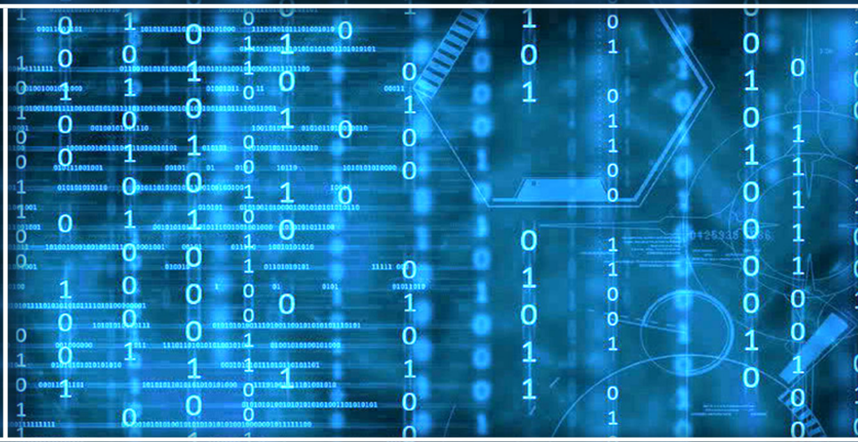
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 9 Issue 11

November 2018



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

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 9 Issue 11 November 2018
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

Exploring Identifiers of Research Articles Related to Food and Disease using Artificial Intelligence

Marco Ross[1], El Sayed Mahmoud[2]
Faculty of Applied Science & Technology
Sheridan College
Oakville, Canada

El-Sayed M. Abdel-Aal[3]
Guelph Research and Development Centre
Agriculture and Agri-Food Canada
Guelph, Canada

***Abstract*—Currently hundreds of studies in the literature have shown the link between food and reducing the risk of chronic diseases. This study investigates the use of natural language processing and artificial intelligence techniques in developing a classifier that is able to identify, extract and analyze food-health articles automatically. In particular, this research focuses on automatic identification of health articles pertinent to roles of food in lowering the risk of cardiovascular disease, type-2 diabetes and cancer. Three hundred food-health articles on this topic were analyzed to help identify a unique key (Identifier) for each set of publications. These keys were employed to construct a classifier that is capable of performing online search for identifying and extracting scientific articles in request. The classifier showed promising results to perform automatic analysis of food-health articles which in turn would help food professionals and researchers to carry out efficient literature search and analysis in a timely fashion.**

Keywords—Natural language processing; text classification; n-grams; bioinformatics; knowledge extraction; nutrition assessment; health promotion; research uptake

I. Introduction

Health professionals in Canada rarely use the results of medical research to promote health and influence policy. This has been shown in a 2007 survey of Canadian health professionals, based on the answers of 928 professionals and managers from Canadian health service organizations. The survey results showed that fifty seven percent of the respondents frequently or very frequently received research results. These received results never or rarely influenced the health professionals' decisions and choices in fourteen percent of the cases. Additionally, they were also never or rarely transformed into concrete applications in another eleven and half percent of the cases [1].

The main reason for these low uptake percentages could be attributed to the outright volume of medical research related to food-health being produced on a regular basis. Large numbers of scientific publications make the selection process of an article about a specific food and disease more difficult. For example, the popular biomedical database MEDLINE, produced by the United States National Library of Medicine, contains over 24 million references of biomedical texts alone [2]. Since 2005, the literature has seen consistent from 16 million articles to 24 million [3], marking a continued increase of approximately 1,800 articles per day from 2005 until present day.

The large number of articles increases the difficulty of finding appropriate papers for a given topic on food and health. It also delays extracting useful information from these articles and perhaps this useful information could be lost in the search process. Such information includes roles of food in health and nutrition, food intake, recommended foods for disease prevention, food protection mechanisms, etc. This information is essential for food or health policies, and food health, nutrient or function claim and food labelling. This research was intended to build an automatic article classifier in the area of roles of food in reducing the risk of three choric diseases: CVD, type-II diabetes and cancer. The study goal was to develop a text classification tool that is capable of performing efficient literature search and analysis in a timely fashion.

A. Motivation

Improving health promotion and disease prevention through diets with the use of artificial intelligence techniques is the main motivation of this research. The use of artificial intelligence techniques in performing literature search and analysis should improve its efficiency. Currently, the link between diet and health promotion and disease prevention is well established with numerous amounts of publications. This requires techniques and tools to extract and analyze data. The current research should make a difference in the way we manage data, develop strategies and conduct research. The stakeholders of health promotion such as health professionals, dieticians, policy makers, and researchers should benefit from this tool.

B. Organization of Paper

The remainder of this paper consists of a literature review, methodology and results.

The literature review focuses on prior research conducted in the fields of text mining, food-healthiness knowledge extraction, and natural language processing (NLP). It will examine the recent literature in these areas as well as full scale surveys and reviews of the general field of food-health and NLP.

The methodology section describes the details of methodologies involved in the work. This includes selecting the training data, identifying n-gram sequence sizes, building the article classifier, testing the classifier and performance metrics used.

The results section will highlight the experimental findings including the analysis of these findings and potential future research.

II. Literature Review

The explosive growth of food-health literature has prompted increasing interest in using text mining techniques to address the information overload faced by domain experts. This is reflected by the conception of articles reviewing this work [4] [5], which target experts in biosciences as their primary audience [6].

The recent proliferation of articles reviewing using text mining for medical applications includes electronic medical records knowledge extraction, epidemic detection through semantic analysis of social media, abbreviations in biomedical text, automatic terminology management in biomedicine, as well as automatic scientific literature analysis as a tool for novel findings and hypotheses from research [6] [7] [8]. The category of automatic scientific literature analysis as a tool for novel findings from research is of most relevance to this work and will be the focus of this study. Automated analysis of scientific literature complements the reading of scientific literature by individual researchers because it allows quick access to information contained in large volumes of documents [7]. Hirschman and others hypothesize that in the future, it is likely that solutions will be developed that produce and test hypotheses against knowledge bases. This type of solution development in the field of bioinformatics relies heavily on researchers having rapid access to a large corpus of literature readily available which may be automatically analyzed and interpreted [7].

A current survey of work in biomedical text mining conducted by Cohen and Hersh in 2005 hypothesized that the biggest challenge to biomedical text mining in the coming 5-10 years would be building systems which are useful to researchers [6]. A literature review by Rebholz-Schuhmann et al. builds on this hypothesis by suggesting future work in this field should be focused on helping researchers in problem solving of specific real-world s. **Figure 1** contains a modified version of a diagram made by Rebholz-Schuhmann et al. which shows the different categories where text mining can help scientific researchers, using food-health relationships as an example [7].

