (Hue-Saturation-Intensity) color model performed best. The SVM and K nearest neighbors (K-NN) were developed for hair color classification [15]. Automated mango ripening stages detection by color classification was performed using DT classifier [16]. Without using CNN based structure, SVM machine learning algorithm was also used for vehicle color recognition [17]. However, Convolutional Neural Network based backbone architecture can recognize vehicle color more efficiently [18]. Those color classification backbone architectures and ML algorithms motivated us to design our model for paddy leaf color classification.

Therefore, we have considered manually comparing LCC as our research problem and wanted to address it by using CNN and DT classifier as our two proposed automated methods. Those two methods can be integrated through a mobile app and thus a smart farming system can be introduced. Hence, by automating the fertilizer recommendation system can help farmers use fertilizers efficiently. This can make a great change over the total economy of a country. Therefore, our proposed solution will economically empower farmers through automating the determination of fertilizers in paddy fields.

The rest of paper is organized as follows. In the Section II, we have discussed the related background of our work. The data collection and validation has been discussed in section III. The overall system model of our proposed scheme, image segmentation and color classification are discussed in Section IV. Experimental analysis is represented in Section V. Its tell us about the fruitfulness of our model. The conclusion is presented in the Section VI. Finally in the Section VII, the acknowledgment is represented.

II. BACKGROUND

Need based nitrogen fertilizers implementation in economical crops like paddy is a concerning issue across the world. The LCC is basically a guideline to supply the necessary amount of nitrogen fertilizer which is optimal for achieving maximum yield. Researchers around the world proposed some systems to recommend nitrogen based fertilizer for paddies as well as other crops. This section reviews some related works on fertilizer recommendation system using image processing and/or LCC.

The suggestion of standard amount of nitrogen fertilizers is essential for rice production. In the critical growth stage, need base nitrogen management is suggested by using the leaf color chart (LCC) [13]. Modeling of demand of on-farm soil nitrogen supply and nitrogen nutrition in rice and wheat is conducted also [19]. Chlorophyll meter, also known as soil plant analysis development (SPAD), is used for determining the chlorophyll concentration on the leaf. SPAD technology is used for real time nitrogen management in rice [20]. In season, need based nitrogen management have demonstrated for four maize genotypes by calibrating leaf color chart [12].

Data acquisition technique from the field is one of the momentous phases of this research. There are some constraints for capturing images which is proposed by IRRI. The farmers need to capture images inside the body shade with proper lighting conditions [5]. Some researchers have used white paper as a background for better segmentation and prepossessing [7], [21]. Accurate segmentation depends on the proper data acquisition process. Some researchers have embedded Otsu's method in mobile application to obtain threshold values where the leaf will be segmented as object pixels from background pixels [7]. The background pixels are the area with white paper and object pixels are the area of the leaf portion. Reducing the limitation of removing background, we have used the DeepLabV3+ segmentation model which is invented by Google. Moreover, it can segment Regions of Interest (ROI) pixels from the leaf whatever the background is [22].

Accurate color level prediction of paddy leaf into 4 categories of LCC can be a fruitful way of this research. Automatic prediction of leaf color level by using digital image processing techniques has been proposed by Singh and Singh [21]. Here, test images are compared with database generated LCC value and then the color level is predicted. For need based nitrogen fertilizer recommendation on Soyabean crops, the fuzzy logic method has been proposed for automatic leaf color level determination [7]. All object pixel values are transferred to Citrus Color Index (CCI) values which is used for the fuzzy logic inference method. Color histogram analysis and pixels Bitwise operations have been proposed for paddy leaf color perception with standard LCC color level [23]. However, the machine learning and the Deep Neural Network based algorithm may bring more accurate results.

Color classification by deep convolutional neural network and machine learning algorithm have been proposed by some researchers for other color classification problems. The CNN based backbone architecture has been proposed for recognition of vehicle color [18]. The support vector machines (SVM) and Kth Nearest Neighbors (KNN) was developed for hair color classification [15] and vehicle color classification [17].

However, from those color classification architectures and ML algorithms, we have designed our own optimal model structure for paddy leaf color classification with the best result.

As it was seen, a number of analyses have been performed for other color classification. However, there are some limitations. For example, the approaches with the application of fuzzy logic suffers from the problem of being computationally intensive i.e. requires much computation power and takes long time. Moreover, in the color histogram analysis and pixels' bitwise operation, the greenness value of input images is compared with the standard greenness value of LCC each time. This creates repetition of comparison which may be eliminated easily through ML algorithms. Therefore, different ML algorithms or color classification neural network methods can learn and solve the limitations more efficiently. Furthermore, the limitation of background removal and segmentation of a leaf can be performed with Google's segmentation model - DeepLabV3+ - which can achieve better accuracy. More on DeepLabV3+ is discussed in Section IV-A.

III. DATA COLLECTION AND VALIDATION

The entire data set is collected from the paddy fields in the Aman season by our own developed application with the tremendous help of Bangladesh Agricultural Research Institute (BARI), Jashore, Bangladesh and the office of the Upazila Agriculture officer near Jashore. We have used our application in Nokia 3 (8MP Camera) and Samsung S8 (12MP Camera) devices for collecting data. In Fig. 2, we have demonstrated the developed application data acquisition process.

Fig. 2. Developed Data Collection Application

The color of the paddy leaf depends on the maturity as well as Nitrogen concentration. However, Nitrogen concentration can determine by analysing only the top most leaves. At first, a total of 560 top most paddy leaf images were taken. The resolution of the images were 500*500 pixels and we used our developed mobile application where each of 2 to 5 LCC level has 158, 162, 124 and 116 number of images respectively. Moreover, the color of the paddy leaf depends on the ambient light. Therefore, those images are captured in the natural daylight condition inside the body shade which condition is proposed by IRRI [4].

Data validation has been carried out by the two principal scientific officers at Bangladesh Agricultural Research Institute (BARI) Jashore. At the time of data collection, we have maintained the LCC data acquisition procedure as well as classified each images comparing to LCC stage. We have

carefully labeled our total data set corresponding to the LCC level with the help of domain experts. We have made our data set public in Kaggle for the academic, educational,and research community [24].

Fig. 3. Four LCC Level of Paddy Leaf

If the color of a paddy leaf is yellowish green, this color corresponds to LCC level 2. On the other hand, if the color of a paddy leaf is dark green, this color corresponds to LCC level 5. The middle 2 levels of the LCC is between the colors yellowish green and dark green and all these four types of paddy leaves are shown in Figure 3. Those four paddy leaf color levels reflect the nitrogen concentration of crops. The leaf color level 2 and 3 represent that the crops need fertilizers. The fertilizer is not needed for leaf color level 4 and level 5 as the nitrogen concentration is maximum in those two color level.

IV. METHODOLOGY

The main aim of this study is to propose a suitable method for automatic categorization of paddy leaves using their digital images. For that at first images of paddy leaves are taken as inputs. After that the DeepLabV3+ semantic segmentation is performed for background removing. The segmented images are used for both the CNN method and the DT method. In the DT method three features are extracted from segmented images. In both methods, the output is the LCC level. The entire flow diagram of our work is depicted in the Fig. 4.

Fig. 4. Proposed Two Methods

A. DeepLabV3+ Segmentation

Object region segmentation from paddy images is implemented by DeepLabV3+ in TensorFlow [22]. This model generates semantic labels to every pixel in the input image and semantic segmentation masks for a specific Region of

Interest (ROI) in paddy leaf. Assigning these semantic labels involves specifying the outline of artifacts. Hence, this imposes much more stringent localization accuracy criteria than other visual entity recognition tasks, such as classifying image levels or recognizing boundary box rates. It explicitly controls the resolution using atrous convolution process. After that feature responses are computed within Deep Convolutional Neural Networks. This model has been driven by Fully Convolutional Network (FCN) frameworks for object detection and semantic segmentation, respectively.

Fig. 5. Semantic Segmentation Process

In this paper, we propose a comparably enabling framework for semantic segmentation by DeepLabV3+ procedure training with our own data set of paddy leafs which is depicted in Fig. 5. As the backbone networks for implementation of this model, we have used MobileNetv2 architecture [25] rather than using other backbone networks such as Xception (server-side deployment) [26]. MobileNetv2 is a fast network structure designed for mobile devices implemented in real field. This model is conceptually intuitive with fast training and inference time for paddy leaf segmentation. Moreover, it offers flexibility and robustness [25].

B. Classification using Decision Tree

As discussed earlier, DT was applied to fruit categorization through color classification and achieved 96% accuracy [16]. Therefore, we apply this classification technique to leaf color classification.

Decision Tree is a supervised machine learning classification method where the data is relentlessly fragmented by a certain parameter. In this approach, a set of training examples are divided into smaller and smaller subsets while at the same time a corresponding decision tree is formed iteratively.

In this classification approach, we have fitted our data set by selecting three features from segmented images. The features are the average values of the three color channels - average red, average green and average blue. The average of each color is calculated by taking the average of the non-zero values of the image pixels. We have augmented initially our 560 paddy leaf images in total 6000 where each LCC level has 1500 images. The augmentation parameters width shift, height shift, shear, zoom and horizontal flip are used which sustain the color pixel information. After that, we have prepared data set based on 6000 images extracting three features. This data set is made public in the Kaggle data set [24] for the academic, educational,and research community.

Classification using a Decision Tree from a data set can be performed by using some properties or attributes of the data. Attributes may be selected based on many criteria and a selection measure is necessary here in our case. An attribute selection measure is a heuristic for selecting the splitting criterion. We have used Information Gain or Entropy as our

attribute selection measure criterion which is also known as ID3 approaches for learning process. Entropy is a degree of randomness of elements or in other words it is a measure of impurity. Mathematically, it can be calculated with the help of the probability of the items as,

$$H = -\sum p(x) log p(x) \qquad (1)$$

Where $p(x)$ is the probability of the item, x. It is a negative summation of probability times the log of the probability of the item, x [27]. Percentage split technique was used to specify the training and testing datasets. 70% data of the whole dataset were used as training data. Testing was done on the remaining 30% data. Before going to train-test split function, we have shuffled our entire dataset that has reduced the bias error of accuracy. Classification using Decision Tree still have some problems especially when the sunlight condition is different.

Sunlight illumination problem is one of the concerning issues of this method. Different day light conditions will change the average values of the image pixels. As a result, miss classification rate may occur and farmers would not be able to detect the exact LCC level. However, to address this problem, we have moved to the convolutional neural network (CNN) model for better accuracy. Deep convolutional neural network works better than normal machine learning algorithms for feature extraction as well as classification.

C. Color Classification Using CNN

Leaf color information contributes significantly to need based nitrogen management. Extracting optimal features depend on the convolutional layers in the CNN model. We have demonstrated the CNN can not only learn for classifications based on shape information but also learn in color distribution. We have designed an optimal CNN model for paddy leaf color classification.

Our CNN architecture consists of seven layers where five layers are for convolution and 2 layers are fully connected layers. In our CNN architecture, the first two layers are convolution layers. Normalization and pooling processes are performed after the convolution process. The architecture of our CNN model is depicted in the Fig. 6.
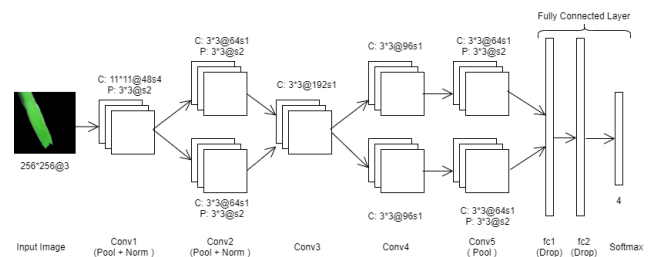

Fig. 6. CNN Architecture of Our System

Features from the images are extracted in the convolution layer. This process is similar to the convolution process in image processing algorithms.If the input image is I_i and the convolution kernel is h, then after the convolution process the output image I_o can be written as,

$$I_o[m,n] = \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} I_i[i,j].h[m,n] \qquad (2)$$

where $I[m, n]$ is the value of the pixel of the image I at the coordinate (m, n). In neural networks, activation factions are used to obtain the output of a node. It determines the resulting value of a node between 0 and 1 or -1 and 1. Moreover, in the convolutional layer, the choice of a specific activation function has vital consequences for the networks. There are different activation functions including $tanh$ and $ReLU$ (Rectified Linear Unit). Activation function $ReLU$ is used for all layers in our CNN networks. The ReLU activation function is used as it does not saturate. When the neuron activates, the gradient is still high (equal to 1). The normalization process is done by equation 3,

$$l_{x,y}^i = K_{x,y}^i / (1 + \alpha/n \sum_{j=i-n/2}^{i+n/2} (K_{x,y}^i)^2)^\beta \qquad (3)$$

Here, $l_{x,y}^i$ is the output of the normalization process and $K_{x,y}^i$ is the result of the layer activation function for convolution at coordinate (x, y). Pooling is performed as the last process in the first two layers. Pooling operation can be of two types: max pooling and mean pooling. The maximum result from the convolutionary process formed by sharp edges will be taken by the max pooling process. On the other hand, the average result of the convolutionary process that sums up the shape in the neighbourhood will be taken by mean pooling. For overlapping pooling, we have used max pooling with size 3x3 and stride 2 in our CNN architecture. The second, fourth and fifth layers are grouped into two independent groups. Convolutional process is also performed in the third and fourth layers without the process of pooling and normalization. The outputs of these two layers are same as input. This is because a 3x3 kernel with pad 1 added has been used for each border. In the fifth layer, only convolution and pooling process is performed without the normalization process.

One long vector is formed by concatenating and flattening the pooling output of the fifth layer from the base networks. This converted long vector is passed to the fully connected layer after the fifth layer. The sixth and seventh layers are two fully connected layers where dropout regularization approach is used to reduce overfitting [28]. Dropout is a strategy that eliminates randomly selected neurons. In the last layer softmax activation function is employed which can be defined in equation 4.

$$p(y^{(i)} = j | x^{(i)}; \theta) = e^{\theta_j^T x^{(i)}} / \sum_{l=1}^k e^{\theta_l^T x^{(i)}} \qquad (4)$$

Where, $x^{(i)}$ is the given input with weight parameter θ and $p(y^{(i)} = j | x^{(i)}; \theta)$ is the probability of $y^{(i)}$ being class j.

Moreover, CNN architecture consists of 7 layers. 11x11@3 kernels for a total of 48 kernels are used in the first layer. The second layer uses 3x3@48 kernels for a total of 128 kernels. On the other hand, the third layer uses 3x3@128 kernels for a total of 192 kernels while 3x3@192 kernels for a total of 192 kernels are used in the fourth layer. Finally, the fifth layer uses 3x3@192 for a total of 128 kernels. In the first, second and fifth layers, the pooling process is employed with the same parameter. The pooling size of 3x3 with 2 pixel stride is used in those layers. The sixth and seventh layers are fully-connected layers. Moreover, in the sixth, seventh and output final layer the

number of neuron is 4096-4096-4 in each layers respectively with dropout regularization method.

We have trained the CNN model by stochastic gradient descent (SGD) as the learning algorithm. We have used 50 images as batch size, three channel image with 256x256@3 resolution, momentum of 0.9 and weight decay of 10^{-6}. The optimization method, SGD, wants to find the minimum or maximum value of some function. SGD will work for all functions with a gradient or first derivative. Moreover, SGD is used by the system for minimizing the error or loss function. The update of the weight parameters based on the function is performed according to equation 5.

$$w_{i+1} = w_i - \alpha \nabla L(z, w_i) \qquad (5)$$

Where the current weight parameters is w_i, the learning rate is α and $\nabla L(z, w_i)$ is the gradient of loss function L with respect to input examples z. To achieve better model convergence, the momentum and the weight decay are employed to the update equation. The final equation of update function in the SGD method is described in equations 6 and 7.

$$\Delta w_{i+1} = \gamma \Delta w_i | (1 - \gamma)(-\alpha \nabla L(z, w_i) \qquad (6)$$

$$w_{i+1} = w_i - \alpha \nabla L(z, w_i) - \alpha \zeta w_i \qquad (7)$$

where γ is momentum variable and ζ is weight decay. Changing momentum and weight decay will speed up the training process.

The prepossessed dataset is used for training in the CNN model. Data augmentation is a powerful technique to reduce overfitting of the model [29], [30]. We have augmented initially our 560 paddy leaf images to 6000 where each LCC level has 1500 paddy leaf images after augmentation These images are also made public in kaggle dataset. Moreover, The percentage split method is used to specify the training and testing sets. Through this method, 70% of the data were used for training and the remaining 30% data were used for testing. Before going to train-test split function, we have shuffled our entire data set that reduces the bias error.

V. EXPERIMENTAL ANALYSIS

Many simulations are conducted to validate our result using two proposed methods. In this section, we describe how we conducted the experiments and demonstrated the simulation outcomes of the proposed two methods. Following metrics are used to measure the performance of our proposed model:

- *Accuracy:* Accuracy is one metric for evaluating classification models. Here, we present accuracy for both CNN method and decision tree method.

- *Loss:* Overall miss classification rate in decision tree method and CNN method.

- *Precision:* Precision can be thought of as a measure of exactness of the model.

- *Sensitivity:* Sensitivity is a measure of completeness of the proposed model.

A. Evaluation of the Decision Tree Approach

The decision tree classifier is significant in machine learning. Evaluation of a model represents how accurately a model classifies the test data set. In the machine learning model, the evaluation belongs to the accuracy score for testing set, miss classification rate, sensitivity and precision scores. We have split our entire 6000 images data set into 70% and 30%. Moreover, before splitting, we shuffled our entire data set which helps us reduce the biasness. We have trained our model by 4200 images data set and tested on the 1800 images data set. Using our standard data set, the decision tree is built with ID3 or information gain algorithm. In this method, the attributes with higher information gain are ranked higher.

In this section, we show, how accurately our model learns and correctly classifies our test data. By using the DT model, we have achieved the test accuracy of 91.22%. The confusion matrix of test data set demonstrates the prediction of each LCC level in Table I.

TABLE I.   CONFUSION MATRIX OF DECISION TREE

Actual Class		Predicted Class			
		Level-2	Level-3	Level-4	Level-5
	Level-2	425 & 26	0	0	
	Level-3	20 & 405	30	0	
	Level-4	0 & 18	405	32	
	Level-5	0 & 0	32	407	

In Table I, we can see the LCC level-2 425 images are correctly classified and only 26 images are misclassifed. Level-2 misclassification rate is 5.76%. In LCC level-3 and level-4, their misclassification rate is 10.99 % in both two levels. Those level misclassification rate is higher than others because of their color level are very close to each other. A little bit daylight change will increase the misclassification rate on these two levels. In LCC level-5, 407 are correctly classified and 32 images are misclassifed. The misclassification rate level-5 is 7.29%. A model's fruitfulness can be found by testing on a totally new data set and accurate class prediction on a new data set. Therefore, We also have checked our model performance by testing 60 totally new images for which each LCC level has 15 paddy images. The DT model achieved 83.33% accuracy in this approach.

There are different evaluation criteria to measure our proposed model's completeness. Here, we have also evaluated our decision tree model by sensitivity and precision score. Precision can be thought of as a measure of exactness of the model. Sensitivity is a measure of completeness of a proposed model. We have calculated the accuracy, sensitivity and precision score of our model by the equations 8, 9 and 10,

$$Accuracy = (TP + TN)/(P + N) \qquad (8)$$

$$Sensitivity = TP/P \qquad (9)$$

$$Precision = TP/(TP + FP) \qquad (10)$$

Where, TP= True Positives, TN= True Negatives, FP= False Positives, P= Totals positive and N= Totals negatives samples. DT model miss classification rate is 8.78%, Precision score is 91.24% and sensitivity score is 91.28%.

B. Evaluation of the CNN Model

In this paper, our aim is to automate the manual LCC system through processing digital images of paddy leaves. To fulfill that expectation, we have proposed a CNN model. In this approach, we have used our 6000 paddy leaf images. We have split the data set into 70% and 30% using train test split function. Therefore, we used 4200 images as our training data and 1800 as our test data set. Then we train our model with this new data set and successfully achieved 94.22% accuracy on test data set.

Model evaluation involves completeness of our proposed model, exactness of our proposed model and accurate classification rate. In this section, we are going to evaluate our CNN model by confusion matrix, precision, and recall score. The confusion matrix is generated from our 1800 test images. From the confusion matrix, we have proved our model miss classification rate is very low in the CNN model. The confusion matrix is shown in Table II,

TABLE II.   CONFUSION MATRIX OF CNN MODEL

Actual Class		Predicted Class			
		Level-2	Level-3	Level-4	Level-5
	Level-2	415	41	0	0
	Level-3	19	433	14	1
	Level-4	0	7	417	11
	Level-5	0	0	11	431

The sensitivity and precision score of our CNN model is 94.27% and 94.31% respectively. We have also checked our model performance by testing 60 new images data set where each LCC level have 15 paddy images. The CNN model have achieved 91.66% accuracy in this approach.

C. Result Analysis

This section presents an analysis of the results for the two proposed methods and the accuracy between two proposed methods. In experimental analysis section, we have shown that the CNN model achieved 94.22% and the DT model achieved 91.22% accuracy by using the same data set. CNN model achieved better accuracy than the DT model. This is due to the fact that sunlight has vital consequences on the paddy leaf images and it changes the pixels' information. As a result, misclassification rate was increased in the DT method. However, in the CNN model our proposed filter and kernel size extract feature in a coherent way that reduces the chances of misclassification. Our proposed CNN model learned in the better way and classified our paddy leaf images with better accuracy.

VI.   CONCLUSION

In this work, we mainly give attention to transform the manual reading of LCC level to an automated method that could be easily deployed in smartphones. Entire work is summarized by three main steps: (1) segmentation, (2) feature extraction, and (3) training and classification. These approaches are performed here using DeepLabV3+ and Convolutional Neural Network. Reducing the sunlight illumination problem while data acquisition can bring out a revolutionary change in this filed. The DeepLabV3+ process can be integrated in Mobile Phone by using the MobileNetV2 backbone network

with TensorFlow. This fruitful solution can also be implemented for other crops like soybean and wheat for nitrogen fertilizer recommendation. Moreover, we have implemented two improved algorithms that accurately classify the Aman paddy leaf according to LCC level with the best 94.22% accuracy in the CNN method and 91.22% in the Decision Tree method.

This automated method of LCC will help the farmer and recommend need based nitrogen fertilizers. The implementation of this system in smart devices will monitor the paddy field automatically that is fruitful for rural area farmers. The farmer will not need manual process to detect the nitrogen level and measuring out the exact amount of fertilizers. Hopefully, Our research will bring an economically fruitful way for Bangladesh as well as ASIA Pacific countries.

VII. ACKNOWLEDGMENT

REFERENCES

[1] D. B. Tari, A. Gazanchian, H. A. Pirdashti, and M. Nasiri, "Flag leaf Morphophysiological Response to Different Agronomical Treatments in a Promising Line of Rice (Oryza sativa L .)," *Natural Resources Research*, vol. 5, no. 3, pp. 403–408, 2009.