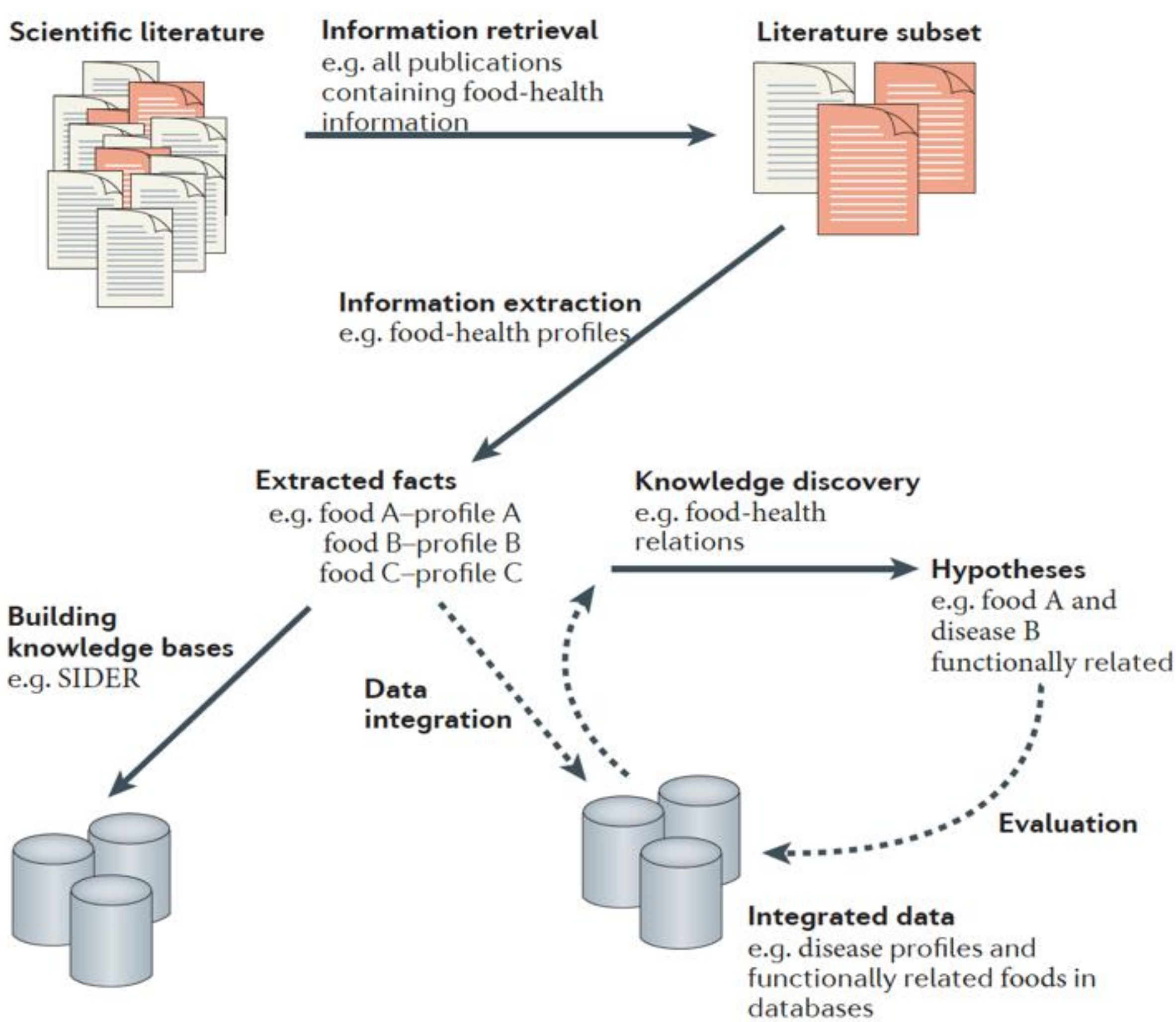


Fig. 1. Categories of text mining solutions.

Figure 1 distinguishes four primary stages in text-mining solutions: information retrieval, information extraction, building knowledge bases and knowledge discovery [8].

Information retrieval could involve a user submitting a query to a search engine and receiving a document fitting to their submitted query in return. Information extraction involves the identification of entities, such as diseases or foods, as well as the identification of complex relationships between these entities [8]. Scientific facts extracted from literature may be used for the purposes of populating databases or data curation. From these extractions, knowledge bases can be built that contain the collected statements together with collected evidence in the form of references to the literature [7] [8]. Knowledge discovery involves identifying undiscovered or hidden knowledge by applying data-mining algorithms to the collection of facts gathered from the literature. From here, text-mining results may be used to suggest new hypotheses automatically which can be used to either validate or disprove existing hypotheses or to help direct future research [8].

This work automatically identifies whether a given medical article is related to food and CVD, type-II diabetes, or cancer, and therefore the category of text mining is most similar to information extraction. This research assumes that databases already exist from which users can query. The ultimate goal of this work was to develop a tool that is able to automatically identify food-health articles relevant to the three proposed diseases which facilitates extracting useful information from medical literature and building knowledge bases.

N-grams has been used for finding identifiers of disease outbreaks in reasons of entering the emergency department room [9] and fining identifiers for customer intent [10]. This work investigates how to use N-grams for finding identifiers for food-health articles related to the proposed diseases.

III. Methodology

A. Classifier

The steps for building the proposed classifier include: (1) creating n-gram lists with various n-gram sizes for each disease category of CVD, type-II diabetes, and cancer (2) determining the effective list of most frequent n-grams in each category (3) identifying the effective n-gram size for detecting the subject of an article. This process is illustrated in **Figure 2.**

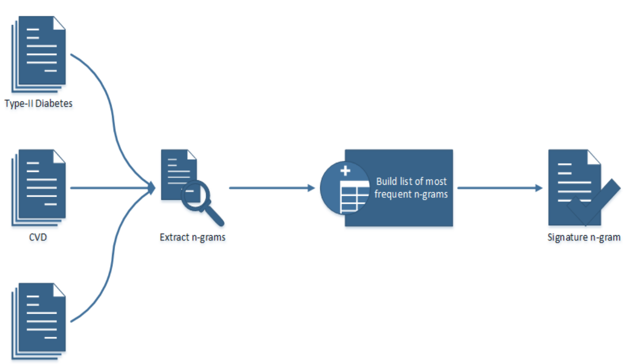


Fig. 2. Steps to identify a unique key for food-health articles related to the diseases: CVD, Cancer and Diabetes.

For the sake of simplicity, the diagram in **Figure 2** shows the process as if one n-gram classifier will be created from all three disease categories, when in fact three *separate* classifiers will be created using this same process.

B. Creating a Classifier

In order to create an accurate food-health article classifier for each disease category, one-hundred peer-reviewed articles which relate to a certain food and that specific disease will be manually selected for each of CVD, type-II diabetes, and cancer, resulting in a total of 300 unique articles. One-hundred articles have been determined to be an appropriate sample size according to [11].

After one-hundred articles are gathered from each of the respective diseases, an n-gram algorithm is applied to the articles in order to extract the n-gram lists from them, thus providing the building blocks of an identifier to be refined in the next steps.

C. Determining Most Frequent N-Grams

After n-grams have been gathered from each of the 4 chosen sequences sizes of n-grams (n=1, n=2, n=3, n=4), the most frequently n-grams in each of the respective articles are used as an identifier for the food-health articles related to that disease. The amount of the commonly found n-grams are determined experimentally as it is not immediately obvious how many unique n-grams will be found, nor is it obvious how many of the top most commonly found ones will be enough to accurately build the classifier. There will likely be hundreds, possibly thousands of unique n-grams and thus the most appropriate allocation of the most commonly found n-grams will be determined once the n-grams have been generated and analyzed experimentally.

D. Determining Best N-Gram Size

The most effective sequence size of the n-grams is determined experimentally. Once again, it is not immediately apparent which size will be the most accurate in classifying a food-health article. n-gram list of larger sequence sizes (e.g. n=5) provide more coherent phrases in natural language, yet they are very specific and unlikely to be commonly found throughout the sample of articles we will use. Likewise, n-grams of much smaller sizes (e.g. n=1) may not be specific enough to differentiate between a food article related to CVD and a food article related to type-II diabetes.

E. Testing Strategy

Each classifier for each disease is tested by using manually selected food-health articles which have not been presented to the algorithm. We used the 70/30 split which is the de facto standard for training and testing machine learning algorithms as seen in [12]. This means that 70% of the data are used to train the algorithm, while 30% are used towards testing it. The 30% that have been used to test are articles which are hidden from the algorithm. If the algorithm is able to correctly classify the articles after training, then it will be considered a success.

F. Performance Metrics

The performance metrics which are used to determine the relevance and accuracy of the algorithm are precision and recall. Using true positive, true negative, false positive, and false negative, determines the accuracy of the classifier. When it receives a medical article as input, does it correctly classify the article or not? That is the only performance metric which will be required in order to determine its accuracy.

G. Data

The data used for this study are manually gathered, peer-reviewed medical articles which specifically discuss health outcomes related to certain types of foods as their subject matter. The articles have been gathered manually because the nature of the data required is very specific and therefore not readily available in large quantities of word corpora such as social media, for example.

The sources of the data are popular medical databases such as PubMed/MEDLINE and Cochrane Library, as well as multidisciplinary scholarly databases such as ScienceDirect, Web of Science, JSTOR, and Google Scholar. These databases contained articles from popular medical journals including the New England Journal of Medicine (NEJM), British Medical Journal (BMJ), JAMA Network, American Diabetes Association, the American Journal of Clinical Nutrition, American Medical Association (AMA), Ovid Lippincott Williams and Wilkins (OLWW), and more. The portals and databases which we accessed these articles through can be found in **Figure 3** below, showing a graphical distribution of the online sources used for gathering the training and testing data.

The data used for both training and testing are pre-processed before being used in the final implementation of the algorithm.

The first stage of preprocessing is converting the medical articles from PDF format to plaintext format, using UTF-8 encoding. The articles are normally retrieved in PDF format, so in order to facilitate extracting n-grams from them, we converted the articles to plain text. An existing Python package 'pdf2txt' which uses another Python package 'pdfminer' is used to batch convert the PDFs to plain text.

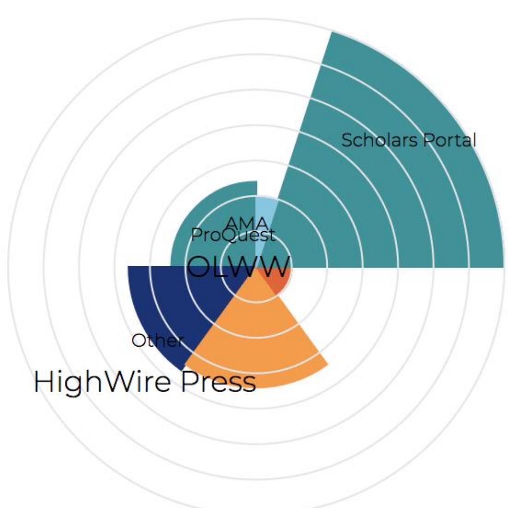


Fig. 3. Data source distribution.

Next, the second stage of preprocessing is normalization which involves tokenizing the text, converting the entire body of text into lowercase, removing non-alphabet characters. in addition, we removed 266 stop words which include unremarkable words such as 'aren't', 'the', 'a', 'as', and 'because', as well as words which were found repeatedly in the articles such as 'journal', 'clinical', 'research', and 'published'.

Mean Magnitude of Relative Error (MMRE) is one of the most widely used evaluation criterion for assessing the performance of software prediction models [13]. This method involves using the *estimated effort* required to develop a software less the *actual effort* to create the software, divided by the actual effort. It is quite similar to the formula for calculating precision in the field of information retrieval. MMRE differs from the method of accuracy determination used in this research primarily due to its application. MMRE is more applicable to determining the accuracy of software estimation when considering man hours and money required to build a software system, whereas precision is more applicable to accuracy of retrieving documents based on a condition.

IV. Results and Analysis

The most frequent Bigrams extracted from food-health articles are three unique identifiers that can be used effectively to enable the automatic identification and classification of the food-health articles related to the three diseases. The *n-gram size* (n=2) and the *length* of the n-grams list (l=800) have been found to be more effective in identifying food-health articles related to any of the three diseases compared to unigrams, trigrams and quadgrams for various n-gram-list lengths.