[2] S. Peng, R. J. Buresh, J. Huang, X. Zhong, Y. Zou, J. Yang, G. Wang, Y. Liu, R. Hu, Q. Tang, K. Cui, F. Zhang, and A. Dobermann, "Improving nitrogen fertilization in rice by site-specific N management. A review," *Agronomy for Sustainable Development*, vol. 30, no. 3, pp. 649–656, 2010.

[3] S. Yosef Tabar, "Evaluation use leaf color chart in rice for nitrogen management," *J Sci Agric*, vol. 3, pp. 66–9, 2013.

[4] IRRI, "About leaf color chart," https://bit.ly/3eU6dt3, 2019, access date: February, 10, 2020.

[5] I. R. R. Institute, "Leaf Color Chart - IRRI Rice Knowledge Bank," https://bit.ly/2SbIqLs, access date: February, 08, 2020.

[6] B. R. K. Bank, "Leaf Color Chart (LCC) for Fertilizer N Management in Rice," http://knowledgebank-brri.org/how-to-use-lcc.php, access date: February, 09, 2020.

[7] K. R. Prilianti, S. P. Yuwono, M. A. Adhiwibawa, M. N. Prihastyanti, L. Limantara, and T. H. Brotosudarmo, "Automatic leaf color level determination for need based fertilizer using fuzzy logic on mobile application: A model for soybean leaves," *Proceedings - 2014 6th International Conference on Information Technology and Electrical Engineering: Leveraging Research and Technology Through University-Industry Collaboration, ICITEE 2014*, 2015.

[8] S. Graeff, J. Pfenning, W. Claupein, and H. P. Liebig, "Evaluation of image analysis to determine the N-fertilizer demand of broccoli plants (Brassica oleracea convar. botrytis var. Italica)," *Advances in Optical Technologies*, vol. 2008, 2008.

[9] D. Malacara, "Color vision and Colorimetry: Theory and applications," *Color Research & Application*, vol. 28, no. 1, pp. 77–78, feb 2003. [Online]. Available: https://doi.org/10.1002/col.10118

[10] A. Mercado-Luna, E. Rico-García, A. Lara-Herrera, G. Soto-Zarazúa, R. Ocampo-Velázquez, R. Guevara-González, G. Herrera-Ruiz, and I. Torres-Pacheco, "Nitrogen determination on tomato (Lycopersicon esculentum Mill.) seedlings by color image analysis (RGB)," *African Journal of Biotechnology*, vol. 9, no. 33, pp. 5326–5332, 2002. [Online]. Available: https://www.ajol.info/index.php/ajb/article/view/92074/81517

[11] Y. Wang, D. Wang, P. Shi, and K. Omasa, "Estimating rice chlorophyll content and leaf nitrogen concentration with a digital still color camera under natural light," *Plant methods*, vol. 10, no. 1, p. 36, 2014.

[12] H. Thind, A. Kumar, M. Vashistha *et al.*, "Calibrating the leaf colour chart for need based fertilizer nitrogen management in different maize (zea mays l.) genotypes," *Field Crops Research*, vol. 120, no. 2, pp. 276–282, 2011.

[13] B. Singh, Y. Singh, J. K. Ladha, K. F. Bronson, V. Balasubramanian, J. Singh, and C. S. Khind, "Chlorophyll meter–and leaf color chart–based nitrogen management for rice and wheat in northwestern india," *Agronomy Journal*, vol. 94, no. 4, pp. 821–829, 2002.

[14] H. Son, C. Kim, and C. Kim, "Automated color model–based concrete detection in construction-site images by using machine learning algorithms," *Journal of Computing in Civil Engineering*, vol. 26, no. 3, pp. 421–433, 2012.

[15] S. Sarraf *et al.*, "Hair color classification in face recognition using machine learning algorithms," *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, vol. 26, no. 3, pp. 317–334, 2016.

[16] F. S. Mim, S. M. Galib, M. F. Hasan, and S. A. Jerin, "Automatic detection of mango ripening stages–an application of information technology to botany," *Scientia Horticulturae*, vol. 237, pp. 156–163, 2018.

[17] N. Baek, S.-M. Park, K.-J. Kim, and S.-B. Park, "Vehicle color classification based on the support vector machine method," in *International conference on intelligent computing*.   Springer, 2007, pp. 1133–1139.

[18] A. Dehghan, S. Z. Masood, G. Shu, E. Ortiz *et al.*, "View independent vehicle make, model and color recognition using convolutional neural network," *arXiv preprint arXiv:1702.01721*, 2017.

[19] C. Adhikari, K. Bronson, G. Panuallah, A. Regmi, P. Saha, A. Dobermann, D. Olk, P. Hobbs, and E. Pasuquin, "On-farm soil n supply and n nutrition in the rice–wheat system of nepal and bangladesh," *Field Crops Research*, vol. 64, no. 3, pp. 273–286, 1999.

[20] V. Balasubramanian, A. Morales, R. Cruz, T. Thiyagarajan, R. Nagarajan, M. Babu, S. Abdulrachman, L. Hai *et al.*, "Adaptation of the chlorophyll meter (spad) technology for real-time n management in rice: a review." *International Rice Research Notes*, vol. 25, no. 1, pp. 4–8, 2000.

[21] A. Singh and M. L. Singh, "Automated color prediction of paddy crop leaf using image processing," *Proceedings - 2015 IEEE International Conference on Technological Innovations in ICT for Agriculture and Rural Development, TIAR 2015*, no. Tiar, pp. 24–32, 2015.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[23] R. John, B. C. Ii, B. S. Ece, S. R. Soltes, and A. V. Tatel, "Android-based image processing application for rice nitrogen management," no. March, 2012.

[24] Torikul, "Kaggle dataset," https://bit.ly/34rE4EM, 2019.

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[27] Chen Jin, Luo De-lin, and Mu Fen-xiang, "An improved id3 decision tree algorithm," in *2009 4th International Conference on Computer Science Education*, July 2009, pp. 127–130.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.

[30] C. Zhang, P. Zhou, C. Li, and L. Liu, "A convolutional neural network for leaves recognition using data augmentation," in *2015 IEEE International Conference on Computer and Information Technology;* *Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing.* IEEE, 2015, pp. 2143–2150.

Performance Comparison of Natural Language Understanding Engines in the Educational Domain

Víctor Juan Jimenez Flores[1]
Faculty of Engineering
Universidad José Carlos Mariátegui
Moquegua, Perú

Oscar Juan Jimenez Flores[2]
Faculty of Engineering
Universidad Privada de Tacna
Tacna, Perú

Juan Carlos Jimenez Flores[3]
Contracts and Services
Southern Peru Copper Corporation
Tacna, Perú

Juan Ubaldo Jimenez Castilla[4]
Faculty of Engineering
Universidad José Carlos Mariátegui
Moquegua, Perú

Abstract—**Recently, chatbots are having a great importance in different domains and are becoming more and more common in customer service. One possible cause is the wide variety of platforms that offer the natural language understanding as a service, for which no programming skills are required. Then, the problem is related to which platform to use to develop a chatbot in the educational domain. Therefore, the main objective of this paper is to compare the main natural language understanding (NLU) engines and determine which could perform better in the educational domain. In this way, researchers can make more justified decisions about which NLU engine to use to develop an educational chatbot. Besides, in this study, six NLU platforms were compared and performance was measured with the F1 score. Training data and input messages were extracted from Mariateguino Bot, which was the chatbot of the José Carlos Mariátegui University during 2018. The results of this comparison indicates that Watson Assistant has the best performance, with an average F1 score of 0.82, which means that it is able to answer correctly in most cases. Finally, other factors can condition the choice of a natural language understanding engine, so that ultimately the choice is left to the user.**

Keywords—*Chatbot; natural language understanding; NLU; F1 score; performance*

I. INTRODUCTION

Nowadays, the business and researchers are progressively perceive the importance of chatbot systems, because they are integrated into daily life, playing roles as assistants to end users [1]. In the educational domain, Kowalski [2] indicates that chatbots can play an important role, because it represents an interactive mechanism, instead of the traditional e-learning systems, where students can constantly ask questions related to a specific field.

On the other hand, most research does not emphasize the used natural language understanding (NLU) engine, or its choice is not very justified. Therefore, this research compares different NLU engines, like Google Dialogflow, Microsoft LUIS, IBM Watson Assistant, Wit.ai, Amazon LEX and Rasa (an open source chatbot framework) and tries to answer, in terms of performance and educational domain, which one to use.

A chatbot is a computer program which uses machine learning technique voice recognition, and natural language processing (NLP) to conduct a intelligent conversation with a person (e.g. Amazon's Alexa and Google's Assistant) [3]. Moreover, one of the main components of a chatbot is the natural language understanding (NLU), which is the ability of a machine to understand human languages, said otherwise, it is the process of converting natural language text into a form that computers can understand [4].

In addition, multiple researches and comparisons were made regarding the different natural language understanding engines available, as mentioned in Section II; however, their limitation was that they used test data. This research tries to fill that gap.

In this research, training data and input messages were extracted from Mariateguino Bot, which was the chatbot of the José Carlos Mariátegui University during 2018, its main function was to attend to the doubts of the students regarding the necessary requirements to carry out administrative procedures. Moreover, the chatbot was able to answer frequently asked questions, support students regarding the admissions process, and provide class schedules.

Mariateguino Bot was made with Dialogflow, a Google service that runs on Google Cloud Platform. Therefore, other platforms, described in Section III, were evaluated.

To determine the performance of the evaluated services, the F1 score was used. F1 score is a performance measure for compare the quality of predictions between systems [5]. In the same way, F1 score is defined as the harmonic mean of precision and recall [6]. It was chosen because it is one of the most practical ways to numerically calculate the performance of an NLU engine and it is widely used in related researches. In addition, according to [7], using the f1 score, the results can be easily compared with previous works, because it is one of the standard metrics for performance measurement.

Finally, this paper is divided into seven sections. Section II gives a brief overview of related works. Section III defines the NLU engines evaluated during the research. Section IV describes the methodology. Section V and section IV describe

the results and discussions. Finally, Section VII and Section VIII describe the conclusions and future work.

II. Related Works

In recent years, many researches have been carried out regarding chatbots and the impact they have on traditional processes, generally in customer service. Some performance related researches are listed below.

Canonico and De Russis wrote a paper titled "A comparison and Critique of Natural Language Understanding Tools" [8], which compares the main cloud-based platforms, from a descriptive and performance based point of view. Their results showed that Watsson Assistant is the platform who performs best.

On the other hand, Braun, Hernandez, Matthes and Langen wrote a paper titles "Evaluating Natural Language Understanding Services for Conversational Question Answering Systems", which presents a method to evaluate the classification performance of NLU services. Their results indicated that LUIS showed the best scores and RASA could achieve similar results.

Unlike the aforementioned researches, this paper makes a comparison between six natural language understanding engines, with input messages and training data data that belonged to a real chatbot from the José Carlos Mariátegui University. Also, the main language of the chatbot was Spanish.

III. Natural Language Understanding Engines

There are many natural language understanding modules that are available as cloud services and major IT players like Google, Microsoft, IBM, Facebook and Amazon have created tools to develop chatbots [9]. Additionally, Rasa was included because Dialogflow training data can be converted to its format and is an open source alternative compared to the other platforms.

A. Dialogflow

According to Sabharwal and Agrawal [10], Dialogflow is one of the services from Google Compute Platform that makes it easy to integrate cognitive virtual agents to traditional applications; also, it uses natural language understanding and natural language processing capabilities to build complex uses cases.

B. LUIS

The Language Understanding Intelligent Service (LUIS) is a Microsoft's bot engine that runs on Azure Cognitive Services [11].

C. Watsson Assistant

The Watson Assistant service enables learning to respond to the customers in a way that simulates a conversation between humans [12]. In addition, Watsson Assistant is a IBM's bot engine.

D. Wit.ai

Wit.ai is a Facebook's bot engine which allows training bots with sample conversations and have your bots repeatedly learn from interrelating with customers [13].

E. Amazon LEX

Amazon Lex is a Amazon's bot engine for building intelligent assistants or chatbots, which provides many AI capabilities like Automatic Speech Recognition (ASR) and Natural language Understanding (NLU) [14].

F. Rasa

Rasa NLU is an open-source NLP library for intent classification and entity extraction in chatbots [15].

IV. Methods

The method of evaluating the classification performance of NLU engines is based on [16].

A. Materials

The NLU engines evaluated during the research were Dialogflow, Wit.ai, LUIS, Amazon LEX and Rasa. Moreover, 100 messages from the Mariateguino Bot conversation history were randomly selected as input data and they were grouped based on the expected intents, thus obtaining 30 intents. To calculate the performance of each platform, the F1 score metric was used, which includes precision and recall. Some input messages from the students and the expected intents are shown in Table I.

TABLE I. Input message examples

Expected Intent	Input Message
req-tramites	solicito constancia de no adeudo
req-tramites	Record académico oficina
INF-universidad-telefonos	Cual es el numero de telefono de servicios academicos de moquegua
INF-universidad-telefonos	Teléfono de UJCM moquegua
INF-universidad-pago-mensualidades	Fecha de pago de pensiones
INF-universidad-pago-mensualidades	Costo de matrícula
INF-universidad-matriculas	hasta cuando me puedo matricular
INF-universidad-inicio-clases	¿Cuando inician las clases?
FAQ-lugar-clases	¿En dónde serán las clases de Derecho?
FAQ-lugar-clases	Donde sera las clases de ingenieria comercial
FAQ-docentes	Docentes de este ciclo?
FAQ-docentes	Plana docente
FAQ-docentes	Profesores de educación
Default Welcome Intent	Hola
Default Fallback Intent	Administracion VI ciclo
Default Fallback Intent	Quiero estudiar de cero
Default Fallback Intent	Quiero saber si hay carreras a distancia

```
{
  - luis: {
        message: "¿Cuando inician las clases?",
        intent: "Default Fallback Intent",
        confidence: 1
    },
  - watson: {
        message: "¿Cuando inician las clases?",
        intent: "inf-universidad-inicio-clases",
        confidence: 0.8172303199768067
    },
  - witai: {
        message: "¿Cuando inician las clases?",
        intent: "inf-universidad-inicio-clases",
        confidence: 0.6489
    },
  - lex: {
        message: "¿Cuando inician las clases?",
        intent: "faqlugarclases",
        confidence: "?"
    },
  - dialogflow: {
        message: "¿Cuando inician las clases?",
        intent: "INF-universidad-inicio-clases",
        confidence: 0.845428911209106
    },
  - rasa: {
        message: "¿Cuando inician las clases?",
        intent: "inf-universidad-inicio-clases",
        confidence: 0.8966025710105896
    }
}
```

Fig. 1. Node.js Application Output

B. Procedure

As a first step, a conversion of the Dialogflow training data to the rest of the research platforms was carried out, using the QBox.ai service, available in https://qbox.a. Then, one hundred messages were randomly selected from Mariateguino Bot conversation history and they were grouped based on the expected intents, thus obtaining 30 intents.

Afterwards, to start testing and obtain data, a Node.js application was created in order to combine the application programming interface (API) from each NLU engine, in such a way that each input message was only entered once and the desired data was obtained in the format shown in Fig. 1. Also, a threshold of 0.5 was programmed for all platforms, so that if the API of the NLU engine returns a confidence less than 0.5, the Node.js application returns the default fallback intent.

In order to evaluate the results, the predicted intent were identified for each input message. In this way, true positives (TP), false positives (FP) and false negatives (FN) were calculated.

As a final step, the performance of the NLU engines was measure in terms of precision, recall and F1 score, given by the following expressions:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

F1 score is defined as the harmonic mean of precision and recall [6].

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

These measures were applied for single intents, then the average F1 score was calculated. For this research, one NLU engine is better than another if it has a higher average F1 score.

V. RESULTS

The results shown in Fig. 2, Fig. 3, Fig. 4 and Table II are the average precision, recall and F1 score of the 30 intents that were evaluated for each natural language understanding engine.

In terms of precision, as Fig. 2 shows, Dialogflow has the highest value (0.83), while LUIS obtained the lowest value (0.46). This means that the majority of cases that Dialogflow marked as positive, were correct.

In terms of recall, as Fig. 2 shows, Watson Assistant has the highest value (0.89). This means that Watson Assistant correctly identified the majority of positive cases from the total number of cases. On the other hand, LUIS obtained the lowest value (0.34).

Finally, in terms of F1 score, calculated from precision and recall, as Fig. 2 shows, Watson Assistant and Dialogflow have the highest value (0.82), while LUIS obtained the lowest value (0.36). The possible cause of the low performance of Microsoft LUIS is discussed in the section VI.

Overall, as can be seen in Table II, Watson Assistant and Dialogflow performed better, while LUIS obtained the lowest performance.

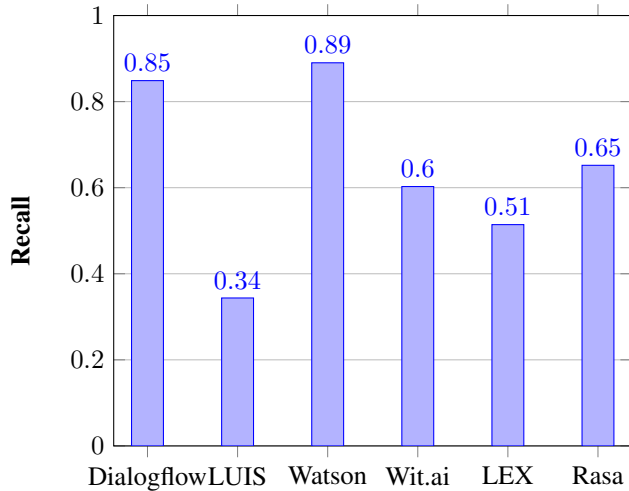Fig. 2. Precision of Natural Language Understanding Engines
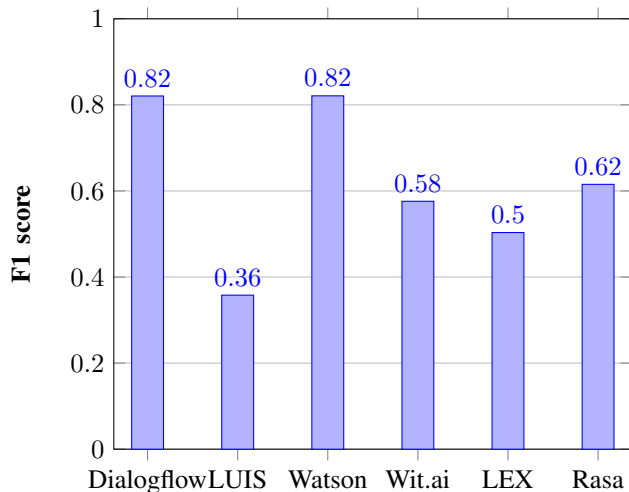
Fig. 3. Recall of Natural Language Understanding Engines

Fig. 4. F1 Score of Natural Language Understanding Engines

TABLE II. F1 SCORES OVERVIEW

NLU Engine	Precision	Recall	F1 score
Dialogflow	0.83	0.85	0.82
LUIS	0.46	0.34	0.35
Watson Assistant	0.81	0.89	0.82
Wit.ai	0.61	0.60	0.58
Amazon LEX	0.55	0.51	0.50
Rasa	0.62	0.65	0.62

VI. DISCUSSION

Despite the fact that Watson Assistant and Dialogflow obtained the same F1 score, Watson Assistant can be considered performed best because the original service with which the chatbot was in production was Dialogflow, so it was constantly improving only on that service.

On the other hand, the low performance of LUIS may be due to the language. Mariateguino Bot was a chatbot made for students in the Spanish language and, despite the fact that LUIS has Spanish in its configuration, it was observed that the intent classification decreases considerably in the presence of input messages that have words with a Spanish accent.

Lastly, the final goal of this research was to compare the main natural language understanding engines and determine which one has the highest performance in the educational domain. Watson assistant was the service with the highest performance; however, for [16], LUIS showed the best results. This difference may be due to the fact that the chatbot domain and language was not the same. Moreover, we agree that Rasa can get better results, after some customization, because, during the present research, its full potential as an open source solution was not exploited. In addition, we agree with [8], which indicates that Watson is the platform that performs best since it can assign the correct intention in most of the cases studied, with a high confidence level.

VII. CONCLUSION

This study presented a performance comparison of Dialogflow, LUIS, Watson Assistant, Wit.ai, Amazon LEX and Rasa services in the educational domain, in order to determine which chatbot solution performs best and provide future researchers with more information on which service to choose. It was concluded that Watson Assistant showed the best performance and its use is recommended for the development of chatbots belonging to the educational domain. However, other factors may affect the choice of a platform that provides the NLU engine, such as the level of usability of the service or pricing plans. Therefore, it will be the company or researcher who decides which service best suits their needs.

On the other hand, the performance obtained by Rasa can be considerably improved with the appropriate settings, keeping in mind that it is an open source chatbot framework with a powerful natural language understanding engine.

VIII. FUTURE WORK

As future work, we plan to evaluate the performance of NLU engines across multiple domains. Similarly, we plan to evaluate the optimal threshold in order to improve performance, since for this research, we only worked with 0.5.

ACKNOWLEDGMENT

REFERENCES

[1] J. J. Bird, A. Ekárt, and D. R. Faria, "Learning from interaction: An intelligent networked-based human-bot and bot-bot chatbot system," in *UK Workshop on Computational Intelligence*. Springer, 2018, pp. 179–190.

[2] S. Kowalski, R. Hoffmann, R. Jain, and M. Mumtaz, "Universities Services in the New Social Ecosystems: Using Conversational Agents to Help Teach Information Security Risk Analysis," in *SOTICS 2011, The First International Conference on Social Eco-Informatics*, 2011, pp. 91–94.

[3] A. Mittal, *Getting Started with Chatbots: Learn and create your own chatbot with deep understanding of Artificial Intelligence and Machine Learning*. Bpb Publications, 2019.

[4] N. Pathak, *Artificial Intelligence for .NET: Speech, Language, and Search: Building Smart Applications with Microsoft Cognitive Services APIs*. Apress, 2017.

[5] Z. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal Thresholding of Classifiers to Maximize F1 Measure," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II*. Springer, 2014, p. 715.

[6] O. Campesato, *Artificial Intelligence, Machine Learning, and Deep Learning*. Mercury Learning & Information, 2020.

[7] G. Arnicans, V. Arnicane, J. Borzovs, and L. Niedrite, *Databases and Information Systems: 12th International Baltic Conference, DB&IS 2016, Riga, Latvia, July 4-6, 2016, Proceedings*, ser. Communications in Computer and Information Science. Springer International Publishing, 2016.

[8] M. Canonico and L. De Russis, "A comparison and critique of natural language understanding tools," *Cloud Computing*, vol. 2018, p. 120, 2018.

[9] P. Hall, V. Venigalla, and S. Janarthanam, *Hands-On Chatbots and Conversational UI Development: Build chatbots and voice user interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*. Packt Publishing, 2017.

[10] B. Galitsky, *Developing enterprise chatbots : learning linguistic structures*. Springer, 2019.

[11] N. Pathak and A. Bhandari, *IoT, AI, and Blockchain for .NET: Building a Next-Generation Application from the Ground Up*. Apress, 2018.

[12] S. Vetter, A. Azraq, S. Chughtai, A. Mashhour, D. V. Nguyen, R. M. Dos Santos, and I. B. M. Redbooks, *Enhancing the IBM Power Systems Platform with IBM Watson Services*. IBM Redbooks, 2018.

[13] J. Seligman, *ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING AND MARKETING MANAGEMENT*, 2018.

[14] S. Tripuraneni and C. Song, *Hands-On Artificial Intelligence on Amazon Web Services: Decrease the time to market for AI and ML applications with the power of AWS*. Packt Publishing, 2019.

[15] S. Raj, *Building Chatbots with Python: Using Natural Language Processing and Machine Learning*. Apress, 2018.

[16] D. Braun, A. Hernandez, F. Matthes, and M. Langen, "Evaluating natural language understanding services for conversational question answering systems," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 174–185.

Date Grading using Machine Learning Techniques on a Novel Dataset

Hafsa Raissouli[1], Abrar Ali Aljabri[2] Sarah Mohammed Aljudaibi[3]
Fazilah Haron[4] and Ghada Alharbi[5]
College of Computer Science and Engineering, Taibah University, Medina, KSA[1,2,3,5]
College of Computer and Cyber Sciences, Prince Muqrin University, Medina, KSA[4]

Abstract—**Dates grading is a crucial stage in the dates' factories. However, it is done manually in most of the Middle Eastern industries. This study, using a novel dataset, identifies the suitable machine learning techniques to grade dates based on the image of the date. The dataset consists of three different types of dates, namely, Ajwah, Mabroom, and Sukkary with each having three different grades. The dates were obtained from Manafez company and graded by their experts. The color, size and texture of the dates are the features that have been considered in this work. To determine the color, we have used color properties in RGB (red, green, and blue) color space. For measuring the size, we applied the best least-square fitting ellipse. To analyze the texture, we used Weber local descriptor to distinguish between texture patterns. In order to identify the suitable grading classifier, we have experimented three approaches, namely, k-nearest neighbor (KNN), support vector machine (SVM) and convolutional neural network (CNN). Our experiments have shown that CNN is the best classifier with an accuracy of 98% for Ajwah, 99% for Mabroom, and 99% for Sukkary. Hence, the CNN classifier has been incorporated in our date grading system.**

Keywords—*Date grading; machine learning; k-nearest neighbor; support vector machine; convolutional neural network*

I. Introduction

The world produces more than four million metric tons of dates annually [1]. Fig. 1 shows the top ten countries in dates production [2]. Each kind of dates (Ajwah, Safawi, Sukkary, etc.) has three different grades. For example, Ajwah has: Ajwah grade 1, Ajwah grade 2, and Ajwah grade 3. The first grade is known with the best quality.

In companies that treat and package dates, sorting and grading is a crucial stage. However, only a limited number of factories in the Middle East adopted automated grading machines to grade dates. The process of grading dates relies on many criteria. The most relevant one is the size of the date. Typically, the first grade comes with the largest size. The existing grading machines grade based on the size criterion. However, grading dates does not base on the size only. The Ministry of Environment Water and Agriculture published in KSA the standards of grading dates. The standards covered eleven kinds of dates (Burji, Khudri, Khalas, Raziz, Sukkary, Shishi, Safawi, Safari, Sagui, Ajwah, and Anbar). The criteria that are considered for grading each type to three grades are: weight, size (height and width), color, shape, humidity content, and sugar level.

According to Manafez international company, the grading machines that base on the size of the date to differentiate between grades have a low grading accuracy. In dates factories,

the grading is manually done by trained experts, which is time consuming. "We are still using manual grading. The main reason is that grading machines are expensive but with a low accuracy" a remark made by the CEO of Manafez International (AlWasel, W. 2018, personal communication, 1 October).

Besides the size of the date, the texture is an important feature. Generally, the first grade of dates, that is the highest quality, is less textured. The texture is an index of the water content of the date. The thirst grade, that is the lowest quality, usually has less water content and hence is drier and more textured. The color is also considered an important criterion in grading dates. Even though the color from grade to grade varies within a small range, considering the color as a feature is helpful.

There are tens of kinds of dates around the world. However, they are similar in shape and within a small range of colors. Therefore, intraclass dates grading presents a challenge. Dates that are of the same type, but different grades are not easy to recognize. Separating different grades manually is time consuming and requires experience and skills. Since this process is needed so frequently in industries, a date grading system will be of great use to reduce time and human intervention.

The purpose of this study is to identify suitable machine learning algorithm to grade dates using a novel dates dataset based on three key features (color, size, and texture). The machine learning algorithms of interest are; the k-nearest neighbor (KNN), support vector machine (SVM), and convolutional neural network (CNN). The rest of the paper is organized as follows: Section II presents a review on the existing work on both classification and grading of dates. Section III demonstrates the development framework. Section IV discusses the experimental settings used. Section V shows the results. A discussion of those results is presented in Section VI, and the conclusion is in Section VII.

II. Literature Review

Researches on dates grading and classification started almost two decades ago. Many studies in the literature have addressed the problem of classification of different types of dates [3] [4] [5] [6] [7] [8]. However, only a handful of researchers focused on dates grading [9] [10]). In the relevant work, intra-class classification is more focussed on types of fruits other than dates. In [11], the authors conducted an extensive review on the quality evaluation of different fruits and vegetables. In [12], the authors have used multilayer perceptron (MLP) and random forest (RF) to grade grapes

while in [13] they have adopted neural networks (NN) to grade apples. Other works include the use of SVM for mango grading [14], NN for tomatoes [15] and SVM for strawberries [16].

The earliest work on dates' research is in 2003 [3]. The authors proposed a system that classifies seven types of dates using multilayer perceptron (MLP) and a statistical method with 100 images for each date type. The most accurate model was the MLP based model. The features used are a combination of physical features (size, shape, texture) and the color feature of the dates. In [4], probabilistic neural networks (PNN) are applied to classify five types of dates. The research used a dataset that consists of 40 images per type. The obtained accuracy was 60% for Bomaan date type, 80% for Khalas, Lolo, and Berhi, and 100% for Fard. In [5], KNN, linear discriminant analysis (LDA) and back propagation neural networks (BPNN) are experimented to classify different types of dates using 140 images per type. The achieved accuracy was 99% using BPNN. The study in [6] used a dataset of 220 images per class. MLP was used with backpropagation (BP) and radial basis function (RBF) to classify dates according to the shape and colour of the date. They obtained the highest accuracy using the RBF, which was 91.1%. The author in [7] proposed a system to classify dates using SVM based on the colour, size, and texture of the dates. They compared two algorithms for texture extraction, namely, local binary patter (LBP) and Weber local descriptor (WLD). The WLD showed slightly better results. This research achieved an accuracy of 98%. In [8], a real time data analytic framework is proposed. It uses 5G technology to instantly analyse dates pictures and classify them using CNN. The input can be in a variety of types: single, rectangular box, round box, piled up, and wrapped. The dataset was extracted from google search engine and consists of 2000 images per class. The achieved accuracy was 99%.

There are only two studies that focused on date grading [9] [10]. The author in [9] used BPNN to grade dates into three grades according to the following features: flabbiness, size, shape, intensity, defects using 400 images per grade. This research achieved an accuracy of 80%. The study in [10] graded the dates into six intra classes (soft small, soft large, semi hard small, semi hard large, hard small, hard large) based on the shape and texture of the dates. A dataset of 960 images was used with SVM, KNN and LDA. The The KNN method provides the best performance with an accuracy of 96.45%. Table I summarizes the related work on both date grading and date classification.

In the light of the literature, two algorithms were used in texture analysis, namely, LBP and WLD. LBP was used in dates grading [10]. In [7], both LBP and WLD were experimented with dates classification and the results showed that WLD performed slightly better than LBP. As for the classifiers, NN was used in [9] [5]. NN performed better than both KNN and LDA in the latter. SVM was used in [10] [7]. SVM showed good results for dates classification [7], but not dates grading [10].

III. The Development Framework

The approach is divided to two phases (see Fig. 2). The first phase starts by uploading images. The grayscale conversion of the image permits it to be thresholded in the next step.

Fig. 1. Top Ten Countries in Dates Production 2019 (from EECO)

TABLE I. Summary of related papers on dates classification and grading

Author and year	Method	Accuracy	Features	Images per type
Hobani, et al. (2003) [3]	• Statistical Method • MLP	• 88.4% to 99.4% • 88.7% to 99.6%	• Moisture • Volume • Weight • Size • Shape • Color	100
Fadel (2007) [4]	• PNN	• 60% to 100%	• Color	40
Haidar, et al. (2012) [5]	• KNN • LDA • NN	• 90% • 96% • 99%	• Color • Size • Shape • Texture	140
Alrajeh and Alzohairy (2012) [6]	• MLP (BP) • MLP (RBF)	• 87.5% • 91.1%	• Shape • Color • Size	220
Muhammad (2015) [7]	• SVM	• 98.1%	• Size • Texture	NA
Hossain, et al. (2018) [8]	• CNN	• 99%	• Size • Texture	2000
Al Ohali (2011) [9]	• BPNN	• 80%	• Flabbiness • Size • Shape • Intensity • Defect	1200
Mohana and Prabhakar (2014) [10]	• KNN • SVM • LDA	• 96.46% • 91.66% • 89.58%	• Size • Shape • Texture	960

Thresholding the image is needed to apply the best least fitting ellipse for size measurements. The color identification and the texture analysis happen on the original images. Feature selection is then performed to reduce the high dimensionality of features by identifying the most relevant ones. Two feature selection methods are experimented: Fisher discriminant ratio (FDR) and sequential feature selection. The obtained feature set serves for KNN and SVM classification. While for CNN, only normalization of the images is done before passing the images to the classifier. In order to implement the actual grading system, the most accurate model out of the three is chosen for phase 2. In phase 2, which represents the grading system, the uploaded image is classified by the trained model that is integrated to the system. Results are saved to the database.
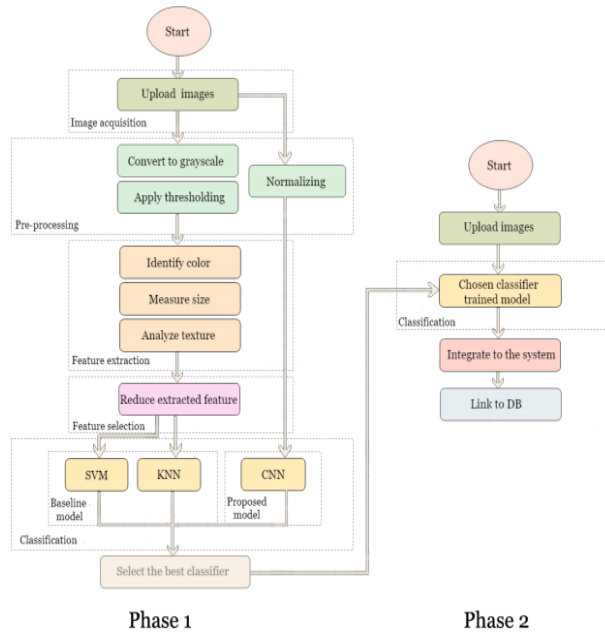
Fig. 2. The Date Grading Development Framework

Fig. 3. The Lighted White-Compartment and the Rotating Dish used to Capture the Dataset

Fig. 4. Samples of the TU-DG Dataset

IV. EXPERIMENTAL SETTINGS

A. Dataset

The dates that were used to construct the dataset are obtained from Manafez International company[1]. Manafez is one of the main exporters of dates in Saudi Arabia. The company provided us with the grades of three types of dates, namely, Ajwah (grade 1, grade 2, and grade 3), Mabroom (grade 1, grade 2, and grade 3), Sukkary (grade 1 and grade 2 only as this type of dates usually comes in two grades). To build the dataset, a video of rotating dates placed on a dish in a lighted compartment was captured. The lighted compartment is illustrated in Fig. 3. A uniform distance of 20 centimeters was fixed with a relatively constant lighting. The distance and lighting were fixed according to the two following experimental observations; changing the distance affects the size of the date in such a way that images taken far from the date appear smaller (see Fig. 6). The size of the date is an important feature in grading. So, changing this parameter may lead to inaccurate results. For the lighting, when enough light is supplied, the details of the date, especially color and texture, are clearly visible. The experiments have shown that low lighting makes the dates appear similar. The wrinkles of the date are not clearly visible. In addition, the color feature is sensitive to the lighting, especially in grading, because the color variance is slight between grades. Fig. 5 illustrates these differences. After capturing the videos, the frames corresponding to the different sides of a date are extracted. The dataset produced, namely Taibah University-Dates Grading dataset (TU-DG), consists of the grades corresponding to three types of dates: Ajwa, Mabroom, and Sukkary. Samples of TU-DG dataset are shown in Fig. 4. Table II shows the number of images in each date type for all the grades in TU-DG dataset. TU-DG dataset is available online on the IEEE dataport.[2]

B. Preprocessing

The preprocessing stage converts the date image to a greyscale image and then threshold it to obtain the region of interest. This serves for applying the best least fitting ellipse to measure the size of the dates. When the thresholded image has some imperfections due to shadows, applying the best least fitting ellipse fails. The algorithm ends up taking ellipses for the date along with other parts of the images that were wrongly thresholded as part of the region of interest. Fig. 7 demonstrates that problem. In order to find a thresholding

---

[1]http://www.manafezinternational.com/

[2]https://ieee-dataport.org/open-access/tu-dg-dataset-date-grading
DOI: 10.21227/qhrr-m850

TABLE II. NUMBER OF IMAGES IN TU-DG DATASET FOR EACH DATE TYPE AND GRADE

Date type	#Grade 1	#Grade 2	#Grade 3	Total
Ajwah	300	345	492	1137
Mabroom	300	500	496	1296
Sukkary	492	458	-	950
Total	1092	1303	988	3383

Fig. 5. Dates Captured in (a) Room Light, and (b) TU-DG Dataset

Fig. 6. A Date Captured with Different Distance, (a)10cm and (b) 30cm

algorithm that works for all the date types, six algorithms were experimented: setting a threshold value manually, weight of intensity difference, active contour, local thresholding, adaptive thresholding, and Otsu thresholding. The number of images that failed in thresholding varies from one method to another. Fig. 9 summarizes the performance of the six algorithms. From the figure, the percentage does not exceed 81% of images that were thresholded correctly for the first five algorithms (setting a threshold value manually, weight of intensity difference, active contour, local thresholding, and adaptive thresholding). Otsu thresholding has succeeded to threshold all the images of all date types. Yet, this type of thresholding resulted in some holes and imperfections in the thresholded image that are recoverable using postprocessing (morphological) techniques. Fig. 8 shows a thresholded image before and after postprocessing. The obtained image after Otsu thresholding shows some small dots (holes) on the white thresholded date. The Matlab built-in functions *imfill* and *bwareaopen* were used to perform the postprocessing. *imfill* fills the small dots while *bwareaopen* ignores any other dots that were not filled by *imfill* function.

C. Features

Most of the related work used the size of the date as in [9] [10]. This is because the size is important to distinguish different types of dates and also different grades. Grade 3 usually appears smaller than grade 1 and 2, while grade 1 and 2 have a slight difference in size. The color was used by many studies in dates classification in [3] [6] and in [10] in dates grading.

The texture was used in [10] for grading and many other studies for classification such as [5], [7], and [8]. The texture is an index of the water contents in the date. Usually grade 1 and 2 have more water contents and hence have less textured skin than grade 3. Some studies used other physical features such as weight, moisture and volume for classification [3]. These features involve manual measurements and are more time consuming. For that reason, the most relevant features

Fig. 7. Best Least Fitting Ellipse Taken for Wrong Areas: (a) Outside and (b) Inside the Date

Fig. 8. Thresholded Image using Otsu Thresholding (a) Before Postprocessing, and (b) After Postprocessing

are extracted in this research that are size, color, and texture. The size was extracted using the best least fitting ellipse. The color was extracted from the RGB color channels of the image. For texture, WLD [17] was used. These extracted features were stored in a matrix with 900 rows, which represents the number of input images, and 873 columns of extracted features. As demonstrated in Table III, 6 features were extracted for size, 771 for color, and 96 for texture. The number of features for each color channel is equal to the values of the channels' histogram bins which is 256. The intensity value for each color channel was also extracted. For the size, the parameters extracted are the length, width, area, perimeter, eccentricity, and equidiameter. WLD descriptor outputs a vector of 32 elements. This descriptor was used with three color channels, RGB, HSV, and YcbCr as in [7].

D. Feature Selection

Feature selection can improve the performance of the model by reducing the dimensionality of the feature set, two feature selection methods were experimented, FDR and sequential.

1) FDR: This method calculates the mean and variance of the variables to increase their separability according to the following equation:

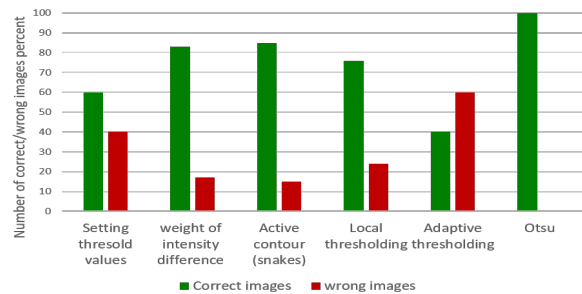$$F = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 - \sigma_2^2} \tag{1}$$

Fig. 9. Percentage of Correct and Wrong Thresholded Images using Different Thresholding Methods

TABLE III. NUMBER OF VALUES FOR EACH EXTRACTED FEATURE

Features	Method used	Description	#Features	Total
Color	Color histogram of RGB channels	Red channel	256	771
		Intensity of red	1	
		Green channel	256	
		Intensity of green	1	
		Blue channel	265	
		Intensity of blue	1	
Size	Best least fitting ellipse	Length	1	6
		Width	1	
		Area	11	
		Perimeter	1	
		Eccentricity	1	
		Equidiameter	1	
Texture	WLD	RGB	32	96
		HSV	32	
		YcbCr	32	
Total number of features			873	

Where μ is the mean and σ^2 is the variance. After performing Fisher discriminant ratio (FDR), the feature selection dimensionality is reduced to two dimensions.

2) Sequential: This selection method selects the best features in the feature set. It takes a number of features to select as an argument and returns a set of best features that is passed to the training process.

E. Classification

1) Baseline Models: A baseline model represents the simplest ways that can be used to obtain results for a given problem. It usually bases on the previous work and serves for comparison with the proposed model. The baseline models in this research are the KNN and SVM. In [7] SVM was used with dates classification and achieved an accuracy of 98%. In [10] KNN and SVM are compared for dates grading. The research shows that KNN gives better results than SVM. KNN and SVM are usually not time consuming in the training phase. They have also relatively simple structure which makes them good to be a baseline model. The baseline models were trained using 300 images per grade for each date type. The loss is estimated using a 10-fold validation for the model. The *FoldLoss* Matlab built-in function calculates the loss for each fold in the validation set and returns the model's loss (or error). This value is between zero and one (for simplicity and ease of comparison with the accuracy value, we state the loss in percentage). The model was tested using 100 images.

- *K-Nearest Neighbor*
  KNN is a classifier that uses similarity to classify new instances. It considers the k closest points in the feature space to assign a class to a given instance. KNN involves several parameter categories such as the value of k (number of neighbors), the distance metric, and the search value. Two search methods are experimented namely, k-d tree and exhaustive search. For the distance metric seven distance metrics were experimented with the exhaustive search method (Minkowski, cityblock, Chebychev, correlation, cosine, Euclidean, and Hamming) and five distance metrics with k-d tree search method that are: Minkowski, Chebychev, cityblock, Euclidean.

- *Support Vector Machine*
  Multi-class SVM was used with RBF kernel. The kernel is required to deal with non-linear problems. This method maps the data points to a higher dimensional space where the classes are linearly separable then converts the dimensions back to the original space. The RBF kernel has two parameters that are C and Gamma. Since those two parameters are not learnable, the Bayesian optimization was used for parameter tuning.

2) Proposed Model: The proposed model is CNN. This classifier requires relatively large dataset. In [18], a dataset of 14828 images was used for a nine class tomato disease classification. This classifier requires more training time compared with the baseline models. CNN has been proven to perform well with many image classification problems. Many studies such as [19] [20] [21] [22] [23] [24] have used CNN to classify fruits. In [25] and [26] CNN was used for vegetables classification. The authors in [18] used CNN for tomato disease classification. The research achieved an accuracy of 99.18%. The study is [27] used CNN in mango classification and achieved an accuracy of 99%. The authors in [28] classified flowers using CNN with an accuracy of 97%. Even though many of the relevant work that applied CNN used big datasets, there are some researches that achieved good results using small datasets. The study in [26] used a dataset of 50 images for cucumber classification task and achieved an accuracy of 96.08%. We have experimented CNN using TU-DG dataset that has 3383 images, and in order to maximize the use of this dataset, a 10-fold cross validation is applied.

3) Evaluation Metrics: Evaluating the result of a classification model using a suitable evaluation metric is important. It shows the performance of a given ML algorithm and eases comparison. Most related works on dates grading used accuracy as an evaluation metric. Since accuracy can be insufficient [29], the loss function is used to support the evaluation of the classification models.

A. Accuracy (ACC)
Accuracy equation is defined by equation 2

$$ACC = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ images} \qquad (2)$$

Accuracy is usually extracted from confusion matrix according to the above equation. A confusion matrix is a table that shows the error of a classification based on the test set. Accuracy works well when the dataset is balanced.

B. Loss Function
The loss function is a method of calculating how well an algorithm models the data. Unlike accuracy that calculates the performance of the model in terms of the test set, the loss function is calculated for the trained model using the validation set. Hence, it is a form of prediction of how well the model will do. In this research, the validation set is divided into ten folds. The loss function calculates the misclassifications of each fold and returns a loss value between 0 and 1, where 0 represents no loss.
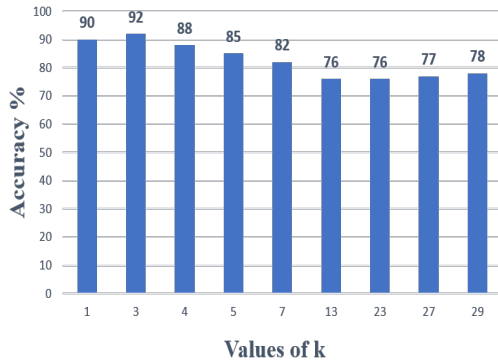
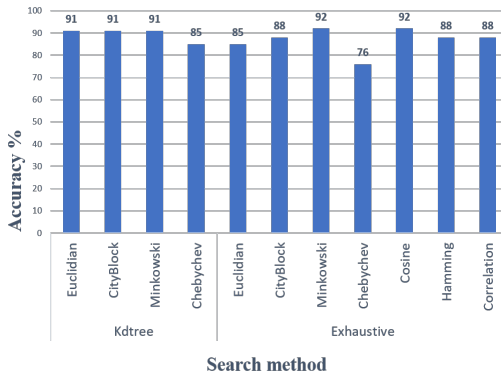Fig. 10. Change of Accuracy with Different Values of k

Fig. 11. K-d Tree and Exhaustive Search Methods with Different Distance Metrics

V. EXPERIMENTS AND RESULTS

A. Baseline Models Results

1) K-Nearest Neighbor: The experiments were conducted first with the row features, that is, without feature selection. Fig. 10 and 11 show the accuracy variation with different values of k and different distance metrics using Ajwa. The results in the figures suggest that the best search method is exhaustive with either cosine or Minkowski distance metric and with k equal to 3. The same parameters worked the best with the other date types (Mabroom and Sukkary). The three types of dates were run in individual experimentations. The accuracy and loss of the KNN classification for the three types of dates is shown in Table VI.