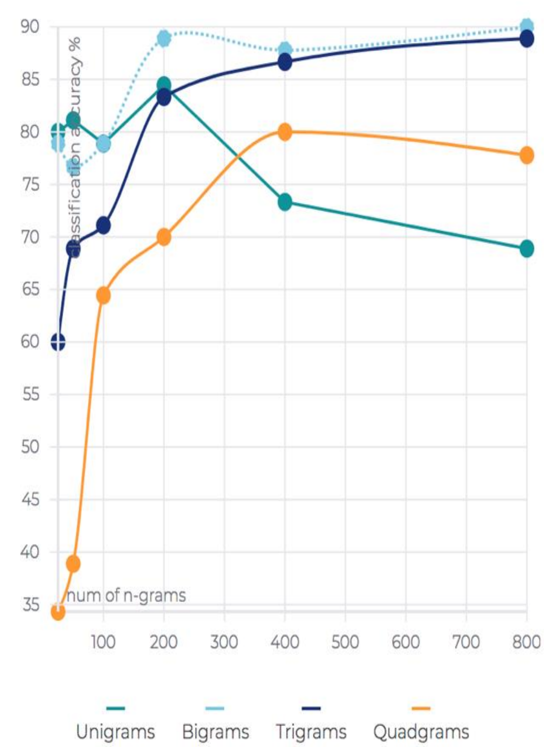


Fig. 4. Overall classification accuracy.

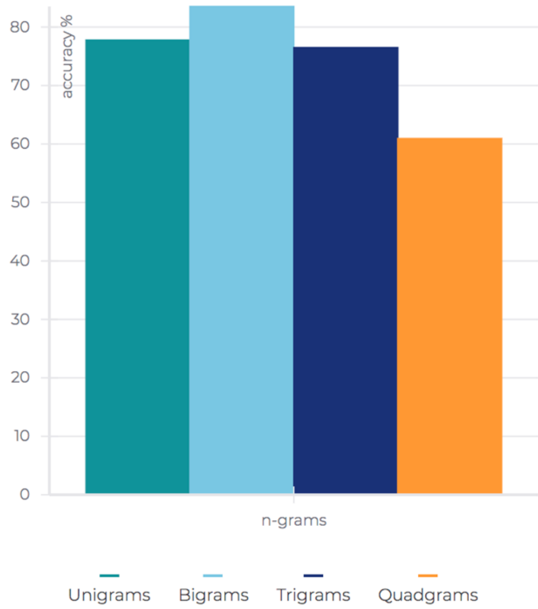


Fig. 5. Average accuracy for various n-gram sizes.

The effectiveness of the most frequent 800 bigrams have been tested by using them as an identifier for developing a food-health article classifier. The highest accuracy of the classifier is 90.00%. The overall average accuracy of all the various combinations for the value of n and number of n's are shown in **Figure 4**. This graph depicts the varying degrees of accuracy resulting from different combinations of these two variables. **Figure 4** shows that most combinations of bigrams with 200 or more n-grams retrieved results with the highest average classification accuracy. Trigrams (n=3) are ~~a close~~ second in accuracy but only with using a set of 200 or more n-grams. Quadgrams (n=4) did poorly particularly with a lower value of the n-gram list length. In contrast to quadgrams, unigrams (n=1) seem to be accurate *only* with a lower n-gram list length, with an accuracy drop off once the length is increased.

When manipulating n-gram size only and not accounting for n-gram list length, bigrams win out with an average accuracy of 83.5%, while unigrams and trigrams have close similar accuracies of 77.8% and 76.5%, respectively. This is shown by **Figure 5** that compares the classifier average accuracy when isolating for n-gram size, without controlling for n-gram list length. Quadgrams showed the biggest reduction in accuracy with an average accuracy of 60.9%. This significant drop in accuracy could be attributed to the decrease in the n-gram list length due to the increase in the n-gram size. This affects the ability of the classifier to find notable differences within the articles. The writing styles of the authors could be another reason behind this drop in the accuracy when finding four words in a row being similar are hard to come by. Additionally, the subject matter of the articles all vary quite a bit even when studying the same diseases. For example, there were a few articles retrieved for the CVD portion of the data which talked about different cardiovascular diseases and their relationship to fish consumption in particular. One article talked about fish consumption and its relationship to risk of myocardial infraction (heart attacks), another spoke about fish and its relation to reduced progression of coronary artery atherosclerosis, and another talked about fish and omega-3 consumption in relation to risk of cerebrovascular disease. This simple example shows that even though an article may study CVD while also talking about fish consumption, it can take many different approaches to doing so. For this reason, quadgrams may be too generic and not as commonly found in order to be an effective method of classifying articles.

Alternatively, when isolating for n-gram list length while not controlling for n-gram size, the overall accuracy of varying degrees of the n-gram list length provides interesting results, shown in **Figure 6**. This figure shows the average classification accuracy for n-gram list length values of 25 through 800. It is interesting to note that while the accuracy gradually increases as we use a higher length value, classification accuracy plateaus after a certain point of 200 n-gram, and only varies by a tenth of a percent between length values of 200, 400 and 800. The respective accuracies of these values are 81.7%, 81.9%, and 81.4%. It is certainly a notable difference from the resulting accuracy of the length values 25 (63.3%) and 50 (66.4%). Perhaps the more notable implication from these values is that increasing the length does not result in higher classification accuracy beyond a certain point. This could be explained by examining how many n-grams are repeatedly found at the bottom of the list when looking at high lengths of n-gram list. Using the cancer training data as an example, we see that using bigrams with a length value of 800, ('breast', 'cancer') is the most frequently found n-gram with a frequency of 1222, with ('cancer', 'risk') coming second with 618 matches. By contrast, the 799^{th} and 800^{th} most commonly found bigrams are ('low', 'folate') and ('lipid, 'metabolism') with a frequency of 15 each. Additionally, the 200^{th} ('cancer', 'patients') and 400^{th} ('dietary', 'indexes') most commonly found bigrams only appear 38 and 23 times, respectively in the entire corpus of training data. This could explain why increasing the length beyond 200 does not drastically increase the classification accuracy, because the data becomes more diluted at this point and contains many more unique n-grams that are very specific to that single test article and may not necessarily be found within the training data.

Another interesting observation from the results of the test data is which individual disease topics had the highest average and highest achievable accuracies. CVD had the highest *achievable* accuracy (HAA) of 100% classification accuracy using n-gram size of 3 and a n-gram list length of 400, which can be noted as (3, 400), while cancer's HAA was 90.0% with a 4-way tie between (2, 100), (2, 400), (2, 800), (3, 800), and diabetes' HAA was 86.7% using (1, 200). This is certainly remarkable because it appears that certain combinations of n-gram size and n-gram list length result in different accuracies for each disease.