- *FDR Feature Selection*
  When applying the FDR feature selection, the numbers were mapped to complex numbers that cannot be used by the classifier. This problem was solved by cutting off the imaginary part of the numbers and keeping the real part only. Table IV shows the accuracy with and without feature selection. The accuracy has dropped almost half.

- *Sequential Feature Selection*
  When applying the sequential feature selection, the accuracy increases using 20 features and then starts to plateau, indicating minimum benefit in increasing the number of features. Fig. 12 shows varying number of features and the resulting accuracy. Based on the

Fig. 12. The Effect of Different Number of Features on Accuracy

TABLE IV. ACCURACY WITHOUT AND WITH FDR FEATURE SELECTION

Date type	Accuracy before feature selection	Accuracy after feature selection
Ajwa	92.12%	72.11%
Mabroom	90.29%	67.38%
Sukkary	88.29%	52.14%

results, this feature selection method also has a negative effect on accuracy. According to the literature review, not every case needs a feature selection. This step can be omitted, and the extracted features can be passed directly to the classifier. This indicates that this problem performs better with raw features.

2) Support Vector Machine: The multiclass SVM was trained using RFB kernel and with default values of C and Gamma. The Bayesian optimization was then used to tune the parameters. For feature selection, the behavior of the model showed again that it performs better with the raw features. Table VI illustrated the obtained results.

From the results, we observe the following:

TABLE V. THE PERFORMANCE OF THE PROPOSED MODEL (CNN)

Date Type	Validation Accuracy	Test Accuracy
Ajwah	99%	98%
Mabroom	100%	99%
Sukkary	99%	99%

TABLE VI. THE PERFORMANCE OF THE BASELINE (KNN AND SVM) AND PROPOSED MODEL (CNN)

	Date type	# training and validation	# testing images	correct classifications	Acc	Loss
KNN	Ajwa	800	100	92	92%	9.88%
	Mabroom	800	100	90	90%	18%
	Sukkary	800	100	88	88%	17%
SVM	Ajwa	800	100	94	94%	10%
	Mabroom	800	100	97	87%	13%
	Sukkary	800	100	90	90%	11%
CNN	Ajwa	810	90	89	98%	2%
	Mabroom	918	102	101	99%	1%
	Sukkary	864	96	95	99%	1%

- The two baseline models, KNN and SVM, have the same accuracy with Ajwah and Sukkary.

- KNN performs better than SVM for Mabroom.

- Ajwah has the highest performance with both KNN and SVM.

- KNN and SVM results are roughly comparable for all the three date types. The performance of a classification method is highly affected by the quality of dataset supplied and the features extracted. The dataset and the methods used in feature extraction were the same for both KNN and SVM. This might be a reason for the comparability of the results of these two classifiers.

B. Results of Proposed Model

For the three date types, the model was trained with a 10-fold cross validation using six convolutional layers with batch normalization and maxpooling. The initial learning rate is set to 0.01 and stochastic gradient descent (SGD) is used for optimization. Table V shows the validation and test accuracy of the model. The results show that both Mabroom and Sukkary were classified with an accuracy of 99% and Ajwa was classified with an accuracy of 98% which is a significantly higher result than the baseline models. Table VI presents the total images used in training and testing, and also the number of correct classifications and testing accuracy for the baseline.

Fig. 13. The Accuracy of KNN, SVM and CNN with the Three Date Types

and proposed models. Fig. 13 shows the accuracy of the baseline models (KNN and SVM) and the proposed model (CNN).

VI. DISCUSSION

In this research, a novel dataset of three grades of three date types was produced (Ajwah (grade 1, 2, and 3), Mabroom (grade 1, 2, and 3), Sukkary (grade 1, 2, and 3)). The dataset was captured under relatively constant lighting and from a distance of 20 cm. Since the details of the date (size, color, and texture) are crucial in grading and vary between the grades within a small range. Changing the lighting or the distance of capturing would affect the appearance of the date details in the image and hence result in images that are hard to grade even by human experts. For the three features that were considered (size, color, and texture), in order to measure the size using the best least fitting ellipse as in [10] and [7], thresholding the image was necessary to define the region of interest. Among the experimented thresholding algorithms such as weight of

intensity difference, active contour, local thresholding, adaptive thresholding, and Otsu thresholding, Otsu thresholding outperformed the other algorithms and successfully thresholded all the images of the three types of date. However, as in [10], postprocessing was still needed to fill the holes inside the date and to ignore the small dots in the background. Following preprocessing, the best least fitting ellipse was applied on the binary thresholded image returning six parameters (length, width, area, perimeter, eccentricity, equidiameter). The color was extracted from the RGB channels along with the intensity of each channel. The texture was extracted using WLD which was applied on three color models (RGB, HSV, YcbCr) [7]. The extracted features were passed to the baseline models which are KNN and SVM. For KNN, different search methods and distance metrics along with different values of k were experimented. Using Minkowski distance and with k=3, the highest accuracy was achieved and ranged between 88% and 92% depending on the date type. For SVM, the RBF kernel was used, and the Bayesian optimization was applied to tune the parameters. The highest accuracy achieved ranged between 86% and 92% depending on the date type. In [10] both KNN and SVM were experimented on dates grading and the study showed that KNN gave much better results than SVM. These results hold somehow in this research, with SVM slightly less than KNN. For the proposed model, the CNN, the images were normalized to have same size across them before passing to the classifier. This model resulted in an accuracy of 98% for Ajwa and 99% for Mabroom and Sukkary. The accuracy of Mabroom and Sukkary is slightly higher than Ajwa because these two kinds show more variation in colour and size from one grade to another. Unlike Ajwa, that has a dark color and less textured skin (refer to Fig. 4). In general, CNN has performed better than KNN and SVM because this model learns high-level features without the need of feature extraction.

VII. CONCLUSION

This paper presents an image based date grading approach using a novel dataset that has three kinds of dates, namely Ajwah, Mabroom and Sukkary. The dataset consists of 3383 images of the three types of dates with their grades; Ajwa (grade 1, grade 2, and grade 3), Mabroom (grade 1, grade 2, and grade3), and Sukkary (grade 1 and grade 2). The size of the date was measured using the best least fitting ellipse after thresholding the image. Among six experimented thresholding algorithms, Otsu thresholding has successfully thresholded all the images of the dataset. The color was extracted from the RGB color space. The texture of the dates was extracted with WLD using three color models that are RGB, HSV, and YcbCr. Three classification techniques were experimented, KNN and SVM as baseline models and CNN as the proposed model. The baseline models achieved an average accuracy of 90% for KNN and 88% for SVM. The proposed model, that is CNN, outperformed the baseline models and achieved an accuracy of 98% for Ajwah, and 99% for Mabroom and Sukkary.

As a future work, this work can be extended in two directions; to build a more comprehensive date types in the dataset to represent the different varieties of dates, and to develop a real-time system that automates the grading process using the CNN algorithm.

REFERENCES

[1] Johnson, D., Al-Khayri, J. and Jain, S., 2015. Introduction: Date Production Status and Prospects in Africa and the Americas. Date Palm Genetic Resources and Utilization, pp.3-18.

[2] Egyptian Economic and Commercial Office in Brazil. 2019. TOP 10 DATE PRODUCING COUNTRIES IN THE WORLD. [online] Available at: ¡https://ecob.com.br/top-10-date-producing-countries-in-the-world/¿.

[3] A. Hobani, A. Thottam, K. Ahmed, Development of a neural network classifier for date fruit varieties using some physical attributes, King Saud University Agricultural Research Center 126 (2003).

[4] M. Fadel, Date fruits classification using probabilistic neural networks, Agricultural Engineering Inter- national: the CIGR Ejournal 7 (3) (2007).

[5] A. Haidar, H. Dong, N. Marvridis, Image-based date fruit classification, IV International Congress on Ultra Modern Telecommunications and Control Systems (2012).

[6] M. Alrajeh, A. Alzohairy, Date fruits classification using mlp and rbf neural networks, International Journal of Computer Applications 41 (10) (2012).

[7] G. Muhammad, Date fruits classification using texture descriptors and shape-size features, Engineering Applications of Artificial Intelligence 37 (2015).

[8] M. Hossain, G. Muhammad, S. Amin, Improving consumer satisfaction in smart cities using edge computing and caching: A case study of date fruits classification, Future Generation Computer Systems 88 (2018).

[9] A. O. Y., Computer vision based date fruit grading system: Design and implementation., Journal of King Saud University - Computer and Information Sciences 23 (1) (2011).

[10] S. Mohana, C. Prabhakar, A novel technique for grading of dates using shape and texture features, Machine Learning and Applications: An International Journal (MLAIJ) 1 (2) (2014).

[11] B. Anuja, B. Atul, Fruits and vegetables quality evaluation using computer vision: A review, Journal of King Saud University - Computer and Information Sciences (2018).

[12] F. e. a. Vazques, Mango classification using convolutional neural networks, Lecture Notes in Computer Science (2009).

[13] X. Li, W. Zhu, Apple grading method based on features fusion of size, shape and color, Procedia Engineering 15 (2011).

[14] C. Nandi, Computer vision based mango fruit grading system, International conference on Innovative Engineering Technologies (ICIET'2014) 28 (29) (2014).

[15] M. Arakeri, Lakshmana, Computer vision based fruit grading system for quality evaluation of tomato in agriculture industry, Procedia Computer Science 79 (2016) .

[16] O. Mahendra, H. Pardede, R. Sustika, K. Suryo, Comparison of features for strawberry grading classification with novel dataset, International Conference on Computer, Control, Informatics and its Applications (IC3INA). (2018).

[17] C. e. a. Jie, Wld: A robust local image descriptor, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2009).

[18] M. Brahimi, K. Boukhalfa, A. Moussaoui, Deep learning for tomato diseases: Classification and symp- toms visualization, Applied Artificial Intelligence 31 (4) (2017).

[19] S. Jeong, H. Yoe, Fruit classification system using deep learning, Journal of Knowledge Information Technology and Systems 13 (5) (2018).

[20] I. Hussain, Q. He, Z. Chen, Automatic fruit recognition based on dcnn for commercial source trace system, International Journal on Computational Science and Applications 8 (2/3) (2018).

[21] H. Muresan, M. Oltean, Fruit recognition from images using deep learning, Acta Universitatis Sapientiae Informatica 10 (1) (2018).

[22] J. Soo-ho, L. Meonghun, Fruit classification system using deep learning, Journal of Knowledge Information Technology and Systems 13 (5) (2018).

[23] I. e. a. Sa, Deepfruits: A fruit detection system using deep neural networks, Sensors 16 (8) (2019).

[24] K. Sanhdhiya, M. Vidhya, M. Saranya, Smart fruit classification using neural networks, International Journal of Trend in Scientific Research and Development 2 (1) (2017).

[25] O. Patil, Classification of vegetables using tensor ow, International Journal for Research in Applied Science and Engineering Technology 6 (4) (2018).

[26] K. e. a. Lin, Deep learning-based segmentation and quantification of cucumber powdery mildew using convolutional neural network, Frontiers in Plant Science 10 (2019).

[27] R. Sriram, A. Tejas, J. Girija, Mango classification using convolutional neural networks, International Research Journal of Engineering and Technology (IRJET) 5 (11) (2018).

[28] H. Hiary, H. Saadeh, M. Saadeh, M. Yaqub, Flower classification using deep convolutional neural networks, IET Computer Vision 12 (6) (2018).

[29] J. Brownlee, Classification accuracy is not enough: More performance measures you can use, Machine Learning Mastery (2014).

Robust Control and Fuzzy Logic Guidance for an Unmanned Surface Vehicle

Marcelo M. Huayna-Aguilar[1], Juan C. Cutipa-Luque[2], Pablo Raul Yanyachi[3]

Electronic Engineering Professional School[1]

Universidad Nacional de San Agustín de Arequipa, Av. Venezuela s/n, Cercado, Arequipa, Peru

Pedro Paulet's Astronomic and Aerospace Institute[2,3]

Universidad Nacional de San Agustín de Arequipa, Cerro San Francisco s/n, Characato, Arequipa, Peru

Abstract—**This work deals with guidance and control of an unmanned surface vehicle which has the mission to monitor autonomously the water quality condition in Peruvian sea onshore. The vehicle is a catamaran class with two slender bodies propelled by two electric thrusts in differential and common modes in order to maneuver in surge and in yaw directions. A multivariable control approach is proposed in order to control these two variables and a fuzzy logic-based guidance tracks predefined trajectories at the sea surface. The conjunction between robust control and guidance algorithms is validated numerically and the results show good stability and performance despite the presence of disturbance, noise sensors and model uncertainties.**

Keywords—*Robust control; guidance; fuzzy; unmanned surface vehicle*

I. INTRODUCTION

The growth, industrialization and concentration of mass population in metropolitan cities have produced risks to vulnerable areas. The Peruvian big cities are close to the Pacific sea, lagoons and rivers, vulnerable to the human activity or natural disasters. Therefore, water quality monitoring is vital for understanding the dynamics of these water bodies. Peruvian public institutions monitor regularly the ocean and lagoons using manned craft and ships through local instrumentation, remote sensing and sampled data analysis [1]. These activities demand steady and high financial support and the use of unmanned surface vehicles (USVs) is becoming more attractive for research institutions due to low-cost, autonomy, and to reduce human risks.

Nowadays, there are many unmanned surface vehicles operating in the world. In [2], the authors present the DELFIM USV designed to carry out mission tasks of data acquisition of marine environment and to serve as a communication hub between a terrain base and multiple autonomous underwater vehicles. The DELFIM has integrated guidance, navigation and control to carry out straightforward its tasks. In [3], the CaRoLime USV presents integration between electronic hardware in order to execute missions for limnology studies in ocean and rivers. The paper also presents a mathematical model obtained using system identification approach and least square estimation optimization. In 2017, the Shanghai University has developed an USV to carry out validation of algorithms applied to hydrography of coastal areas [4]. There are more information about the development and the application of these vehicles that can be found in [5].

The unmanned vehicles is commonly designed with an architecture composed of embedded electronic hardware and software to make possible its maneuvering autonomously or semi-autonomously. The guidance and control algorithms allow the vehicle to move in maritime environment following predefined trajectories set remotely or through an embedded digital memory. The control system of these vehicles should guarantee better performance and stability despite the presence of disturbances, electrical sensor noise and model uncertainties. There are many robust control strategies applied to these unmanned maritime vehicles. An $\mathcal{H}_\infty$ robust mixed sensitivity approach is applied to an autonomous underwater vehicle in [6], an $\mathcal{H}_2$ robust control approach is applied to an autonomous underwater vehicle (AUV) characterized by output disturbances and time delay in [7], a nonlinear robust control is applied to other unmanned maritime vehicle for hull ship inspection in [8], and another modified robust control algorithm for an USV is proposed in [9].

Relative to the guidance, the work of [10] has been a source of inspiration for nowadays trends. A robust guidance control, based in $\mathcal{H}_\infty$ control and Fuzzy guidance, is proposed in [11] to carry out three-dimensional inspection maneuvers of an AUV. Other authors propose a waypoint guidance using fuzzy logic to generate controller command signals [12]. In [13], the authors proposed an improvement to the classical line-of-sight (LOS) guidance algorithm with and integral action and adapting to different cruise speeds. Path following based in LOS is still an active research area to evaluate the USV performance subject to disturbance, noise sensor, parameter variation, and nonlinear constraints imposed by actuators and other unmodeled dynamics [14], [15], [16].

This paper deals with $\mathcal{H}_\infty$ robust control and fuzzy logic guidance system for the EDSON-J vehicle, an USV of the UNSA (*Universidad Nacional de San Agustín de Arequipa*). A multivariable and centralized controller is synthesized using mixed sensitivity approach and a guidance system based on Mamdani type fuzzy logic rule. The main contribution of this paper is to combine the robustness properties of an advanced controller synthesized according to frequency domain requirement with a conventional guidance algorithm based on LOS. The organization of this paper follows: section I presents the introduction over guidance and control in unmanned maritime vehicles, section II presents the mathematical model representation of the EDSON-J USV, section III presents the robust and multivariable control approach, section IV presents the guidance based on fuzzy and LOS approaches, section V presents the numerical results using a simulator and section

VI provides the conclusions.

II. EDSON-J MODEL

The EDSON-J is a catamaran class vehicle, which is under development at the UNSA since 2019 to carry out inspection and monitoring tasks in the Pacific sea coastal and lagoons of Arequipa (south region of Peru country). The vehicle has a length of 3m and width of 1.6m, draft and free-board of 0.6m.

The unmanned vehicle has been modeled using the rigid body dynamics and hydrodynamics forces interaction between the hull structure and the fluid according to the nomenclature given in [17]. Fig. 1 shows the EDSON-J USV reference frames, an inertial-frame fixed at earth and other body-frame attached at the geometric center of the vehicle. The inertial-frame considers three longitudinal position (x, y, z) and three angular displacements (ϕ, θ, ψ), the body-frame considers three linear velocities (u, v, w) and three angular velocities (p, q, r). Unlike underwater vehicles, it is assumed the USV has motions constrained to the water surface and are neglected heave z, roll ϕ and pitch θ motions. Then, the EDSON-J nonlinear model can be summarized in kinematic and dynamic equations, respectively:

$$\dot{\eta} = J\nu, \tag{1}$$

$$M\dot{\nu} + C(\nu)\nu + D(\nu)\nu = b\tau, \tag{2}$$

where the $\eta = \begin{bmatrix} x & y & \psi \end{bmatrix}^T$ is the earth frame position vector, the $\nu = \begin{bmatrix} u & v & r \end{bmatrix}^T$ is the body frame velocity vector. The control input vector is representing by $\tau = \begin{bmatrix} n_c & n_d \end{bmatrix}^T$ indicating the common mode n_c and differential mode n_d propeller actions. $J(\eta)$ is the coordinate transformation matrix between the earth and body frames, M is the rigid body mass plus the added mass of the vehicle, $C(\nu)$ includes rigid body and Coriolis force terms, $D(\nu)$ is the damping matrix, b is the coefficient related to the control input τ. These matrices are expressed following:

$$J(\eta) = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

$$M = \begin{bmatrix} m - X_{\dot{u}} & 0 & 0 \\ 0 & m - Y_{\dot{v}} & mx_G - Y_{\dot{r}} \\ 0 & mx_G - Y_{\dot{r}} & I_{zz} - N_{\dot{r}} \end{bmatrix}, \tag{4}$$

$$C(\nu) = \begin{bmatrix} 0 & C_{12}(\nu) & C_{13}(\nu) \\ C_{21}(\nu) & 0 & C_{33}(\nu) \\ C_{31}(\nu) & C_{32}(\nu) & 0 \end{bmatrix}, \tag{5}$$

$$D(\nu) = -\begin{bmatrix} D_{11}(\nu) & 0 & 0 \\ 0 & D_{22}(\nu) & 0 \\ 0 & 0 & D_{33}(\nu) \end{bmatrix}, \tag{6}$$

where:

$$
\begin{aligned}
C_{12}(\nu) &= -mr, \\
C_{13}(\nu) &= -mx_G r + Y_{\dot{v}}v + \frac{Y_{\dot{r}} + N_{\dot{v}}}{2}, \\
C_{21}(\nu) &= mr, \\
C_{23}(\nu) &= -X_{\dot{u}}u, \\
C_{31}(\nu) &= mx_G r - Y_{\dot{v}}v - \frac{Y_{\dot{r}} - N_{\dot{v}}}{2}r, \\
C_{32}(\nu) &= X_{\dot{u}}u,
\end{aligned}
\tag{7}
$$

TABLE I. EDSON-J USV NONLINEAR MODEL PARAMETERS.

Parameter	Value	Unit
m	250	kg
x_G	0.08	m
I_z	204.1000	kg/m^2
$X_{\dot{u}}$	-2.4706	kg
X_u	-0.2912	kg/s
$X_{|u|u}$	-27.6262	kg/m
$Y_{\dot{v}}$	-247.0649	kg
Y_v	-123.5324	kg/s
$Y_{\dot{r}}$	-370.5973	kg.m
$Y_{|v|v}$	-38.9275	kg/m
$N_{\dot{r}}$	-748.3102	kg.m^2
$N_{\dot{v}}$	-370.5973	kg.m
N_r	-741.1947	kg.m^2/s
$N_{|r|r}$	-262.7609	kg.m^2

$$
\begin{aligned}
D_{11}(\nu) &= X_u + X_{|u|u}|u|, \\
D_{22}(\nu) &= Y_v + Y_{|v|v}|v|, \\
D_{33}(\nu) &= N_r + N_{|r|r}|r|.
\end{aligned}
\tag{8}
$$

The EDSON-J model parameters can be obtained using the rigid body dynamics and the slender body theory through empirical and semi-empirical relations [18]. The computed parameter values for the EDSON-J nonlinear model are presented in Table I.

III. EDSON-J ROBUST CONTROL

This section presents the multivariable robust control approach based on $\mathcal{H}_\infty$ mixed sensitivity. The nonlinear EDSON-J model expressed in (1) and (2) is linearized using Taylor expansion series around the cruise speed of $u = 2$ m/s. To guarantee the linear controllability and observability system properties, some states are neglected and the reduced system has the state variables $\mathbf{X} = \begin{bmatrix} u & v & \dot{\psi} \end{bmatrix}^T$, control input variable $\mathbf{U} = \begin{bmatrix} n_c & n_d \end{bmatrix}^T$ and output variable $\mathbf{Y} = \begin{bmatrix} u & \dot{\psi} \end{bmatrix}^T$. The linear EDSON-J model is expressed in the state space form as follows:

$$G := \begin{cases} \dot{\mathbf{X}} = \mathbf{AX} + \mathbf{BU} \\ \mathbf{Y} = \mathbf{CX} + \mathbf{DU} \end{cases} \tag{9}$$

where the matrices are as follows:

$$\mathbf{A} = \begin{bmatrix} -0.3294 & 0 & 0 \\ 0 & 0.0800 & 0.0858 \\ 0 & -2.3951 & -1.5949 \end{bmatrix}, \tag{10}$$

$$\mathbf{B} = \begin{bmatrix} 0.2109 & 0 \\ 0 & 0.1145 \\ 0 & -0.8347 \end{bmatrix}, \tag{11}$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{12}$$

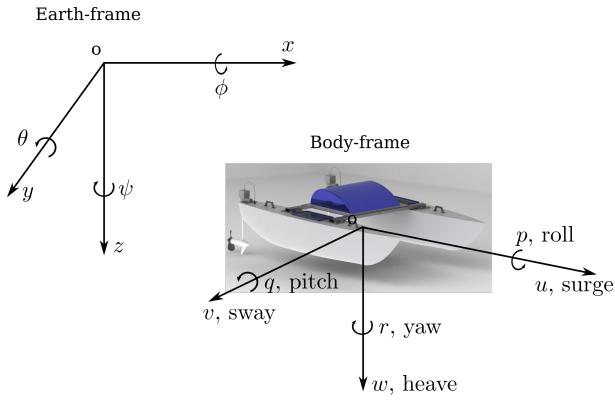$$\mathbf{D} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{13}$$

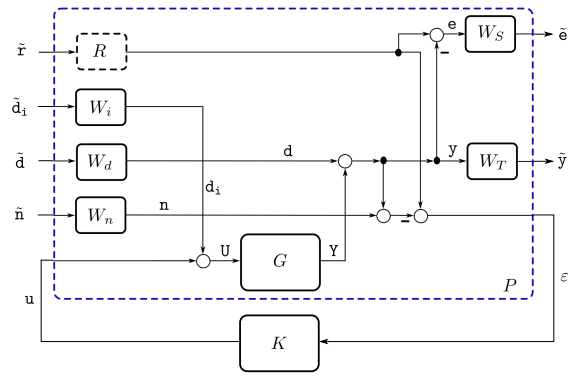Fig. 1. Coordinate Frames for the EDSON-J vehicle.