Across the 26 different combinations of n-gram size (1-4) and n-gram list length (25-800), CVD alone had the highest average accuracy of 87.4% while cancer and diabetes lagged behind with 69.3% and 69.2%, respectively. Thus, we could conclude that the CVD training data was either more unique or that the diabetes and cancer data was not unique enough. The latter seems to be the more likely, as the classifier was not able to distinguish between the test data belonging to diabetes more often than it incorrectly classified it. That is to say, when a diabetes article was not correctly recognized by the classifier as a diabetes article, it was because the test article had **an equal number of matches** from both the diabetes and CVD training data, not because it flat out incorrectly guessed the subject matter of the article. This could be because diabetes and CVD tend to have many overlapping terms and risk factors in medicine.

For the cancer testing data, the low overall accuracy could be explained through the fact that there is not very much research available linking food to cancer, and when there is, there are so many different types of cancers that these articles study including breast cancer, lung cancer, colorectal cancer, kidney cancer, and prostate cancer. This may have led to a failure to classify the articles correctly on a consistent basis due to the training data being so diverse.

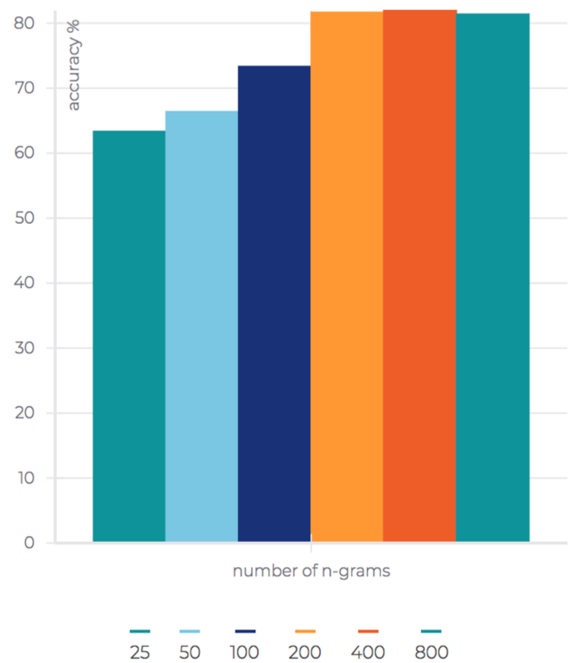


Fig. 6. Average accuracy for various n-gram list lengths.

V. Conclusion

This research is the first report to describe the use of natural language processing and artificial intelligence techniques to extract and analyze data from literature via an automatic classifier. The developed classifier could change the way we manage data, develop strategies and conduct research. The classifier tool would be useful for a broad range of stakeholders including health professionals, dieticians, policy makers and researchers. More research is underway to further develop this classifier into one that is able to find trends in food and health, in order to develop novel hypotheses and support existing ones. Additionally, some features will be built in to filter articles on the basis of inclusion/exclusion criteria provided by authorities.

VI. Next Steps

The established identifiers are the fundamental step of the automatic extraction of useful information from the food-health articles related to specific diseases. The next steps will focus on analysis and mining the contents of the identified articles for specific disease. Data warehousing, big data techniques will be investigated to store and organize the extracted data in multidimensional databases. These databases could be used by food or nutrition researchers and other stakeholders to identify research gaps and to guide future strategies in food and health for both private and public sector.

Acknowledgment

This work is resulted from Marco Ross's undergraduate thesis of the Honours Bachelor of Computer Science (Mobile Computing). This research was supported by the Centre for Mobile Innovation (CMI) – Sheridan College, ON, Canada, through the work study program. We thank the thesis advisory committee members at Sheridan College for their useful feedback on the work.

References

[1] O. Belkhodia, N. Amara, R. Landry and M. Ouimet, "The Extent and Organizational Determinants of Research Utilization in Canadian Health Services Organizations," *Science Communication,* vol. 28, no. 3, pp. 377-417.

[2] "MEDLINE Fact Sheet," U.S. National Library of Medicine, 2018. [Online]. Available: https://www.nlm.nih.gov/pubs/factsheets/medline.html. [Accessed 5 April 2018].

[3] L. Hunter and K. Cohen, "Biomedical Language Processing: What's Beyond PubMed?," *Molecular Cell,* vol. 21, no. 5, pp. 589-594, 2006.

[4] "Reviews on Text Mining in Biomedicine," Biomedical Literature and Text Mining Publications, 2006. [Online]. Available: http://blimp.cs.queensu.ca/cateR_1.html. [Accessed 11 April 2018].

[5] S. Ananiadou and J. McNaught, "Text mining for biology and biomedicine," *Scitech Book News,* p. 286, 2006.

[6] A. Cohen and W. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics,* vol. 6, no. 1, pp. 57-71, 2005.

[7] L. Hirschman, G. A. P. C Burns, M. Krallinger, C. Arighi, K. Bretonnel Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala, A. Lourenço, R. Nash, A.-L. Veuthey and T. Wiegers, "Text mining for the biocuration workflow," *Database: The Journal of Biological Databases and Curation,* vol. 1, 2012.

[8] D. Rebholz-Schuhmann, A. Oellrich and R. Hoehndorf, "Text-mining solutions for biomedical research: enabling integrative biology," *Nature Reviews.Genetics,* vol. 13, no. 12, pp. 829-839, 2012.

[9] E. S. Mahmoud and S. Deborah, "Identifying Syndromic Fingerprints in Reason Fields in Emergency Department or Telehealth Records using N-grams for Similarity Analysis," *Advances in Disease Surveillance,* vol. 4, p. 55, 2007.

[10] S. Akulick and E. S. Mahmoud, "Intent detection through text mining and analysis," in *Future Technologies Conference (FTC),* Vancouver, 2017.

[11] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft and J. Popp, "Sample size planning for classification models," *Analytica Chimica Acta,* vol. 760, pp. 25-33, 2013.

[12] J. B. L. S. K. Weinberger, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *Journal Of Machine Learning Research,* vol. 17, no. 1, pp. 207-244, 2009.

[13] T. Foss, E. Stensrud, B. Kitchenham and I. Myrtveit, "A simulation study of the model evaluation criterion MMRE," *IEEE transactions on software engineering,* vol. 29, no. 11, pp. 985-995, 2003.

A Method for Implementing Probabilistic Entity Resolution

Awaad Alsarkhi, John R. Talburt
Department of Information Science
University of Arkansas at Little Rock
Little Rock Arkansas, 72204, USA

***Abstract*—Deterministic and probabilistic are two approaches to matching, commonly used in Entity Resolution (ER) systems. While many users are familiar with writing and using Boolean rules for deterministic matching, fewer are as familiar with the scoring rule configuration used to support probabilistic matching. This paper describes a method using deterministic matching to "bootstrap" probabilistic matching. It also examines the effectiveness of three commonly used strategies to mitigate the effect of missing values when using probabilistic matching. The results are based on the experiment using different sets of synthetically generated data processed using the OYSTER open source entity resolution system.**

Keywords—Entity resolution; probabilistic matching; deterministic matching; Boolean rules; scoring rule; missing values

I. Introduction

The OYSTER open source entity resolution system **(https://bitbucket.org/oysterer/oyster/)** was designed to support both deterministic and probabilistic matching. Probabilistic matching is performed using a scoring rule based on agreement and disagreement weights [18]. Generating the weights typically requires more effort and analysis than using Boolean rules to implement deterministic matching. The scoring rule is similar to the Boolean rule which specifies a similarity function and optional data preparation function for comparing the values of identity attributes between two entity references [2], [3], [30]-[34]. The linking and review decision for each pair of entity references is determined by setting up a score threshold. Pairs scoring at or above the threshold are linked, and those below are not linked. Moreover, pairs scoring below the threshold, but above a user defined "review score" are written to an exceptions report for clerical review and possible remediation.

Each identity attribute has at least one agreement weight and at least one disagreement weight. In some cases, there can also be a third weight called a "missing weight" to be used instead of the agreement or disagreement weight in the case both values of the identified attribute are missing.

Basically, the power of the probabilistic matching is the ability to assign weights to particular values of identity attributes. Typically, individual weights are only assigned when the identity attribute has an uneven distribution of value frequencies, and then, only to the values with the highest frequencies, e.g. the top 10% of frequencies. Value-frequency weights are less effective for identity attributes with an even distribution of value frequencies low frequencies such as dates-of-birth or telephone numbers.

For example, in the context of a school system, if many different students have the first name "Mary," then the weight of agreement for "Mary" should be smaller than the agreement weight for a name used by only a few students, perhaps even one student. In general, the magnitude of the weight is proportionate to its ability to discriminate between references to different students. Weight calculations are discussed in detail in a later section of the paper.

Another advantage of the scoring rule is its natural support for the "clerical review" or "link remediation" process. This is a process by which pairs of references with minimal support for linking are reported for manual review by domain experts. This is easily done when using the scoring rule by simply flagging any pair of references with a total score very close to the match threshold, i.e., "near matches" for review. Although very labor intensive, the clerical review is often implemented in high-risk applications where high-levels of linking accuracy are required, such as patient medical records.

The organization of this paper is as follows:

- Section II discusses prior work related to this study.
- Section III describes the "bootstrap" method for creating a scoring rule.
- Section IV describes the calculation of the weights and three missing value treatments approaches based on that database by applying run of OYSTER is configured invariants algorithm designed for feature extraction.
- Section V gives the experimental results.
- Section VI compares results from Boolean Rules and Scoring Rule
- Section VII and VIII summarizes the conclusions of this work.

II. E Literature Review

Numerous systems have been used to solve some data quality issue can problems weaken the effectiveness of probabilistic matching [4], [5], [6]. One of the influential methods is OYSTER algorithms. While OYSTER does include the capability to implement probabilistic matching, many users are not familiar with its setup and operation. The purpose of this paper is to provide a step-by-step guide for users interested in experimenting with probabilistic matching using OYSTER. Particularly, the matching outcome of the scoring rule is lower when the sources have a high level of inconsistent value representations, such as misspellings, aliases, especially missing values. It should be noted here, while probabilistic matching has the ability to significantly improve record linking accuracy, it is not absolute. Certain data quality problems in the reference data can impair the effectiveness of probabilistic matching [10], [11], [19], [20]-[24]. In particular, the problems of inconsistent representation of identity attribute values and missing values have the most impact.

```
<ScoringRule Ident="Example1" MatchScore ="12" ReviewScore="11">
    <Term Item = "Fname" SIMILARITY="SoundexOrNickName"
AgreeWgt="11.27"
WgtTable="/Users/Alassad/Desktop/Oyster/LScoringIndex/Input/WgtTabl
eFirst.txt" DisagreeWgt="-8.30" />
    <Term Item ="Lname" SIMILARITY="Exact" DataPrep="Scan(LR,
Letter, 0, ToUpper, SameOrder)" AgreeWgt="4.20"
WgtTable="/Users/Alassad/Desktop/Oyster/LScoringIndex/Input/WgtTabl
eLast.txt" DisagreeWgt="-19.93"/>
    <Term Item="Nbr" SIMILARITY="Exact" DataPrep="Scan(LR,
DIGIT, 0, KeepCase, SameOrder)" AgreeWgt="10.42" DisagreeWgt="-
2.96"/>
    <Term Item="DOB" SIMILARITY="Exact" DataPrep="Scan(LR,
DIGIT, 0, KeepCase, SameOrder)" AgreeWgt="5.94" DisagreeWgt="-
1.31"/>
</ScoringRule>
```

Fig. 1. Example attributes script for scoring rule.

Figure 1 shows an example Attributes Script defining three identity attributes “Fname” (student first name), “Lname” (student last name), DOB (Date of Birth), and “Nbr” (street number of student address). Each run of OYSTER is configured by three XML Scripts: Run Script, Source Descriptor, and Attributes Script. The Run Script sets the overall purpose of the run called the “Run Mode.” The Run Modes for OYSTER are

- Merge-Purge
- Identity Resolution
- Identity Capture
- Identity Update
- Reference-to-Reference Assertion
- Structure-to-Structure Assertion
- Reference-to-Structure Assertion
- Structure Split Assertion

The two Run Mode configuration relevant to this paper are Merge-Purge and Identity Capture [3]. Both read a set of entity references and link the references together according to the match rule (either Boolean or Scoring) specified in the Attributes Script and transitive closure. The only difference is that Identity Capture also creates an Identity Knowledgebase whereas Merge-Purge does not. Identity Capture is the first step in using OYSTER for identity management applications such as MDM [5], [9], [13], [16].

The Source Descriptor Script is used to define the source and layout of the entity references to be linked by the system. The Source Descriptor can describe several different file sources including comma separated value (CSV) and fixed-width field formats. Because the focus of this paper is on the configuration of the Scoring in the Attributes Script, we will always assume the Run Mode is Identity Capture and the reference source is CSV.