Fig. 2. Two Ports Configuration to $\mathcal{H}_\infty$ Synthesis [6].

Fig. 2 presents the two ports structure used to synthesize the robust controller through mixed sensitivity approach, where G is the linear model, W_* are the weighting functions, R is the reference pre-filter and K is the synthesized robust controller. The extended plant P and the controller K form the closed loop system $T_{\mathtt{zw}}$; which has the exogenous input $\mathtt{w}$ composed of reference signal $\tilde{\mathtt{r}}$, input disturbance $\tilde{\mathtt{d}}_{\mathtt{i}}$, output disturbance $\tilde{\mathtt{d}}$ and sensor noise $\tilde{\mathtt{n}}$; and has the exogenous output $\mathtt{z}$ composed of tracking error $\tilde{\mathtt{e}}$ and output response $\tilde{\mathtt{y}}$.

A closed loop matrix $T_{\mathtt{zw}}$ is measured using $\mathcal{H}_\infty$ norm to guarantee good performance and stability [19], [6]:

$$\|T_{\mathtt{zw}}\| = \left\| \begin{matrix} W_S S \\ W_T T \end{matrix} \right\|_\infty, \tag{14}$$

where $S = (I + KG)^{-1}$ is the sensitivity matrix, $T = (I - S)$ is the complementary sensitivity matrix, and I is an identity matrix, W_S is the weighting sensitivity matrix and W_T is the weighting of complementary sensitivity matrix. Assuming that the matrices involved satisfy necessary detectability and stabilizability conditions, and based on well-known results, there is an optimal controller $K(s)$ so that a closed loop function between w and z satisfies [19]:

$$\|T_{\mathtt{zw}}\|_\infty = \gamma, \tag{15}$$

where γ is a real number associated to a suboptimal control problem.

The $\mathcal{H}_\infty$ mixed sensitivity approach was formally used in former work to control an autonomous underwater vehicle [6]. In this paper, the authors uses the same approach applied to the EDSON-J USV where the input pre-filter and weighting matrices R, W_i, W_d and W_n are identity matrices of $I_{2\times 2}$. The output weighting sensitivity matrices are diagonal matrices:

$$W_S = \text{diag}\left\{ \frac{0.5s+1}{s+0.0001}, \frac{0.5s+1}{s+0.0001} \right\}, \tag{16}$$

$$W_T = \text{diag}\left\{ \frac{s+1}{0.0001s+5}, \frac{s+1}{0.0001s+5} \right\}. \tag{17}$$

IV. EDSON-J GUIDANCE

The structure of the robust guidance control for EDSON-J vehicle is described in Fig. 3 where the robust control guarantees a good tracking of the cruise speed command u_d and a good tracking of the yaw rate command $\dot{\psi}_d$. This last signal will be generated by the guidance algorithm composed of two subsystems. This section presents the guidance algorithm of the vehicle that integrates the Lookahead-based steering approach and a fuzzy logic controller to generate yaw rate command considering information of LOS.

A. Lookahead-based Steering

The goal of this subsystem is to generate the steering command (yaw angle) using the definiciton of LOS and Lookahead-based sterring algorithm. The geometric representation of line of sight (LOS) is as shown in Fig. 4, where (x,y) is the actual position of the vehicle, (x_k, y_k) and (x_{k+1}, y_{k+1}) are waypoints that define a straight-line path. The path-tangential angle can be expressed as follows [10]:

$$\alpha_k := \text{atan2}(y_{k+1} - y_k, x_{k+1} - x_k), \tag{18}$$

where the atan2 is the four-quadrant version of $\arctan$ constrained to $[-\pi/2, \pi/2]$. The USV coordinates in the path-fixed reference frame (x_e, y_e) can be written as:

$$\begin{bmatrix} x_e \\ y_e \end{bmatrix} = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}^T \begin{bmatrix} x - x(k) \\ y - y(k) \end{bmatrix} \tag{19}$$

where x_e is the along-track distance and y_e is the cross-track error. If the vehicle follows a path, only the cross-track error is relevant because the vehicle converges to the straight-line when $y_e \to 0$. Considering the second row of expression (19), the cross-track error can be re-arranged as:

$$y_e = -(x - x_k)\sin(\alpha_k) + (y - y_k)\cos(\alpha_k), \tag{20}$$

After defined the LOS guidance, a lookahead-based algorithm can be expressed as:

$$\chi_d = \alpha_k + \arctan(\frac{-y_e}{\Delta}) \tag{21}$$

where the second term of expression is the velocity-path relative angle, which ensures the velocity towards at point located ahead a distance $\Delta > 0$. The angle χ_d is transformed to the yaw angle command ψ_d:
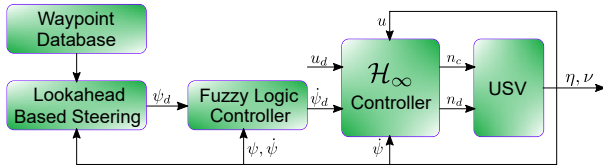
$$\psi_d = \chi_d - \beta \tag{22}$$
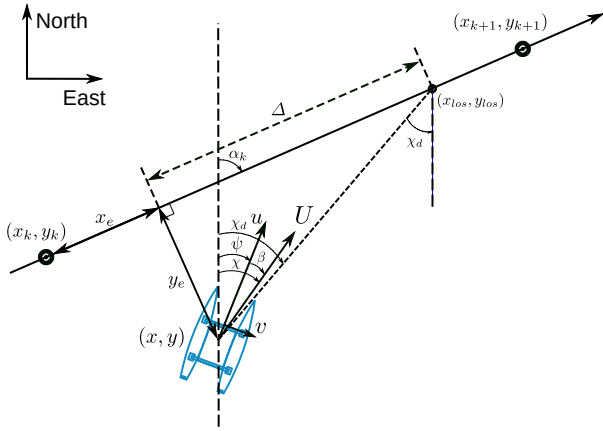
Fig. 3. USV Robust Guidance Control.

Fig. 5. The Yaw Error $\tilde{\psi} = \psi_d - \psi$ Input Membership Functions.

Fig. 4. LOS Guidance Geometry.

Fig. 6. The Yaw Rate $\dot{\psi}$ Input Membership Functions.

where,

$$\beta = \arcsin(\frac{v}{U}) \qquad (23)$$

B. Fuzzy Logic Yaw Rate Controller

The fuzzy logic rule reads the yaw angle command ψ_d from the LOS block and the actual yaw angle ψ from the inertial sensor, and computes the error $\tilde{\psi} = \psi_d - \psi$ as one input. The second input is the actual yaw rate $\dot{\psi}$ from the inertial sensor. This block uses a Mamdani fuzzy inference to generate the yaw rate command $\dot{\psi}_d$ which feeds the robust control system as reference signal (Fig. 3).

Fig. 5 and 6 present the membership functions relative to the two inputs. These inputs use five Gaussian membership function defined as: negative medium NM, negative small NS, zero ZE, positive small PS, and positive medium PM. The one output uses seven triangular membership functions defined as: negative big NB, negative medium NM, negative small NS, zero ZE, positive small PS, positive medium PM, positive big PB (see Fig. 7). It uses the centroid defuzzification which is represented by the surface in Fig. 8.

Fig. 7. Output Yaw Rate Desired.

V. RESULTS

The proposed alternative for good tracking and robust performance is validated numerically using the EDSON-J non-linear model simulator considering three degree of freedom.

A. Robust Controller

The suboptimal gamma value achieved is $\gamma = 1.0067$ and the value of the K controller is reduced to nine states. Fig. 9 presents the surge sensitivity S response with low value

Fig. 8. The Surface of the Fuzzy Inference System.

Fig. 9. Sensitivity S and its Weighting Inverse $1/W_S$ for Surge Velocity u.

Fig. 11. Complementary Sensitivity T and its Weighting Inverse $1/W_T$ for Surge Velocity u.

Fig. 10. Sensitivity S and its Weighting Inverse $1/W_S$ for Yaw Rate $\dot{\psi}$.

Fig. 12. Complementary Sensitivity T and its Weighting Inverse $1/W_T$ for Yaw Rate $\dot{\psi}$.

at low frequencies with slope of 20 dB/dec and it does not cross its weighting given by $1/W_S$. Fig. 10 presents the yaw rate sensitivity S response with low value at low frequencies with slope of 20 dB/dec and it does not cross its weighting given by $1/W_S$. Both results achieve crossover frequencies of 1.12 rad/s and is considered as good tracking and rejection to environmental disturbances, such as waves, currents and wind. Fig. 11 presents the surge complementary sensitivity T response with low value at high frequencies with slope of -20 dB/dec and it does not cross its weighting given by $1/W_T$. Fig. 12 presents the yaw rate complementary sensitivity S response with low value at high frequencies with slope of -20 dB/dec and it does not cross its weighting given by $1/W_T$. Both results achieve crossover frequencies of 6.2 rad/s and is considered as good stability and robustness to model uncertainties and noise rejection.

The closed loop matrix given in (14) achieves the desired shaping. Fig. 13 and 14 shown that the $\mathcal{H}_\infty$ norm is limited to values less than 1.

Fig. 13. Transfer Matrix $W_S S$.

Fig. 14. Transfer Matrix $W_T T$.

Fig. 15. Circular Path Following.

B. *Fuzzy Logic Guidance*

The guidance of EDSON-J is validates with LOS ratio of 5m and with zigzag and circular trajectories. The cruise speed command is keeping to $u_d = 2$ m/s meanwhile the vehicle follows the desired path and Fuzzy logic controller generates the yaw rate command $\dot{\psi}_d$ feeding the Robust controller.

Fig. 15 shows the vehicle follows a desired circular path starting at (0.0) and has a maximum tracking error of 3.52 m, an acceptable value assuming that the LOS ratio is 5 m. The reduction of LOS distance can cause degradation in the Fuzzy logic controller performance and may require more energy consumption of the propellers. Fig. 16 presents the state variables responses u, v and $\dot{\psi}$ during the performance of this path following. The surge velocity u achieves the desired value set by u_d, the sway velocity v has transitory oscillation due to the fuzzy logic rules and steering command given by Lookahead-base steering algorithm. However, at steady state the oscillatory behavior is limited keeping the vehicle velocity at values of 0.08 m/s. The way rate response $\dot{\psi}$ has similar oscillatory transition and in steady state goes to a constant value of 0.013 rad/s. These values are suitable considering the circular path with 150m of ratio.

Finally, a zigzag path of 20m amplitude is considered in Fig. 17. The vehicle follows the desired trajectory with good performance and the maximum tracking error is 0.5 m. Fig. 18 presents the u, v and $\dot{\psi}$ responses during the performance of this path following. The surge velocity u achieves the desired value set by u_d, the sway velocity v has transitory oscillation due to the fuzzy logic rules and steering command given by Lookahead-base steering algorithm. Unlike the last test, this sway velocity oscillate in order to do the turning maneuver in both directions. The yaw rate response $\dot{\psi}$ has also an oscillatory behavior, a maximum value of 0.48 rad/s is achieved in transitory response.

The frequency domain analysis of the robust control approach achieves all the specifications for the USV, a multivariable and coupled system. The Fuzzy guidance, in conjunction with the robust controller, guaranteed good tracking of straight and curved trajectories. The time domain responses show good performance without compromising stability and maintaining

Fig. 16. Nonlinear Model Responses u, v and $\dot{\psi}$ during circular path following.

the effort control level between its physical limitations. These initial results validate the proposed guidance and control schema to be implemented in the EDSON-J USV, which has the main mission to carry instrumentation and measure periodically water quality of the Peruvian sea.

VI. CONCLUSION

A robust guidance control is presented for the EDSON-J, an unmanned surface vehicle under development at the
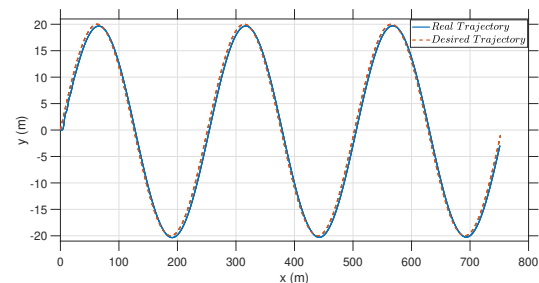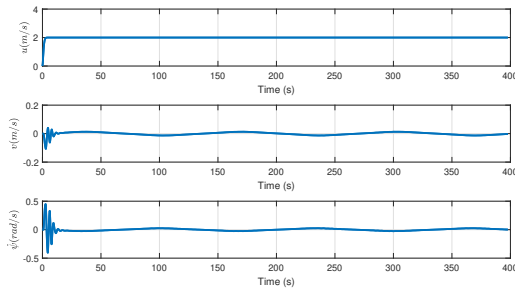

Fig. 17. Zigzag Path Following.

Fig. 18. Nonlinear Model Responses u, v and $\dot{\psi}$ during Zigzag Path Following.

Acknowledgment

References

[1] W. Elliott, R. Gonzales, N. Blas, A. Ramírez, C. Maldonado, M. Flores, and M. Jacinto, "Seguimiento de las pesquerías y calidad ambiental 2001-2005. imarpe huacho," *Inf Inst Mar Perú*, vol. 39, no. 1-2, 2012.

[2] J. Alves, P. Oliveira, R. Oliveira, A. Pascoal, M. Rufino, L. Sebastiao, and C. Silvestre, "Vehicle and mission control of the delfim autonomous surface craft," in *2006 14th Mediterranean Conference on Control and Automation*. IEEE, 2006, pp. 1–6.

[3] S. Wirtensohn, J. Reuter, M. Blaich, M. Schuster, and O. Hamburger, "Modelling and identification of a twin hull-based autonomous surface craft," in *2013 18th International Conference on Methods & Models in Automation & Robotics (MMAR)*. IEEE, 2013, pp. 121–126.

[4] Y. Peng, Y. Yang, J. Cui, X. Li, H. Pu, J. Gu, S. Xie, and J. Luo, "Development of the usv 'jinghai-i'and sea trials in the southern yellow sea," *Ocean engineering*, vol. 131, pp. 186–196, 2017.

[5] C. Zhou, S. Gu, Y. Wen, Z. Du, C. Xiao, L. Huang, and M. Zhu, "The review unmanned surface vehicle path planning: Based on multi-modality constraint," *Ocean Engineering*, vol. 200, p. 107043, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0029801820301177

[6] J. C. Cutipa-Luque, D. C. Donha, J. L. D. Dantas, L. M. [de Oliveira], and E. A. [de Barros], "Robust control of an underactuated auv," *IFAC Proceedings Volumes*, vol. 45, no. 27, pp. 138 – 143, 2012, 9th IFAC Conference on Manoeuvring and Control of Marine Craft. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474667016312186

[7] L. Qiao, S. Ruan, G. Zhang, and W. Zhang, "Robust h2 optimal depth control of an autonomous underwater vehicle with output disturbances and time delay," *Ocean Engineering*, vol. 165, pp. 399–409, 2018.

[8] C. Z. Ferreira, R. Cardoso, M. E. M. Meza, and J. P. J. Ávila, "Controlling tracking trajectory of a robotic vehicle for inspection of underwater structures," *Ocean Engineering*, vol. 149, pp. 373 – 382, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0029801817307540

[9] H. Zheng, J. Wu, W. Wu, and Y. Zhang, "Robust dynamic positioning of autonomous surface vessels with tube-based model predictive control," *Ocean Engineering*, vol. 199, p. 106820, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0029801819309187

[10] M. Breivik and T. I. Fossen, "Guidance laws for autonomous underwater vehicles," *Underwater vehicles*, vol. 4, pp. 51–76, 2009.

[11] J. C. C. Luque and D. C. Donha, "Auv robust guidance control," *IFAC Proceedings Volumes*, vol. 41, no. 1, pp. 85–90, 2008.

[12] D. Pearson, E. An, M. Dhanak, K. von Ellenrieder, and P. Beaujean, *High-level fuzzy logic guidance system for an unmanned surface vehicle (USV) tasked to perform autonomous launch and recovery (ALR) of an autonomous underwater vehicle (AUV)*. IEEE, 2014.

[13] L. Wan, Y. Su, H. Zhang, B. Shi, and M. S. AbouOmar, "An improved integral light-of-sight guidance law for path following of unmanned surface vehicles," *Ocean Engineering*, vol. 205, p. 107302, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0029801820303437

[14] T. I. Fossen and A. M. Lekkas, "Direct and indirect adaptive integral line-of-sight path-following controllers for marine craft exposed to ocean currents," *International Journal of Adaptive Control and Signal Processing*, vol. 31, no. 4, pp. 445–463, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/acs.2550

[15] W. Caharija, E. I. Grøtli, and K. Y. Pettersen, "Semiglobal exponential stability of a counter-current and co-current guidance scheme," *IFAC-PapersOnLine*, vol. 51, no. 29, pp. 274 – 280, 2018, 11th IFAC Conference on Control Applications in Marine Systems, Robotics, and Vehicles CAMS 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896318322043

[16] B. Qiu, G. Wang, and Y. Fan, "Predictor los-based trajectory linearization control for path following of underactuated unmanned surface vehicle with input saturation," *Ocean Engineering*, vol. 214, p. 107874, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0029801820308416

[17] T. I. Fossen, *Handbook of marine craft hydrodynamics and motion control*. John Wiley & Sons, 2011.

[18] J. C. Cutipa Luque, "Identificação e controle de um veículo submersível autônomo sub-atuado." Ph.D. dissertation, Universidade de São Paulo, 2012.

[19] S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control: Analysis and Design*. John Whiley & Sons, 2005.

Passenger Communication System for Next-Generation Self-Driving Cars: *A Buddy*

M Talha Bin Ahmed Lodhi[1]

Control, Automotive and Robotics Lab
aff. with NCRA
Department of Computer Science and Information Technology
Mirpur University of Science and Technology (MUST)
Mirpur-10250, AJK Pakistan

Faisal Riaz[2]

Control, Automotive and Robotics Lab
aff. with NCRA
Department of Computer Science and Information Technology
Mirpur University of Science and Technology (MUST)
Mirpur-10250, AJK Pakistan

Yasir Mehmood[3]

Department of Computer Science and Information Technology
Mirpur University of Science and Technology (MUST)
Mirpur-10250, AJK Pakistan

Muhammad Farrukh Farid[4]

Control, Automotive and Robotics Lab
aff. with NCRA
Department of Computer Science and Information Technology
Mirpur University of Science and Technology (MUST)
Mirpur-10250, AJK Pakistan

Abdul Ghafoor Dar[5]

Department of Computer Science and Information Technology
Mirpur University of Science and Technology (MUST)
Mirpur-10250, AJK Pakistan

Muhammad Atif Butt[6]

Control, Automotive and Robotics Lab
aff. with NCRA
Department of Computer Science and Information Technology
Mirpur University of Science and Technology (MUST)
Mirpur-10250, AJK Pakistan

Samia Abid[7]

Control, Automotive and Robotics Lab
aff. with NCRA
Department of Computer Science and Information Technology
Mirpur University of Science and Technology (MUST)
Mirpur-10250, AJK Pakistan

Hasan Ali Asghar[8]

Control, Automotive and Robotics Lab
aff. with NCRA
Department of Computer Science and Information Technology
Mirpur University of Science and Technology (MUST)
Mirpur-10250, AJK Pakistan

Abstract—With the rapid emergence of autonomous vehicles, there is a need to build such communication systems which help the passengers to communicate with autonomous vehicles (AVs) robustly. In this regard, this research work presents a multimodal passenger communication system. The communication system is known as "buddy" for AVs. Buddy is an all in one control system for AVs which incorporates touch, speech, text, and emotion recognition methods of interaction. Buddy makes it easy for passengers to interact with AVs. It enables the communication between the passengers and the AV which eventually provides a safe driving experience. Moreover, we have proposed and developed our own simulator two evaluate the performance of our proposed passenger communication system. We have also conducted extensive infield-tests to test the effectiveness of the proposed system. The extensive rigor analysis validates the results and hence the significance of the proposed passenger communication system.

Keywords—*Autonomous Vehicles (AVs); passenger communication system; the simulation engine*

I. INTRODUCTION

Communication is an important part of human beings to convey their messages to each other. Popularizing communication in the vehicles is key in the proliferation of this rapidly emerging technology. Autonomous vehicles (AVs) are the latest type of robots and there is a need to build a communication system that helps the passengers to communicate with AVs. This is signified by the work of Wang et al. [1] which presents methods of preparation for the mass emergence of AVs. They show how AV capabilities can be extended to underwater army submarines thus signifying the impact of AVs in military applications. Hence, the need arises to engineer autonomous systems with Human-Computer Interaction (HCI) capabilities.

HCI is a field that helps humans to interact with machines. Extensive research has been done in HCI to make systems autonomous. Valeria et al. [2] have evaluated the factors played by psychology in designing systems. The role of HCI is vast and applies to many unique types of systems.

Moreover, in the HCI-based research, different fields have been explored at large. For example, Venuto et al. [3] have presented the state of the art HCI system for the remote control of a mechatronic actuator, such as a wheelchair or a car. Leminen et al. [4] have proposed a project detailing an HCI for a smart home geared for research purposes. Pons et al. [5] worked on interactive software geared for animals, referred to as ACI (animal-computer interaction). However, the role of HCI has not been explored to tailor novel interaction solutions for a passenger to AVs.

A lot of research has been done to make HCI systems for controlling AVs. Woo et al. [6] have worked on remote driving tools for AVs which are a collaboration between remote driving tools and sensor fusion displays. Xiao et al. [7] have proposed sensor fusion methodologies while Poveda et al. [8] have developed hybrid source seeking controllers specifically for AVs. However, these HCI systems cannot compute the effective state of the passenger sitting in AV.

Considering the above-mentioned limitations, we have tailored affective computing inspired passenger to AV (P-AV) interaction system. Since there is a need for building a communication system that helps the passengers to communicate with AVs, we propose a passenger communication system (PCS) which consists of 4 modules as shown in Fig. 1. For the testing and evaluation of the proposed system in different environments, we have proposed and developed our simulator. Moreover, the extensive testing of the proposed system has been done on EMO (AV) to test the performance of the proposed PCS.

The rest of this paper is organized as follows. Section II presents the literature review. The system methodology has been formally defined in Section III. The simulation results are described in Section IV. Finally, the conclusions are presented in Sections V.

II. Literature Review

Communication in autonomous vehicles' is an important emerging technology. We discuss its importance in different threads as follows. Hernandez et al. [9] describes how different types of interactions can help to measure and manage the stress of a driver. For this purpose, they join both wearable technology and business into the steering wheel of a vehicle which allows no upsetting spotting. Ragot et al. [10] describe how human emotions can recognize by physiological signals. They trained models using SVM classifier and compare emotion recognition accuracy by using laboratory sensor "Biopac MP150" and wearable sensor "Empatica E4". P. Karthikeyan et al. [11] focused on identifying the stress analysis with the advanced processing of physiological signals which is based on the mental arithmetic task. Chang et al. [12] have proposed physiological signals for an emotion recognition system based on the support vector recognition.

A driver's driving may affect due to emotions and non-driving associated reasoning tasks which may cause traffic accidents. Boril et al. [13] focused on the task of classification of emotion and cognitive load of real driving scenarios to estimate "speech production-based" and "Cepstral-based acoustic" features. They applied SVM and GMM classifiers that gave 79% and 95.2% classification performance in the task of classification of emotion as neutral vs. negative and

cognitive classification respectively. Tarnowski et al. [14] presented a 3D face model to calculate the features. The results of these features are based on facial expressions. With 2D images, light condition, and movement of head positions play a sensitive role to recognize emotions using cameras. For features classification, the "K-NN" classifier and "MLP neural network" are performed. The experiments showed the best results of these seven emotional states. However, we can also recognize the emotions of passengers in AV via facial expressions. In another work, Hickson et al. [15] presented a novel way for data collection and also provide a novel approach to increase the accuracy of CNN deep learning through personalization. Currently, the emotion recognition system is suffering in facial image conversion performance. To overcome this problem, Wang et al. [16] proposed a novel approach for emotion recognition system using "Jaya algorithm". This algorithm has been used to train the dataset and guaranteed that it won't stick with the training set to local optimum point and has fewer necessities over hyper-parameters. In this article, Eyben et al. [17] have presented a novel approach to improve the intelligent measures for driving safety in automatic driving systems. Gordon McIntyre and Roland Gocke [18] discussed particular problems faced in emotion recognition system and try to deal with natural way. They presented a novel approach which incorporates semantic descriptions and feature sets of computer vision. Computer vision process is used to detect facial expression emotions with combination muscle movements' codebook. Automatic speech recognition and computer vision are a dealer in machine learning and pattern matching. Their framework, effective communication consists of the generic model with the ontology domain. They recognize emotions by facial expressions, but they didn't consider a passenger's emotions of an AV.

III. System Methodology

This section provides details about the proposed agent architecture of the PCS as follows:

A. Proposed Agent Architecture of Passenger Communication System

In this section, we present the proposed agent architecture of the PCS as shown in Fig. 1. Our agent consists of six modules which includes the sensory module, microsoft cognitive services, artificial amygdala, dialog generation, strategy selection and actuator module. Initially, our agent receives the perceptual data from the sensory module. From the obtained data, it will analyze the facial expressions of the passenger/client. The data is then forwarded to the microsoft cognitive services which reads the emotions from the facial expressions of the passenger/client. The results generated by this module are then forwarded to the artificial amygdala module which processes the fear level experienced by the passenger/client. After examining the fear evaluated by the passenger/client two processes will be executed:

- Strategy Selection Module
- Dialog Generation Module

First, the strategy selection module creates a direct link from the artificial amygdala module. It picks a strategy upon
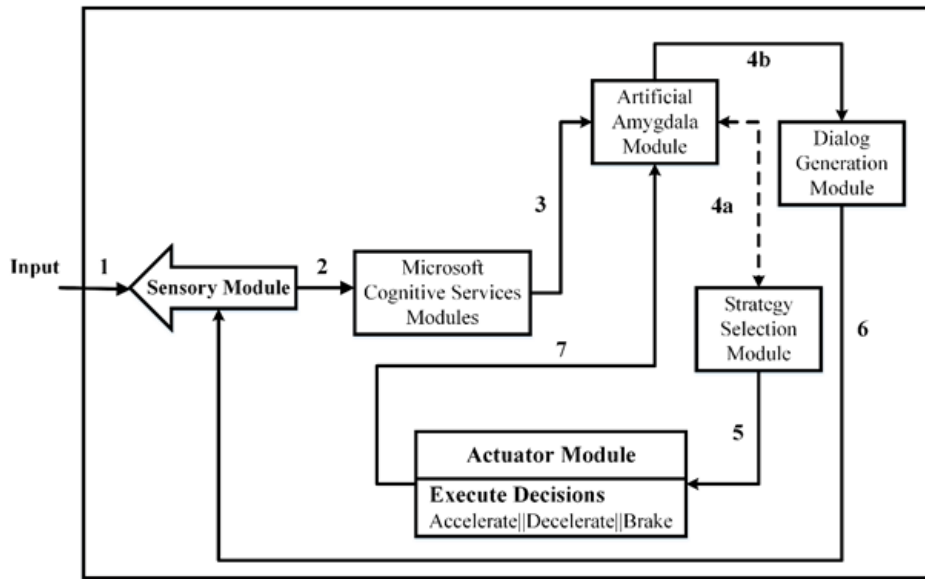
Fig. 1. The Proposed Agent Architecture of the Passenger Communication System

---

Algorithm 1 Proposed algorithm of PCS

1: **procedure** PROCESSINPUT(TEXT)
2:     $action \leftarrow LoadActionFile()$
3:     $command \leftarrow Action.Parse(text)$
4:     **if** $Command! = Empty$ **then**
5:         **return** Command
6:     **if** $Command == Empty$ **then**
7:         Throw Exception
8: **procedure** PROCESSSPEECH(VOICE_ INPUT)
9:     $InitializeSpeechModule$
10:     $Text \leftarrow SpeechModule.Recognize(voice\_input)$
11:     $Command \leftarrow CallProcessInput(text)$
12:     **return** Command
13: **procedure** RECOGNIZEEMOTION()
14:     **Initialize** CameraModule
15:     $Photo \leftarrow CameraModule.GetFrame()$
16:     $Result \leftarrow CognitiveServices(Photo)$
17:     **if** $Result.Contains(Face)$ **then**
18:         **return** Result.Face.DominantEmotion
19:     **if** $Command == Empty$ **then**
20:         Throw Exception

Fig. 2. Simulation Screen of the Controller Integrated into "Buddy" App

IV. EXPERIMENTAL SETUP AND RESULTS

In this section, we discuss the results of our proposed PCS (known as "Buddy"). For our PCS "Buddy" can be used on different platforms we have first targeted desktop application. Our desktop application is fully functional on any distribution of windows. Here, we briefly discuss the functionalities of PCS as follows:

A. PCS

Our proposed PCS is very user-friendly. It communicates with users to provide a comfortable environment in which a user can say whatever they want as shown in Fig. 2. We have implemented our language processor called PML which uses defined Regular Expression (REG-EX). REG-EX understands what the user currently desires for at time 't'. The user can type or say it and our language processor shows results accordingly. The language processor is capable enough to process what the user says and gives results e.g. "I want to go to Lahore". It checks in its predefined REG-EX and tells whether you have entered a valid statement or not.

After exploring the Home Section, next comes the con-

which the agent drives itself and then it sends the strategy to the actuator module which in turn accelerates/decelerates or applies the Brake. It continuously checks the fear level from the artificial amygdala and selects the strategy after checking what sort of emotion is being experienced by the passenger/client. Secondly, after checking the emotion of the passenger/client the agent tries to soothe the passenger/client with its pattern matching language (PML) in the dialog generation module and assess the expression of the passenger/client through the input obtained from the sensory module. The sensory module consists of multiple sensors that are commonly deployed with AVs like sonar, camera, Lidar. The step-wise procedure of the proposed PCS is provided in the Algorithm 1.
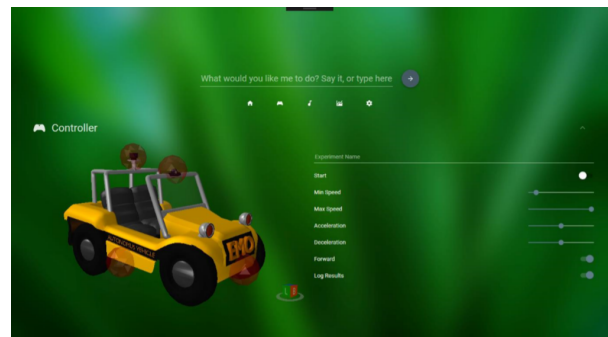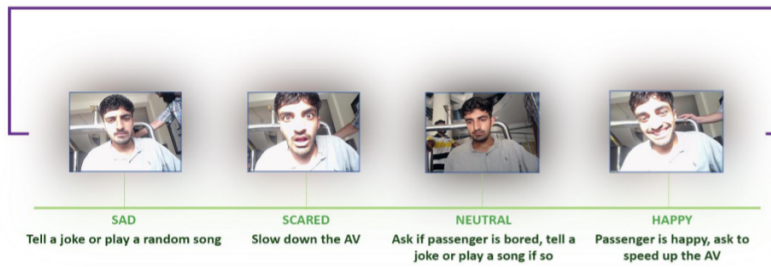
Fig. 3. Various Considered Facial Expressions of the Passenger
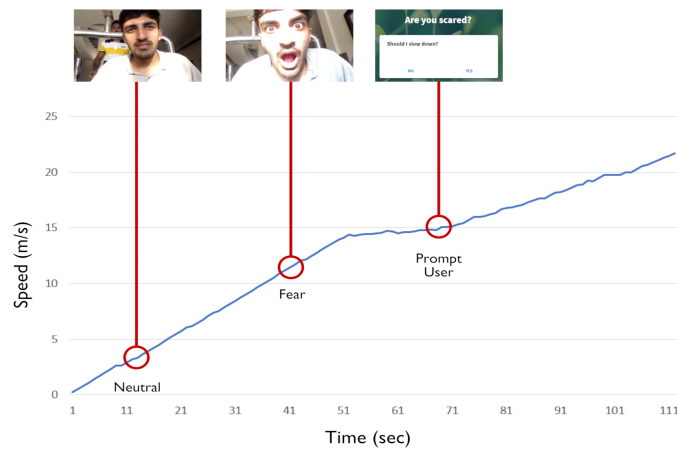

Fig. 4. Effect of 'Fear' on Driving Strategy

Fig. 5. Effect of Feeling 'Fear' on Driving Strategy when a Passenger wants the AV to Accelerate

troller module. This module is directly related to any sort of autonomous agent mainly "vehicles". This module provides the 3D display of the desired autonomous agent. In Fig. 2 on the left side of the screen we have implemented the 3D model of EMO the first autonomous vehicle of Pakistan. As we can see in the above figure there are red circles on certain parts of EMO. These red circles indicate that there is no lidar, camera or sonar yet attached to the system. At the bottom of

the screen is a cube that is used for 360°-degree rotation of EMO.

In the rightmost corner, we have set an experiment logger. We simply have to write the name of the experiment that we want to log e.g. "BIKE_AV_COLLISION_EXPERIMENT1". The experiment will start logging after the log results button is turned ON. Otherwise, the experiments will not be logged. The Start Button is used to turn ON the AV. We can also

adjust the max and min speed as well as acceleration and deceleration of the AV. The next feature is about moving the AV either Forward or Backwards as the user desires.

We have performed the emotion recognition test on the bases of the facial expression of the passenger and examined the driving strategy being employed by the AV to sustain a safe driving experience. In this regard, several images are provided to the PCS and the adopted strategy is noted against each recognized emotion. Fig. 3 represents various facial expressions of the passenger and the messages passed to the AV in return. Fig. 4 represents the result of the "fear" emotion on the driving strategy being employed by the AV. Here, the x-axis represents the recording time of the simulation (in seconds) while the speed of the vehicle is depicted in the y-axis. Initially, when the system recognized that the passenger is feeling "neutral" the AV started accelerating its speed. At 41 seconds, when AV reached its maximum acceleration the passenger started feeling fear; the PCS sensed the "fear" and generated the prompt "Are you scared? should I slow down?". Upon receiving a positive response, the AV started decelerating itself. At 75 seconds, the vehicle reached its least acceleration, the PCS generated a prompt to confirm if the passenger is happy with the speed and continues with the same speed on receiving the consent of the passenger. In the same way, Fig. 5 represents the driving strategy adopted by the AV when the passenger felt "fear". However, in this scenario, when the prompt generated "are you scared?" on sensing the "fear" from the facial expression of the passenger, the passenger withheld to decrease the speed of the vehicle and consequently, AV continued to increase its speed keeping road conditions under consideration.

Whereas, the driving strategy being adopted by the AV in response to the "happy" emotion is represented in Fig. 6. Initially, the facial expressions of the passenger were "neutral". Hence, the AV started accelerating its speed considering the road conditions. While accelerating (at 43 seconds), the PCS recognized the "happy" emotions of the passenger and generated a prompt "you seem happy! want to speed up?" to procure the assent of the passenger to accelerate the AV. In this scenario, the passenger agreed to increase the speed and consequently, the AV started accelerating while considering the road situations. However, Fig. 7 also represents the driving strategy adopted in response to the "happy" emotion. But, in this scenario, the passenger refused the PCS to increase the speed of the vehicle, and hence the AV has maintained the same speed considering the varying conditions of the road.

V. Testing and Evaluation

To assess the performance of an AV in a particular environment, many simulation engines have been developed. The existing simulators used by different driving groups are the racing simulators mainly built on the gaming engines; however, these simulators do not generate the exact results and do not provide the required feedback. Moreover, the existing simulators do not different environment scenarios like urban, desert, snowy conditions, etc.

The proposed simulation exhibits the features of full 3D support, behavior space, and enables to integrate human emotions in AVs. Besides, it allows us to integrate multiple sensors on AVs i.e. camera, SONAR, LiDAR, etc. Our developed simulator consists of its own custom road/map designer which makes it easier to generate different road scenarios to test the performance of an AV. A 3D map created using our map designer is shown in Fig. 8. Several environments have been designed which allow the testing of AVs in different scenarios based on their location. These include environments as shown in Fig. 10.

Whereas, we have assessed the perfomance of our proposed system in different scenarios as shown in Fig. 10. To set up experiments, the speed of AVs randomly set to [2-4]. Where, each test has been executed 5 times and saved in the log file

In the first scenario (Fig. 10(a)), a single path was given to the AV and it managed to reach its destination without any deviation. While in the second scenario, a location was fed to the simulator; however, two paths were available to reach the destination as shown in Fig. 10(b). When the vehicle reached the T-zone where it had to decide which path to select in order to continue its operation, it detected an obstacle on one of the route and continued to pursue its operation on the other path. While. in the 3rd scenario (Fig. 10(c)), a destination was fed to the simulator and multiple ways were available to reach the destination, however, using its optimal path selection algorithm, our proposed system optimally selected a path to continue a smooth journey without any obstacles.

A. Discussion

Our Sim-Engine is fully functional and has been deployed in the testing of first autonomous vehicle of Pakistan EMO. Our Sim-Engine is capable of carrying out multiple scenarios based on different aspects and also it is capable of choosing different driving strategies when faced with different circumstances such as what to do when a large truck is in front of you. Furthermore, our Sim-Engine also has the capability of checking out passenger's emotions that what he/she is feeling at that time and based on that it will generate a response which will help the user to calm down or get happy. Our Sim-Engine also logs all the data (results) which are being collected at that particular instance. A comparison study of the proposed simulator is given in Table I.

VI. Conclusion

Communication in AVs is a very hot research area. However, very few systems have been proposed to enable communication between passengers and the AV in order to provide a comfortable driving experience. Hence, software-based solutions are of immense significance to enable communication between passenger and AV. Though, considering the need of a reliable PCS for the AVs, this research work presents a novel PCS to test various autonomous agents. The proposed PCS is very user-friendly and can be embedded in any sort of autonomous agent, however, we have mainly targeted AVs. The performance of the proposed PCS has been verified through in-field experiments and evaluating its effects on the driving strategy being employed by the AV. In the future, we aim to investigate performance of the proposed PCS to ensure the road safety considering different constraints.
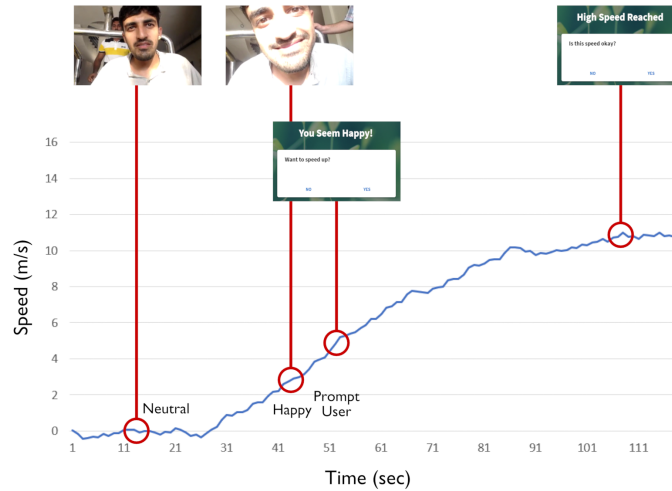
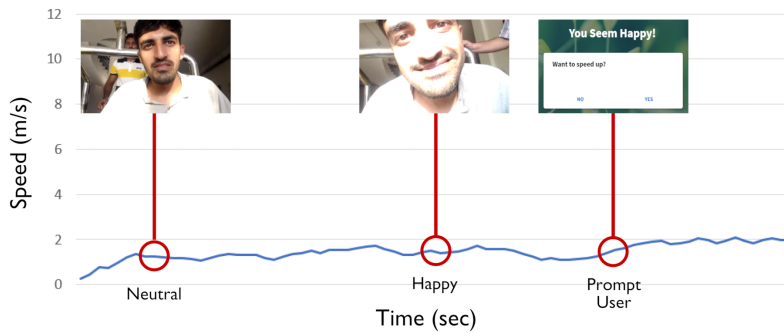Fig. 6. Effect of Feeling 'Happy' on Driving Strategy

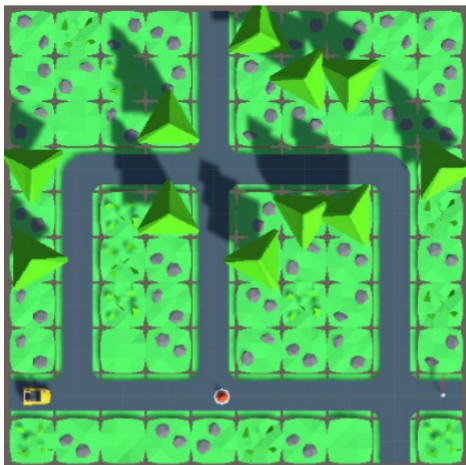Fig. 7. Effect of Feeling "Fear" on Driving Strategy when a Passenger doesn't want the AV to Accelerate

Fig. 8. Map Generated by our Custom Map Designer

REFERENCES

[1] Wang, Wen-hua, Ya-zhen Du, Lin-lin Wang, Yi Huang, Yu-xin Yao, and Hao Gao. "Development of the New Concept of Sandglass-Type Floating Production, Storage and Offloading System." In The 28th International Ocean and Polar Engineering Conference. International Society of Offshore and Polar Engineers, 2018.

[2] Righi, Valeria, Sergio Sayago, and Josep Blat. "When we talk about older people in HCI, who are we talking about? Towards a 'turn to community' in the design of technologies for a growing ageing population." International Journal of Human-Computer Studies 108 (2017): 15-31.

[3] De Venuto, Daniela, Valerio F. Annese, and Giovanni Mezzina. "An embedded system remotely driving mechanical devices by P300 brain activity." In Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017, pp. 1014-1019. IEEE, 2017.

[4] Leminen, Seppo, and Mika Westerlund. "Categorization of innovation tools in living labs." Technology Innovation Management Review 7, no. 1 (2017).

[5] Pons, Patricia, Alicia Carrion-Plaza, and Javier Jaen. "Remote interspecies interactions: Improving humans and animals' wellbeing through mobile playful spaces." Pervasive and Mobile Computing 52 (2019): 113-130.

[6] Woo, Jinseok, Janos Botzheim, and Naoyuki Kubota. "A socially

TABLE I. A COMPARISON OF THE PROPOSED SIM-ENGINE WITH STATE OF THE ART SIMULATORS

Features/Simulators	NETLOGO	CARLA	Orchestra	NVIDEA	Appolo	Unity	Sim Engine (Unity-Based)
Specific to AVs		✓	✓	✓	✓		✓
Full 3D Support		✓	✓	✓	✓	✓	✓
Emotion Enabled							✓
Behavior Space	✓						✓
Environment Variations		✓			✓		✓
Custom Map Designer					✓		✓
Inbuilt Sensor Emulation		✓	✓	✓			✓
Open Source	✓	✓					✓
Cloud Based				✓	✓		✓
Windows Support	✓	✓	✓	✓	✓	✓	✓
Mac Support	✓				✓	✓	✓
Linux Support	✓				✓	✓	✓

Fig. 9. Environment Variations: (a) Forest View ; (b) Desert View ; (c) Rain View, (f) City View(c) Snowy View, (f) Mud View

Fig. 10. Environment Variations: (a) Single Path for an AV ; (b) Two Paths for an AV ; (b) Multiple Paths for an AV

interactive robot partner using content-based conversation system for information support." Journal of Advanced Computational Intelligence and Intelligent Informatics 22, no. 6 (2018): 989-997.

[7] Xiao, Fuyuan. "Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy." Information Fusion 46 (2019): 23-32.

[8] Poveda, Jorge I., Mouhacine Benosman, Andrew R. Teel, and Ricardo G. Sanfelice. "A hybrid adaptive feedback law for robust obstacle avoidance and coordination in multiple vehicle systems." In 2018 Annual American Control Conference (ACC), pp. 616-621. IEEE, 2018.

[9] Hernandez, Javier, Daniel McDuff, Xavier Benavides, Judith Amores, Pattie Maes, and Rosalind Picard. "AutoEmotive: bringing empathy to the driving experience to manage stress." In Proceedings of the 2014 companion publication on Designing interactive systems, pp. 53-56. ACM, 2014.

[10] Ragot, Martin, Nicolas Martin, Sonia Em, Nico Pallamin, and Jean-Marc Diverrez. "Emotion recognition using physiological signals: laboratory vs. wearable sensors." In International Conference on Applied Human Factors and Ergonomics, pp. 15-22. Springer, Cham, 2017.