III. Generating Weights and Configuring the Scoring Rule

A. Scoring Rule Configuration

The Attributes Script has two primary functions. The first is to designate the identity attributes in the entity references being processed. The second is to define how the identity attributes are to be matched. OYSTER currently supports two types of matching definitions, Boolean rules for deterministic matching and the scoring rule for probabilistic matching [7], [8]. The Scoring Rule configuration defines the Match Score to be 5 and the Review Score to be 4. All three of the identity attributes are used in the scoring rule. The first <Term> of the rule defines the agreement (“Similarity”) as “SoundexOrNickName”. This means that two first names will be in agreement if they generate the same SOUNDEX code (e.g. “Philip” and “Phillip”) or if they are nicknames of each other (e.g. “Robert” and “Bob”).

The first <Term> also defines a weight table for frequently occurring first name values named “WgtTableFirst.txt”. The weight table contains a set of key-value pairs. The key is the attribute value, e.g. “Mary”, and the value is the corresponding agreement weight, e.g. “4. In the weight table, the key and value must be separated from each other by a single tab character. When a weight table is given, the OYSTER logic is as follows. If the values agree, then the table is searched for both attribute values. Note: the lookup operation for attribute values in the table is not letter case sensitive. If both attribute values are found in the table, then it uses the larger of the two agreement weights. If only one value is found, it uses the one agreement weight found. If neither value is found, then it uses the default agreement weight given in the <Term> element. In this case, the default agreement weight is “9.08”. Of course, if the attribute values disagree, then in the <Term> element the disagreement weight of “2.45” is used.

The second <Term> element for “Lname” operates in much the same way except that it uses a “DataPrep” function in addition to a Similarity function. The DataPrep function is a data transformation that is applied to the raw attribute value and the result of the DataPrep transformation is passed to the Similarity function. The DataPrep function used here is the “SCAN” function. The parameters of the SCAN direct it to search the characters of the attribute value from left-to-right (LR), extract only letters (Letter), extract all letters found (0), change all letters to upper case (ToUpper), and keep the letters in the same order asthey were found (SameOrder). For example, “O’Malley” and “OMalley” would both transform to “OMALLEY”.

An important note about the use of weight tables: If a DataPrep function is used in the <Term>, then the lookup key used for the weight table is the output of the DataPrep transformation function. If a DataPrep function is not used, then the lookup key is the raw, untransformed attribute value from the input file, but ignoring case. It is important to understand this logic to properly construct an agreement weight table. For example, if a <Term> were to define the DataPrep function as “SOUNDEX”, then the keys defined in the weight table for this <Term> should be SOUNDEX output hash codes, not the input names. So, if the input value is “John” and the DataPrep is SOUNDEX, then the system will try to lookup for “J500” the SOUNDEX hash of “John” in the table. On the other hand, if DataPrep is SCAN (Letters to uppercase), then the system would try to lookup “JOHN” in the weight table [12], [15], [29].

B. Selecting Identity Attributes

The first step in selecting identity attributes is to profile the data. The primary candidates for identity attributes are those that have a high completeness and high uniqueness score. Attributes with low completeness or low uniqueness can be used as a secondary identity attribute to support the primary attributes.

The accuracy of the scoring rule is lower when the sources have high levels of inconsistent value representations (misspellings, aliases, etc.) and missing values. Also be sure to watch for placeholder values. These can bias the weight calculations.

The best candidates for value-level weights are identity attributes with an uneven distribution of the value frequencies, e.g. like names where several have high frequencies and many other have low frequencies. Attributes with evenly distributed values are best assigned only attribute-level weights

C. Linking the Records using Another Process

The calculation of the agreement and disagreement weights follows the method developed by Fellegi and Sunter (1969). In their method, the agreement weight is based on the ratio of two conditional probabilities. The probability equivalent references will agree on the attribute or attribute value divided by the probability non-equivalent references will agree on the attribute or attribute value. The problem is that we don’t know which references are equivalent, i.e. are referencing the same entity. In fact, if it were already known which references were equivalent and non-equivalent then there would be no purpose in setting up the matching rules.

One solution to this problem it to manually determine the correct linking for some subset of the data, i.e. create a “truth set”. However, this can be exact costly in terms of time and effort. A second solution is to approximate the true linking with another ER process. For this paper, we follow the second solution by running OYSTER with a set of Boolean (deterministic) rules. The clusters created by this process are then assumed to be the equivalent references for the purpose of calculating the agreement and disagreement weights.

D. Calculating Weights

From cluster profile of the Boolean linking process (found in the OYSTER .log file), calculate the total number of equivalent pairs and total number of non-equivalent pairs.

This is done by taking each line of the cluster profile and calculating the number of pairs based on the cluster size. For example, if the profile has 8 clusters of size 5, then each cluster of size 5 represents (5x (5-1)/2 = 10 equivalent pairs

Next multiple is the number of pairs in each cluster by the number of clusters. For example, if there are 10 cluster of size 5, then 8 x 10 = 80 equivalent pairs

Sum these count for each line in the cluster profile to get the total number of equivalent pairs. Let E represent this number.

Calculate the total number of possible pairs by taking the file size (total number of records) N and calculating N x (N-1)/2. Let T represent this number. For example, if the file contains 100 records, the T = 100x99/2 = 4,950 pairs

Calculate the total number of non-equivalent pairs by subtracting the total number of equivalent pairs found from the cluster profile from the total number of possible pairs. Let U represent this number, i.e.

U = T – E.

Next, join the Link File output from the OYSTER Boolean rule output (. link file) back to the source data by record identifier. This will create a new file called the weight analysis table. Each row of the weight analysis table should have a column for the

- Record identifier
- OYSTER identifier
- Value of each identity attribute in the record

Next, select the first identity attribute to process. Create a new column in the weight analysis table to hold the standardized value of the first attribute. Standardization is the process of removing variation from the values of the selected attribute. At a minimum you should convert all names to upper case letters so that "ANN" and "Ann" will come together.

If you want to remove even more variation, you could replace the name with the SOUNDEX code, NYSIIS, or other phonetic hash code. If you are planning to use a weight table, please read the previous section on Weight Table Logic. The setup of your weight table depends upon whether you use SOUNDEX for DataPrep or for Similarity. For example, SOUNDEX would make "ANN" and "ANNE" group together. If you used SOUNDEX for the Similarity, then you still have to keep the original name value together with its standardized or hashed value to populate the weight table. In the current implementation of the Scoring rule, if you use DataPrep, it looks up the DataPrep value in the table, but if you don’t use DataPrep it looks up the source value while ignoring case. So if you are calculating value-level weights using SOUNDEX or some other hash, it will be simpler if you use the hash function as the DataPrep so that the keys in the weight table are also hash values.