[11] Palanisamy, K., M. Murugappan, and S. Yaacob. "Multiple physiological signal-based human stress identification using non-linear classifiers." Elektronika ir elektrotechnika 19, no. 7 (2013): 80-85.

[12] Chang, Chuan-Yu, Jun-Ying Zheng, and Chi-Jane Wang. "Based on support vector regression for emotion recognition using physiologi-

cal signals." In The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1-7. IEEE, 2010.

[13] Boril, Hynek, Seyed Omid Sadjadi, and John HL Hansen. "UTDrive: Emotion and cognitive load classification for in-vehicle scenarios." In The 5th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems. 2011.

[14] Tarnowski, Paweł, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz J. Rak. "Emotion recognition using facial expressions."

Procedia Computer Science 108 (2017): 1175-1184.

[15] Hickson, Steven, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. "Eyemotion: Classifying facial expressions in VR using eye-tracking cameras." In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1626-1635. IEEE, 2019.

[16] Wang, Shui-Hua, Preetha Phillips, Zheng-Chao Dong, and Yu-Dong Zhang. "Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm." Neurocomputing 272 (2018): 668-676.

[17] Eyben, Florian, Martin Wöllmer, Tony Poitschke, Björn Schuller, Christoph Blaschke, Berthold Färber, and Nhu Nguyen-Thien. "Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car." Advances in human-computer interaction 2010 (2010).

[18] McIntyre, Gordon, and Roland Göcke. "The composite sensing of effect." In Affect and Emotion in Human-Computer Interaction, pp. 104-115. Springer, Berlin, Heidelberg, 2008.

A Hybrid Model based on Radial basis Function Neural Network for Intrusion Detection

Marwan Albahar[1], Ayman Alharbi[2], Manal Alsuwat[3], Hind Aljuaid[4]
Umm Al Qura University[1,2], Taif University[3,4]

Abstract—An Intrusion Detection System (IDS) is a system that monitors the network for identifying malicious activities. Upon identifying the unusual activities, IDS sends a notification to the network administrators to warn about the hackers' hostile activities. To detect intrusion, signature-based systems are considered to be one of the most effective methods. However, they cannot detect new attacks. Additionally, it is costly and challenging to keep the attack signatures database up to date with known signatures, which constructed a significant drawback. Neural networks are capable of learning through input patterns and have the potential to generalize data. In this paper, we propose a hybrid model based on Directed Batch Growing Self-Organizing Map (DBGSOM) combined with a Radial Basis Function Neural Network (RBFNN) detecting abnormalities in the network. Based on our experiment, the proposed model performed well and has resulted in satisfactory performance measures compared to Self-Organizing Maps and Radial Basis Function Neural Network (SOM&RBFNN) model.

Keywords—*Intrusion detection; neural network; radial basis function; directed batch growing self-organizing map*

I. INTRODUCTION

The immense flow of network traffic has created opportunities for attackers to breach privacy and harm the integrity of the network and its users. In such circumstances, intrusion detection has become a crucial part of computer security to ensure that attacks and intrusion activities can be detected [1]. Intrusions are threats to network systems that may come in different forms, such as damaging the systems and making it unavailable, information examination, and information manipulation [2]. There are two kinds of intrusions passive and active, where passive intrusions are surreptitiously and without detection, whereas active intrusions lead to change and effect to network resources. Intrusion Detection System (IDS), an influential approach, is designated to detect normal and malicious activities in computers. It is a system that defends the network by identifying suspicious activities while analyzing the network traffic to detect and counter intrusions [2] timely. An IDS has two main methods of detecting intrusions: signature-based IDS and anomaly-based IDS. Signature-based detection is utilized to searching for a "signature" pattern or known attacks. This type of IDS, it requires regular updates to currently common signatures or identities to ensure that the intruders' database is current. Nevertheless, attackers can change small things in signatures, so the databases cannot recognize it. Consequently, a new attack type may not be picked up because the signature doesn't exist in the database. Moreover, the larger the databases, the more processing to analyze each connection and verify it. In contrast to signature-based IDS, anomaly-based detection is employed to detect known and unknown attacks depending on learning their behavior by specifying

observations that deviate from a basic model in a computer network and inform the network's administrator to take necessary actions. The main advantage of anomaly-based detection is the ability to detect unknown attacks [1,3]. As a result of their importance, different systems have been proposed for intrusion detection by many researchers. Among these systems, machine learning models and specifically neural networks can effectively detect malicious activities on a network by getting trained using enough intrusion detection recorded data [4-7]. Non-neural network machine learning models such as SVM have specific limitations such as low repetition attack detection rates, detection instability, and training process complexity [7]. Neural network models have been used for anomaly detection, such as autoencoders and variational autoencoders (VAEs). Autoencoders are composed of encoder and decoder networks that are sequentially connected. An encoder can compress the input data and a decoder to reconstruct the input data. Autoencoders attempt to reduce the error in reconstruction (i.e., the difference between the decoder output and the original input). This error is used as an anomaly score to detect anomalies [8-9]. Small reconstruction errors correspond to normal data, while larger reconstruction errors correspond to anomalous data [8]. A VAE, which is a directed probabilistic graphical model (DPGM), combines Bayesian inference with the autoencoder framework. It provides a probability measure rather than an error of reconstruction as an anomaly score. As probabilities are more principled and objective than reconstruction errors and do not require model specific thresholds for anomalies to judge, VAEs outperform autoencoders in intrusion detection. However, for natural inference and learning, VAE assumes complicated data distributions can be modeled with a smaller group of latent variables whose probability density distributions are Gaussian. At the same time, studies show the data distribution is usually multi-modal, and a single Gaussian distribution cannot be considered before the latent space [8]. As a result, some other methods, such as the GGM-VAE method are proposed to compensate for this shortcoming. In GGM-VAE, gated recurrent unit (GRU) cells are used to find the correlations between data. Then, the GGM-VAE method uses Gaussian Mixture prior in the latent space to characterize the multi-modal data using a series of Gaussian distributions and applies VAE. While GGM-VAE yields a better detection rate over VAE, similar to previously discussed deep networks, it needs extensive data for training. On the other hand, Radial basis function neural networks (RBFNNs), which are capable of classifying patterns in nonlinear space and linear encoding, can estimate non-parametric multi-dimensional functions through a limited set of educational information [7]. Also, RBFNNs are rapid, comprehensive, and yields highly accurate in training [7]. New research by Mohammadi and Amiri also verifies their

effectiveness in detecting intrusion in the networks [7]. This research first uses self-organizing maps (SOMs) to cluster the data. It then uses the cluster centers to determine the number of radial basis functions and set the parameters of these functions in RBFNN. SOMs can profoundly reduce the dimensionality of the data and therefore reduce the computational runtime of the process. However, one issue with SOMs is that they need a predefined structure, including determining the number of clusters/neurons. To handle this limitation growing SOMs (GSOMs) and researchers present its variations. One of the recent variations of GSOMs is directed batch GSOMs (DBG-SOMs) which not only resolve the limitation of SOMs but also introduces a suitable mechanism to discover an appropriate growth position and assigning new neurons initial weight vectors [7]. A batch learning strategy for growing self-organizing maps was employed in DBGSOM to resolve SOM and GSOM models' issues. Consequentially, DBGSOM has a better ability to conservation the topology and reduce exposure for twisting and tangling in the network [10]. Furthermore, RBFNNs are rapid and comprehensive in training and produce high precision in detecting intrusion [7]. In this study, we measure the effectiveness of combining DBGSOM and RBFNN to build an effective anomaly-based intrusion detection system. Then, we compare its efficacy to SOM with RBFNN. We first explain SOMs and its limitations and then discuss DBGSOM before representing the RBFNN process. Then, we apply our proposed method on three publicly available datasets. Finally, based on our experimental results, we show that our hybrid model outperforms the SOM&RBFNN model.

The paper proceeds as follows: In Section II, we provide related works. In Section III, we describe DBGSOMs and RBFNN used in this work. In Section IV, we introduce our proposed model. Next , we introduce the dataset utilized in Section V. In Section VI, we shows our experimental results, and Section VII concludes and gives our future works.

II. Related Work

There is a large number of published researches that prove the effectiveness and ability of machine learning especially neural networks in intrusion detection techniques, some of them will be summarized in the following: A recent study by Vinayakumar et al. [11], shows convolutional neural networks (CNNs) and its various architectures styles generally perform good efficiency compared to classical machine learning classifiers [12]. They modeled network traffic with various learning methods including multi-layer perceptron (MLP), CNN, CNN-recurrent neural network (CNN-RNN), CNN-long short-term memory (CNNLSTM) and CNN-gated recurrent unit (GRU) and reported their results for the NSL-KDD dataset. Their reported detection rate seems promising. However, since these methods need millions of known good and bad network connections for training, and obtaining a right deep model usually needs trying complex architecture, which they consequently require more computational cost, they cannot be yet the best option to be employed for the job. In another study, Li et at. used a representation learning method of graphic conversion to transform intrusion data into image shape and then apply CNN on the transformed features to detect anomalies [12]. In this study, they, also, only applied their method on the NSL-KDD dataset and achieved Impressive results. However, clearly

the converting the data to the image form proposed is time-consuming and the method also needs huge data for training to perform well.

Researchers in [1] applied a long short-term memory (LSTM) model to discover intrusion and utilized the CIDDS001 dataset for assessing the LSTM model's performance and they discovered that it outperforms on SVM, MLP, and Naïve Bayes techniques concerning to multiclassification problem. Regarding the self-organization map (SOM), Sadeq and Ahmad studied the effectiveness of combining the SOM with a backpropagation neural network (BPNN) to reveal intrusion systems [2]. The proposed approach is divided into two-stage. In the first classification stage, SOM was utilized for categorizing normal traffic from attacks. Then, in the second classification stage, BPNN is utilized for categorizing the attacks into DoS, R2L, Probe, and U2R. They applied this proposed approach for dataset NSL KDD and the outcomes showed that the performance and precision of the intrusion detection system had improved.

In 2019, JIN et al. [13], proposed a new model based on the simple recurrent unit (SRU) and deep convolutional generative adversarial networks (DCGANs) to detection intrusion in the network. DCGAN was utilized to extract features from the raw-data directly and then create sufficient and balanced data samples from current attack samples. In addition to retaining the information that appeared in raw sample data. Because of the dependency that exists in LSTM, it has been replaced by SRU to allow the parallelization. Which led to improved speed of training 10 times than LSTM. Extensive experiments have been done on the dataset NSL-KDD to verify the efficiency of this model and it achieved satisfactory results in classification performance and accuracy in intrusion detection.

In addition, a previous study by Alia et al. [14], a radial basis function (RBF)-based intelligent intrusion detection system within the framework of approximation theory was applied to the dataset NSL-KDD through the k-means algorithm to detect the attack and classify its type under one category: DoS, Probe, R2L, and U2R. The results of attack classification were high for rare attacks (R2L and U2R) whereas low for frequent attacks (DoS and Probe).

In [7], a hybrid self-learner intrusion detection model was proposed using self-organized neural network algorithms and perceptron networks. The authors highlighted the advantages of the hybrid model, such as the self-learner ability of intrusion detection and memory storage. The authors also pointed out that RBF neural networks can learn quicker and more comfortable than MLP networks. Moreover, the authors showed the error rate of the proposed algorithm with the back error propagation method has a lower error rate, which reveals its higher accuracy than back error propagation method.

III. Background

A. Directed Batch Growing Self-organizing Maps (DBG-SOMs)

SOMs are the foundation of some machine learning models in anomaly-based intrusion detection systems [6, 7]. This is because SOMs are unsupervised and do not need any user feedback, while their output maps can be visualized and help

understand how data are clustered [6]. SOM is also an excellent tool for dimension reduction as the SOMs' weights can be considered as cluster centers and be good representatives of data. There are two necessary steps to implement the SOM algorithm: initially, finding the winner by competition between neurons. After that, the weight vector is adapted to the winning neurons and their topological neighbors. One of the main limitations for SOM is needing a pre-defined structure and learning parameter before initializing the training process. To resolve this issue, growing SOMs (GSOMs) were introduced, which require fewer epochs, offer a flexible structure, and allow the capability to learn the nonlinear manifolds at high dimensional feature space [10, 15]. Three main aspects in which the GSOM models are different are determining when and where to add new neurons and assigning proper weight vectors. Fig. 1 shows the initialization of the GSOM network that begins with four neurons at a rectangular network, guaranteeing that all neurons have a similar lattice state in the initialization stage. GSOM models fill all available positions around the candidate

Fig. 1. Initial Structure of GSOM with Four Neurons A, B, C and D on a Square Grid [15].

neuron. Consequently, because of misconfiguration and map twisting that can result from unexpected network growth and incorrect neuron addition and the initialization of weight, they reduce the topology preservation quality of the map despite their dynamic nature. A new variation of the GSOM model called directed batch GSOMs (DBGSOMs) [15] is recently proposed to resolve the issues of GSOM models by introducing a batch learning strategy for GSOMs. DBGSOMs have a better ability to conservation the topology and reduce exposure for twisting and tangling. This model directs the network growth appropriately within the feature space by looking at the cumulative error around candidate boundary neurons. As a result, DBGSOM offers appropriate mechanisms for finding a suitable growth position and assign primary weight vectors to new neurons; moreover, it has been verified that it has a better clustering capability than GSOM and SOM. It requires less time for learning data points manifold compared to GSOM because it generates fewer neurons [15].

B. Radial Basis Function Neural Network (RBFNN)

RBFNN is a forward network type based on function approximation theory. RBFNN has consisted of input, hidden, and output layers (Fig. 2). The hidden layer RBF of

RBFNN mostly utilizes one of the following nonlinear functions [16]: Gaussian function, poly-quadratic function, inverse poly-quadratic function, and thin plate spline function.
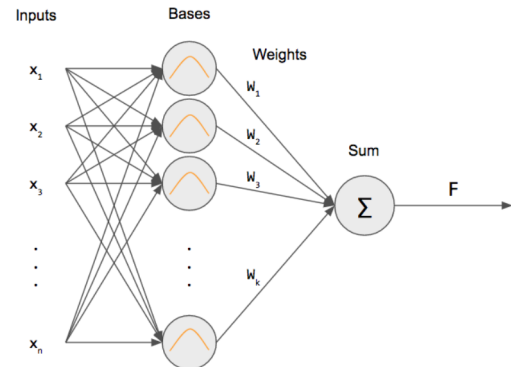

Fig. 2. The Framework of a Radial basis Function Neural Network (RBFNN)

Among these functions, Gaussian functions are the more popular ones. The RBF form used in the RBFNN is not essential to network performance while choosing c_i and σ_i (i.e., mean and standard deviation in the Gaussian function) is key to the whole network 's performance [16]. Clustering algorithms like K-means clustering and self-organizing maps can be used to cluster the data, and then use the clustering/SOM outputs for the Gaussian function in RBFNN. When using K-means clustering, the cluster centers c_i determine the mean values in the Gaussian function. The standard deviation of data in each cluster can be considered as σ_i in the Gaussian radial basis function. When self-organizing maps are used, the number of SOM neurons is the number of the clusters, and the final weights of the SOM networks can be used as the c_i for the Gaussian radial basis functions. Then, for standard deviations, for the cluster center i, we can compute the distance of the cluster i to the other cluster centers and then use $\frac{1}{2}$x minimum distance as the corresponding standard deviation. This lets RBFNN knows where to place the Gaussian functions.

IV. PROPOSED METHOD

We discussed the advantages of DBGSOM over SOM and explained the RBFNN method. As mentioned earlier, RBFNNs are fast and comprehensive in training and yield high precision. As Mohammadi and Amiri showed in [7], a combination of SOM with RBFNN is useful in detecting intrusion in a network. Therefore, in this paper, we use DBGSOM in conjunction with RBFNN for anomaly-based intrusion detection. The diagram for our hybrid model process is shown in Fig. 3.

V. DATASET

In the following, we explain the datasets which were used in our experiments. The three datasets are NSL-KDD, UNSW-NB15, and CICIDS2017 datasets, publicly available, and can be downloaded from their designated websites.

A. NSL-KDD

NSL-KDD is an intrusion detection benchmark data set suggested for resolving several inherent issues in the KDD'99
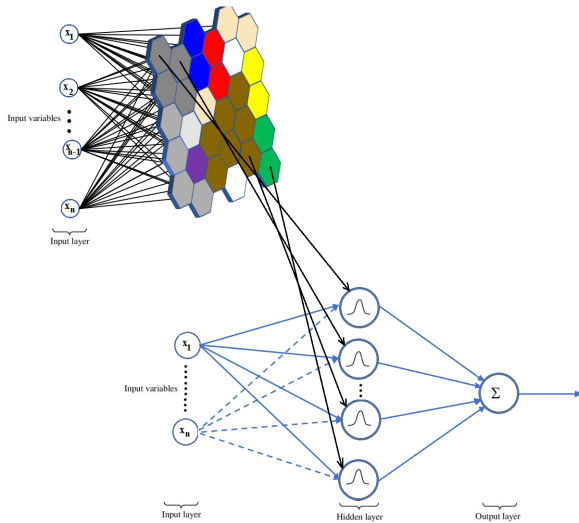
Fig. 3. The Proposed Model Process. Data are First Mapped to K Neurons/Clusters using SOM, and then Obtained Weights for each Neuron is used as the Mean Value to Calculate the Gaussian Function in RBFNN. The Standard Deviation for each Gaussian Function is also Considered as $\frac{1}{2}$x Minimum Distance of the Distances from the Corresponding Neurons in SOM to other Neurons. The Number of Neurons/Clusters in SOM/DBGSOM is equal to the Number of Radial basis Functions in RBFNN.

data set. The number of records in the NSL-KDD train and test sets is sensible. This benefit makes it possible to run experiments on the full set without randomly selecting a small portion. Consequently, the assessment findings will be consistent and comparable to various research work. Compared with KDD, the NSl-KDD dataset does not contain duplicated records in the train set, so classifiers are not biased to more frequent records. Duplicate records were also removed in NSL-KDD, and therefore, the learners' performance is not biased towards the methods with better detection rates on the repeated records. In our experiment, we use the "KD-DTrain+_20Percent" version of this dataset for training, and the "KDDTest+" data are utilized for testing. The set of training includes 25192x41 data, and the testing set contains 22544x41 data.

B. UNSW-NB15

The UNSW-NB 15 dataset's raw network packets were generated by the IXIA PerfectStorm tool at the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) to create a hybrid of real, modern normal activities and contemporary synthetic attack behaviors [18]. There are nine types of attacks in this dataset: name, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shell-code, and Worms. The Argus, Bro-IDS tools are utilized, and twelve algorithms are developed to generate a total of 49 features with the class label. A section of this data set was created as a training group (i.e., UNSW_NB15_training-set.csv) and testing set (i.e., UNSW_NB15_testing-set.csv). The training set contains 82333x43 data, and the testing set contains 175342x43 data. This partition is also used in our experiments.

C. CICIDS2017

The dataset CICIDS2017 includes benign and most up-to-date common attacks, which resemble real-world results (PCAPs). The top priority in building this dataset has been generating realistic background traffic. Data is captured in five days from Monday to Friday. The dataset for each day based on the day of the week, type of activity, and size of data is summarized as follows:

- Monday, Normal Activity, 11.0G
- Tuesday, attacks + Normal Activity, 11G
- Wednesday, attacks + Normal Activity, 13G
- Thursday, attacks + Normal Activity, 7.8G
- Friday, attacks + Normal Activity, 8.3G

Because of the hardware limitation, we only used the "Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.xlsx" from the MachineLearningCSV version. This data only contains DDoS attack and normal attack. It has 225745 activities, and each activity has 78 features (i.e. 225745x 78). Among these data, 97718 data is normal traffic, and the rest (i.e. 128027) is attack data. We randomly chose .2 data from normal data (i.e. 19544x78) and the same number of data from attack data for training, and we use the rest of the data for testing. Therefore, we have 39088x78 data for training and 186657x78 data for testing.

VI. EXPERIMENT SETTINGS AND RESULT

A. Data Pre-processing

Intrusion datasets contain different types of values. The information regarding each scenario (normal or attack) can be both numerical values and categorical values. For categorical values, datasets use text values. For example, each record in NSL-KDD dataset looks like the following data:
"0 tcp ftp_data SF 491 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 0 0 0 0 1 0 0 150 25 0.17 0.03 0.17 0 0 0 0.05 0 normal".

In this dataset, four columns which are related to tcp, ftp_data, SF, and $normal$ contain categorical values. In contrast, the other columns contain numerical values (we should note the last column lists the label for each record by indicating whether it is a $normal$ condition or an $attack$ one). To use the categorical features of these columns in our machine learning, we set numerical value for each category. For example, the second column $Protocol\_type$ contains three categories including tcp, $icmp$, and udp which are respectively set to 1, 2, and 3. We normalize the numerical values in each column by subtracting the mean value of that column and dividing it by the respective standard deviation. However, we notice in all datasets that $most$ of the values in each column are similar to each other, and some values might vary much compared to the rest. Based on this observation, we decide to use each column's median value, which relates to our dataset more closely instead of the mean value.

B. Experimental Results

In this section, we evaluate our anomaly-based intrusion detection method using the three publicly available datasets:
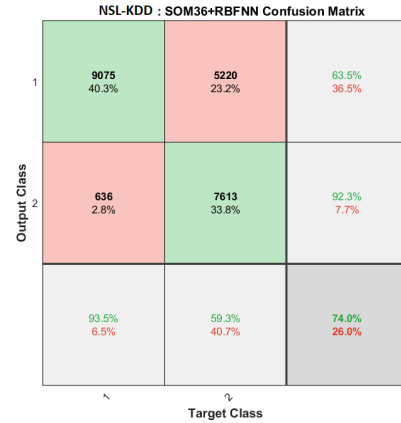
NSL-KDD, UNSW_NB15, and a subset of CICIDS2017 datasets. We have conducted our experiments using a Windows laptop with Core i7 CPU 2.70GHz and 16 GB RAM. We applied the proposed model, and (SOM36&RBFNN) **SOM was implemented with 36 pre-defined clusters followed by RBFNN**. The methods are implemented in MATLAB. We used the built-in MATLAB function for SOM (i.e., the $selforgmap$ function to create the map and the train function to train the SOM map), DBGSOM MATLAB codes are provided by the authors [10], and we implemented the RBFNN method. Table I, Table II, and Table III summarize results for the three datasets. Our hybrid model outperforms (SOM36&RBFNN) in all three cases. The row related to $DBGSOM\_Number\_Of\_Clusters$ in each table indicates the number of clusters (i.e. the final number of neurons) in each run of the proposed model. Since some neurons do not contain any data or they might have very few data, we remove any cluster which contains less than 0.001 of data. While this only removes unnecessary neurons, it speeds up RBFNN as it significantly reduces the number of radial basis functions. In each table, the $DBGSOM\_Final\_Number\_Of\_Clusters$ row indicates the final number of DBGSOM clusters after removing these neurons. Clearly, for all the three datasets, our hybrid model surpasses its (SOM&RBFNN) peer. Also, for each dataset, we included a confusion matrix of one run of the method in Fig. 4, 5 and 6. Confusion matrices for the three datasets also verify that the proposed model outperforms the (SOM&RBFNN) model.
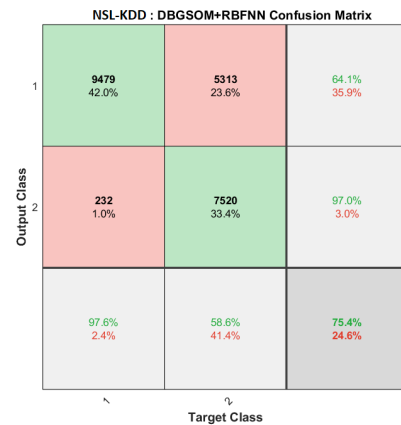
C. Network Traffic Data Visualization

As it was mentioned before, one advantage of SOMs and consequently DBGSOMs is that they can visualize high dimensional data in two dimensions. In Fig. 7, 8 and 9, we illustrate the 2D visualization obtained by SOM and DBGSOM for the three examined datasets. These figures, which are the hit maps (where each data hit on the SOM map), are useful in showing how intrusion or traffic data are different from while connected to each others.

D. Computational Runtime

To compare the CPU time consumed by the DBG-SOM+RBFNN method with the SOM+RBFNN, we need to compare the CPU time by SOM with DBGSOM the RBFNN method would have the same calculations when SOM employed with the same number of neurons that DBGSOM ends with. For the sake of this experiment, we fixed the number of epochs for SOM, DBGSOM, and RBFNN to 100. Table IV summarizes the average runtime for applying DBGSOM and SOM on the three datasets. The first column indicates the name of the dataset, the second column shows the number of neurons found by DBGSOM, which is also used by SOM, and the third and fourth columns list the CPU time by DBGSOM and SOM in seconds. From the table, we can see that SOM is faster than DBGSOM because DBGSOM has spent more CPU time than conventional SOM due to the step of inserting neurons, which involves the correct growth position and the weight initialization but has the benefit of conserving data topology on the map [10]. We also included the average runtime of implementing RBFNN as the next step after employing either SOM or DBGSOM in the fifth column. Since applying both

(a) SOM36+RBFNN

(b) DBGSOM+RBFNN

Fig. 4. Confusion Matrix for the NSL-KDD Dataset obtained by running the (SOM&RBFNN) and the Proposed Hybrid Models

our hybrid and the (SOM&RBFNN) models to train a model is offline, we may be interested in knowing the CPU time for testing the model on test data. Therefore, we calculate the CPU time that our system needed to evaluate all the test data, classify them as either normal or attack, and report the result in the sixth column. The next column indicates the average runtime our system needed to evaluate each data. The CPU is taken by all test images divided by the number of test images. The CPU time for testing one record of data for all three datasets is very small. This indicates the proposed method can be employed in a real-time application and would even be speeded up by using high-config systems. The last two columns of the table show the dimension of our training data and testing data. The number of neurons found by DBGSOM indicates why the CPU time for testing one record of data in each dataset is different.

It is significant to highlight the fact that SOM's batch learning principles can save training time dramatically. Besides, the procedure of growing and the ultimate grid structure are independent from the arrangement of the input vector presentation.

TABLE I. CLASSIFICATION RATE FOR THE COMPARED METHODS ON THE NSL_KDD DATASET IN 5 RUNS. DATA ARE NORMALIZED; SF VALUE FOR DBGSOM IS .01. AFTER DBGSOM COMPUTES CENTERS, CLASSES WITH LESS THAN .001 OF DATA ARE REMOVED.

Methods	Run#1	Run#2	Run#3	Run#4	Run#5	Average
SOM36+RBFNN	74.31	74.94	74.96	77.72	75.74	75.53
Proposed hybrid model	75.27	76.60	75.71	75.60	76.66	75.97
DBGSOM_Number_Of_Clusters	263	294	299	269	259	276.8
DBGSOM_Final_Number_Of_Clusters	166	191	181	175	165	175.6

TABLE II. CLASSIFICATION RATE FOR THE COMPARED METHODS ON UNSW_NB15 DATA SET IN 5 RUNS. DATA ARE NORMALIZED, SF VALUE FOR DBGSOM IS .01. AFTER DBGSOM COMPUTES CENTERS, CLASSES WITH LESS THAN .001 OF DATA ARE REMOVED.

Methods	Run#1	Run#2	Run#3	Run#4	Run#5	Average
SOM36+RBFNN	75.69	74.71	82.38	81.47	76.63	78.18
Proposed hybrid model	91.57	91.47	85.72	90.50	85.85	89.02
DBGSOM_Number_Of_Clusters	652	698	658	626	638	654.40
DBGSOM_Final_Number_Of_Clusters	339	373	354	353	366	357.00

TABLE III. CLASSIFICATION RATE FOR THE COMPARED METHODS ON THE SUBSET OF THE CICIDS 2017 DATASET (I.E., FRIDAY-WORKINGHOURS-AFTERNOON-DDOS.PCAP_ISCX) IN 5 RUNS. DATA ARE NORMALIZED; SF VALUE FOR DBGSOM IS .01. AFTER DBGSOM COMPUTES CENTERS, CLASSES WITH LESS THAN .001 OF DATA ARE REMOVED.

Methods	Run#1	Run#2	Run#3	Run#4	Run#5	Average
SOM36+RBFNN	98.86	98.82	97.73	98.45	88.44	96.46
Proposed hybrid model	99.27	99.17	99.70	99.22	97.39	98.95
DBGSOM_Number_Of_Clusters	277	245	277	302	293	279
DBGSOM_Final_Number_Of_Clusters	153	150	152	170	165	159

TABLE IV. THE AVERAGE RUNTIME FOR APPLYING DBGSOM AND SOM ON THE THREE DATASETS.

Dataset	Neurons	DBGSOM	SOM	RBFNN	Testing on all test image set	Testing per test image	Training Dimension	Testing Dimension
NSL-KDD	304	252.5041	173.3397	432.4043	3.992026	0.0191805	25192x41	22544x41
UNSW-NB15	833	2260.498	1635.984	3942.145	86.45262	0.0224827	82332x42	175341x42
CICIDS2017	324	761.0038	466.5995	1338.836	67.031675	0.0071727	39088x78	186657x78

With this technique, in each training cycle, the training vectors are offered to the network once, and neurons cumulative fault calculated immediately after the step of weight adaptation. Thus, the neurons network has the opportunity for growth from more than one boundary node which causes difficulty in managing the growth process that leads to dead neurons, which are that don't represent any input vector at the end of the training, and unexpected growth of the network and therefore raises the computational cost. Our hybrid model presents many rules for network growth and that fills just one position around each boundary neuron. Because of the narrow neighborhood function in GSOM, allocation weight vectors to recent neurons have a dramatic influence on map tangling and twisting and must be regarded for serving the quality and smoothness of the network. In DBGSOM, the cumulative error for detecting eligible boundary neurons is not only considered but also for assigning appropriate weight vectors and network location to new neurons.

VII. CONCLUSION

We propose to apply DBGSOM together with RBFNN for detecting anomaly-based intrusion in the network. DBGSOM employs a batch learning strategy for GSOMs to resolve the issues of SOM and GSOM models. It has a better ability to conservation topology. It reduces exposure for twisting and tangling while offers suitable mechanisms to discover a proper growing position and designating initial weight vectors for the new neurons. RBFNNs, on the other hand, are fast and comprehensive in training, and yields high precision in detecting intrusion. We implemented the DGBSOM+RBFNN method in

MATLAB and applied it on three publicly available datasets: NSL-KDD, UNSW-NB15, and CICIDS2017. Our extensive experiment indicates that the proposed method outperforms the SOM+RBFNN method for anomaly-based intrusion detection.

Future works intend to integrate the proposed model with a novel regularization technique that utilizes the standard deviation for classifying and detecting intrusions efficiency. We will incorporate the regularization technique to discourage learning from complex model. As a result, we expect more generalization from the machine learning model to accurately perform on unseen data.
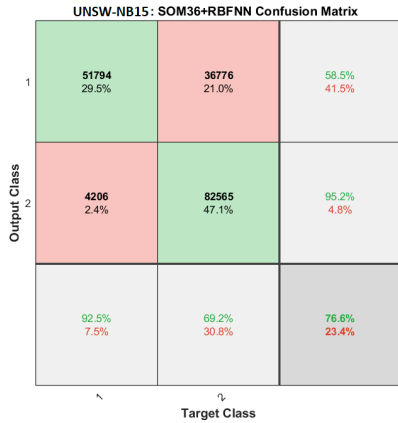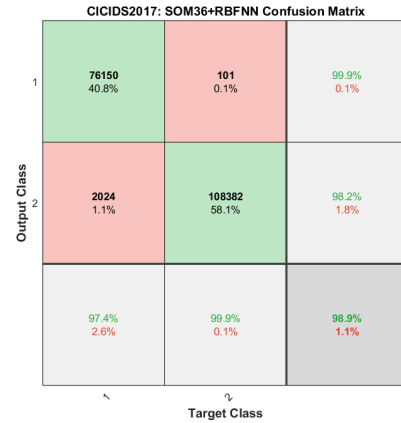
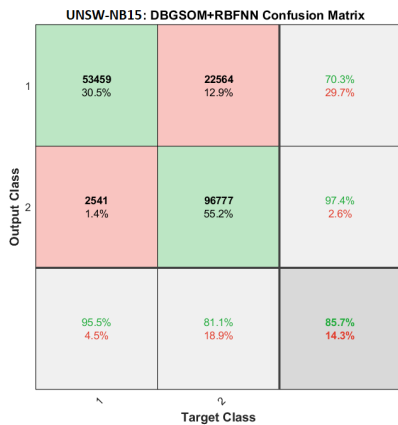**Conflicts of interest**:The authors declare that there is no conflict of interest.

REFERENCES

[1] S. A. Althubiti, E. M. Jones, and K. Roy, "LSTM for Anomaly-Based Network Intrusion Detection," in 2018 28th International Telecommunication Networks and Applications Conference (ITNAC), 2018.

[2] S. AlHamouz and A. Abu-Shareha, "Hybrid Classification Approach Using Self-Organizing Map and Back Propagation Artificial Neural Networks for Intrusion Detection," in 2017 10th International Conference on Developments in eSystems Engineering (DeSE), 2017.

[3] https://www.dnsstuff.com/intrusion-detection-system

(a) SOM36+RBFNN

(b) DBGSOM+RBFNN

Fig. 5. Confusion Matrix for the UNSW-NB15 Dataset obtained by running the (SOM36&RBFNN) and the Proposed Hybrid Models.

(a) SOM36+RBFNN

(b) DBGSOM+RBFNN

Fig. 6. Confusion Matrix for the CICIDS2017 Dataset obtained by running the SOM36+RBFNN and Proposed Hybrid Models.

[4]   T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, "Deep Recurrent Neural Network for Intrusion Detection in SDN-based Networks," in 2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft), 2018.

[5]   R. C. Staudemeyer, "Applying long short-term memory recurrent neural networks to intrusion detection," South Afr. Comput. J., vol. 56, no. 1, pp. 136–154, 2015

[6]   S. Albayrak, C. Scheel, D. Milosevic, and A. Muller, "Combining Self-Organizing Map Algorithms for Robust and Scalable Intrusion Detection," in International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06).

[7]   S. Mohammadi and F. Amiri, "An Efficient Hybrid Self-Learning Intrusion Detection System Based on Neural Networks," International Journal of Computational Intelligence and Applications, vol. 18, no. 1, p. 1950001, Mar. 2019.

[8]   Y. Guo, W. Liao, Q. Wang, L. Yu, T. Ji, and P. Li, "Multidimensional time series anomaly detection: A gru-based gaussian mixture variational autoencoder approach," in Proceedings of the 10th Asian Conference on Machine Learning (ACML18), Beijing, China, Nov. 14-16, 2018.

[9]   J. An and S. Cho. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. Technical Report, SNU Data Mining Center, 2015. http: //dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf

[10] M. Vasighi and H. Amini, "A directed batch growing approach to enhance the topology preservation of self-organizing map," Applied Soft Computing, vol. 55, pp. 424–435, Jun. 2017

[11] R. Vinayakumar, K. P. Soman and P. Poornachandran, "Applying convolutional neural network for network intrusion detection," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, 2017, pp. 1222-1228.

[12] Z. Li, Z. Qin, K. Huang, X. Yang, and S. Ye, "Intrusion Detection Using Convolutional Neural Networks for Representation Learning," in Neural Information Processing, Springer International Publishing, 2017, pp. 858–866.
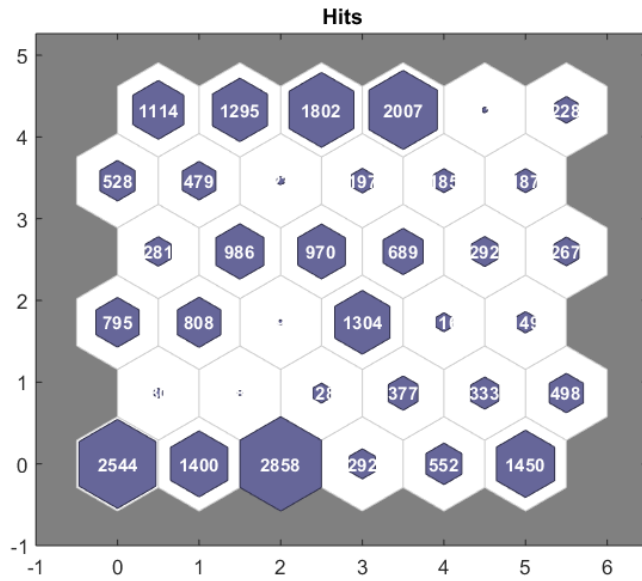
[13] J. Yang, T. Li, G. Liang, W. He, and Y. Zhao, "A Simple Recurrent Unit Model Based Intrusion Detection System With DCGAN," IEEE Access, vol. 7, pp. 83286–83296, 2019.

[14] A. AbuGhazleh, M. Almiani, B. Magableh, and A. Razaque, "Intelligent Intrusion Detection Using Radial Basis Function Neural Network," in 2019 Sixth International Conference on Software Defined Systems (SDS), 2019
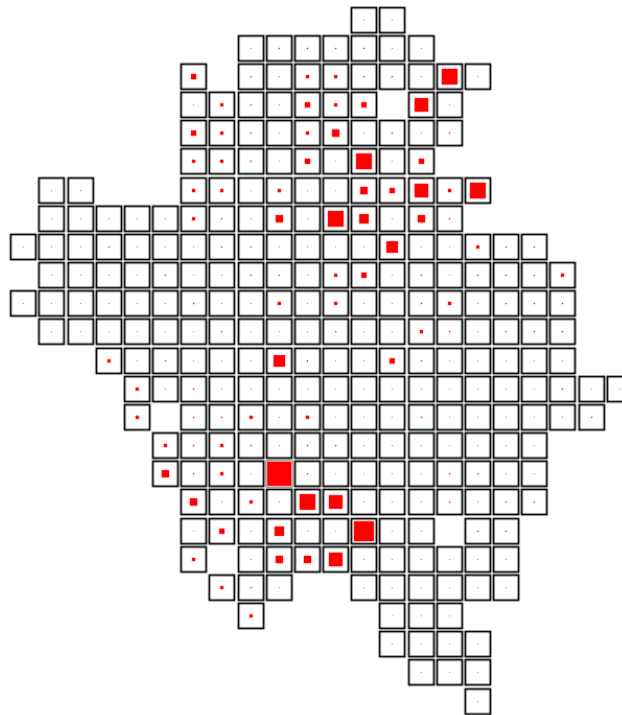
[15] D. Alahakoon ; S.K. Halgamuge ; B. Srinivasan, "Dynamic self-organizing maps with controlled growth for knowledge discovery", IEEE Transactions on Neural Networks, Volume: 11, Issue: 3, pp.

601-614, 2000

[16] Q. Zhang and F. Wilson, RBNN Application and Simulation in Big Data Set Classification, Journal of Intelligent& Fuzzy Systems (JIFS), pp. 1-9, 2019.

[17] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[18] N. Moustafa, "Designing an online and reliable statistical anomaly detection framework for dealing with large high-speed network traffic." Ph.D. dissertation, University of New South Wales, Canberra, Australia, 2017.
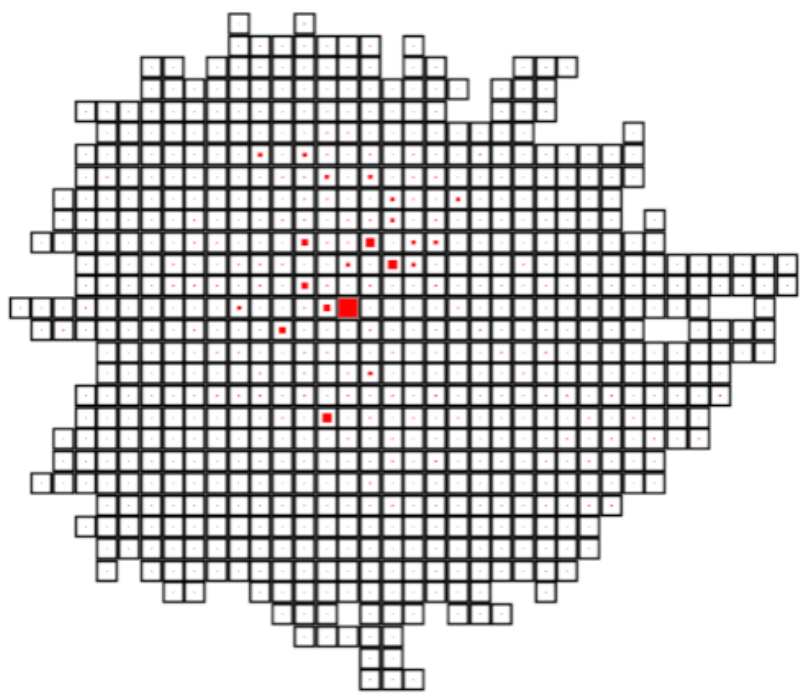
(a) SOM36

(b) DBGSOM

Fig. 7. Hit Maps from the NSL-KDD Data Set obtained by (a) SOM with 36 Neurons, (b) DBGSOM with 310 Neurons. For SOM, each Cell Number shows the Number of Data assigned to those Cell/Neurons and for DBGSOM, the Number of Data assigned to each Cell/Neuron is Proportional to how much that Cell is Colored.
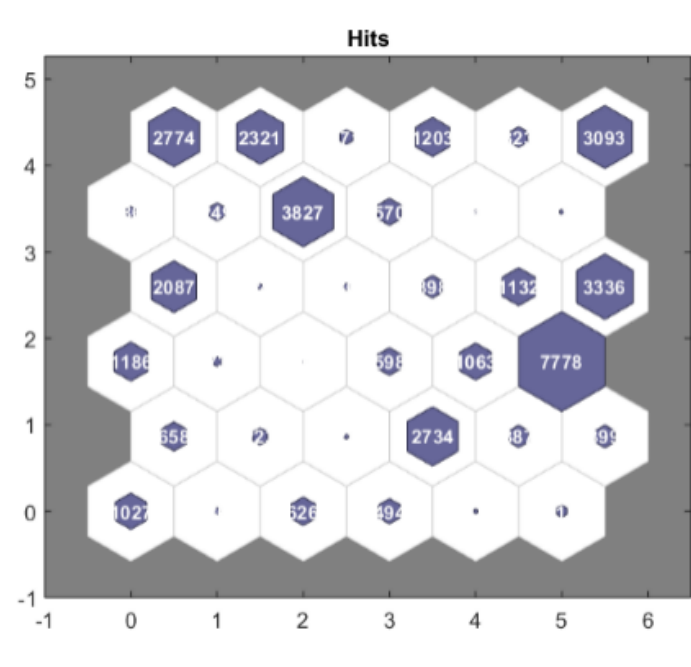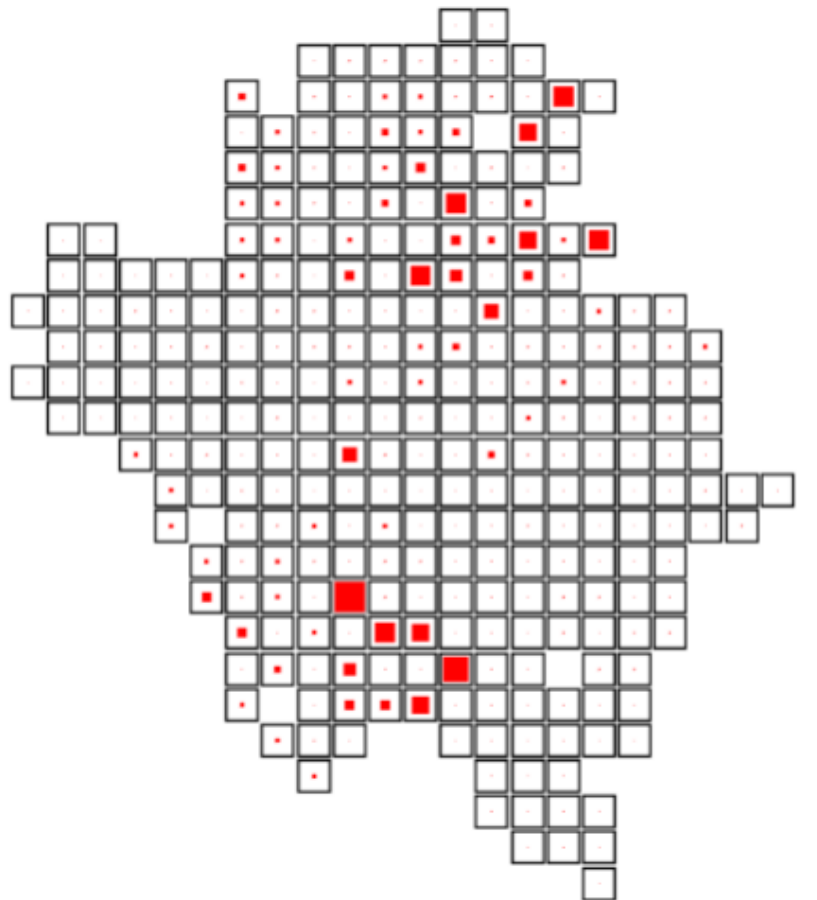
(a) SOM36

(b) DBGSOM

Fig. 8. Hit Maps from the UNSW-NB15 Data Set obtained by (a) SOM with 36 Neurons, (b) DBGSOM with 310 Neurons. For SOM, each Cell Number shows the Number of Data assigned to those Cell/Neurons and for DBGSOM, the Number of Data assigned to each Cell/Neuron is Proportional to how much that Cell is Colored.

(a) SOM36

(b) DBGSOM

Fig. 9. Hit Maps from the CICIDS2017 Data Set obtained by (a) SOM with 36 Neurons, (b) DBGSOM with 310 Neurons. For SOM, each Cell Number shows the Number of Data assigned to those Cell/Neurons and for DBGSOM, the Number of Data assigned to each Cell/Neuron is Proportional to how much that Cell is Colored.