After creating the column of standardized values for the attribute in the weight analysis table, do a primary sort of the table by the standardized value of the attribute, and a secondary sort by the OYSTER identifier. This will give you groups of consecutive rows with the same standardized value, and at least one subgroup with the same OYSTER identifier. If it only has one subgroup, then the group and subgroup are the same.

For the first subgroup, count the number of records having the same standardized value and same OYSTER identifier. Call this number A1V1N1 for Column 1, Value 1, Number of rows in Subgroup 1.

Now, calculate the number of equivalent pairs agreeing on standardized value V1 for this subgroup. Call this number C1V1E1 meaning Column 1, Value 1, Equivalent pairs in Subgroup 1. A1V1E1 is calculated by

A1V1E1 = A1V1N1 x (A1V1N1-1)/2

If there is more than one subgroup for standardized value V1, repeat this calculation for each subgroup of V1. If there is a second subgroup, then C1V1N2 would represent Value 1, Number of rows in Subgroup 2. The second subgroup would contribute another

C1V1E2 = C1V1N2 x (C1V1N2-1)/2

Equivalent pairs agreeing on V1.

Calculate the total number of equivalent pairs agreeing on V1 (call this number C1V1E) by summing the equivalent pairs for all subgroups, i.e.

C1V1E = C1V1E1 + C1V1E2 + ….

Next calculate the total number of pairs agreeing on the first standardized value V1 in Column 1 by counting the number of rows having the same standardized value V1. Call this value C1V1N. Note V1N should be the same as the sum of count of each subgroup, i.e.

C1V1N = C1V1N1 + C1V1N2 + …

The total pairs agreeing on V1 (call this C1V1A) is calculated by

C1V1A = C1V1N x (C1V1N-1)/2.

Next calculate the number of non-equivalent pairs agreeing on value V1 (call this value C1V1U). This number is calculated by

C1V1U = C1V1A – C1V1E.

For example, suppose C1 is the first name attribute, and if C1V1 is the standardized value “JOHN” and there are 5 records with this values that standardize to “JOHN” such as “John” or other variations. Suppose there are 2 subgroups of this value. The first attribute C1 has been selected for attribute-level or value-level weights, we need to make some other calculations for the entire attribute.

The first is the attribute’s overall disagreement weight D. This is done by repeating the previous steps for all of the values of the attributes. In particular, we need to Sum all pairs agreeing on the values of C1, as

C1A = C1V1A + C1V2A + …

Sum all equivalent pairs agreeing on the values of C1, as

C1E = C1V1E + C1V2E + …

Calculate the number of non-equivalent pairs agreeing on the values of C1, as

C1U = C1A – C1E

Then the

Disagree Weight C1

= log2{(1 – (C1E/E)) x (U/(1 - C1U))}

The agreement weight for C1 will depend upon whether it will have value-level weights or only a single attribute-level agreement weights. If there is only one agreement weight for the entire attribute C1, then it is calculated as

Agree Weight C1 = log2{(C1E/E) x (U/(C1U)}

using the same numbers as described above.

However, in the case of value-level weights, then formula is the same, but the values of C1E and C1U should only be the sum of the pair counts for the values not used for individual weight calculations. In other words,

C1A = C1VxA + C1VyA + …

where Vx is a value of C1 not given an individual weight, and Vy is a value of C1 not given an individual weight and so on. Another way to look at this is to think of removing the frequent values from the weight analysis table and calculating the overall agreement weight for the remaining values.

The same process described above is repeated for every identity attribute.

E. Constructing the OYSTER Scoring Rule

Following the same example, as shown in Figure 1, each one of the attributes from the weight analysis table will be used to define a <Term> in the Scoring Rule. The attribute name will be the value for the Item in the <Term>

For the <Term>, define a DataPrep and Similarity function in alignment with the DataPrep and Similarity functions used in the Boolean rules. Even though the scoring rule will allow you to use the same attribute in more than one scoring rule <Term>, we recommend only using each attribute once to define a <Term> in the Scoring Rule with one Similarity Function and one optional DataPrep function.

If a scoring rule <Term> defines only attribute-level weights, then the agreement weight (AgreeWgt value) should be set to the overall agreement weight as calculated in the previous section. Similarly, the disagreement weight (DisagreeWgt value) should be set to the overall disagreement weight for the attribute.

A <Term> can also define an optional Missing weight. The Missing weight is added to the overall score instead of the AgreeWgt or DisagreeWgt values whenever both of the values being compared are missing. Although the missing weight is often set to zero, it is possible to calculate a weight by treating “missing” as a value just as you would a normal name value.

While more formal research is needed to determine the best strategy for scoring missing values, there is some preliminary evidence the best strategy is to not assign a special value for missing value agreements. If no Missing weight value is given, then the DisagreeWgt value will be used if either or both values of the attribute are missing.

If a scoring rule term uses value-level weights, then the AgreeWgt value should be set to the overall agreement weight for only the rows in weight analysis not used to calculate individual value weights. On the other hand, the disagreement weight is still calculated as the overall disagreement weight for all values of the attribute. In addition, a <Term> using value-level weights must use the value of WgtTable to define a path to the text file containing the agreement weights to be used for the frequently occurring values. The weight table is simple. Each row should only have an attribute value followed by only ONE-tab character followed by the agreement weight. Any rows starting with two exclamation marks (!!) are ignored as comments.

In addition to setting the parameters for each <Term> item, the match score (MatchScore value) must be set at the <ScoringRule> level. An optional review score (ReviewScore value) value can also be set. The review score is a value less than match score. When a pair of references generates a score less than the match score, but greater than or equal to the review score, the two references will be written to a Clerical Review report. The purpose of the clerical review is to allow the user to see the pairs of references scoring just below the match score and determine if they are true negative pairs or false negative pairs.

If the review score is not given, then no clerical review report will be produced. If a review score is given, it should not be set too far below the match score. If too many pairs are written to the Clerical Review Report, it may significantly degrade the performance of the system. The goal is to only produce and review pairs of records scoring close to the value of the match score.

On exception to this rule is when using the Clerical Report for analysis and quality assurance. By running the scoring rule with a small test file, setting the match threshold actual high, and the review score low, every pair of references will be written to Clerical Review Report. This allows the user to analyze and check the scores being generated for every pair of references in the test file.

F. *Setting the Optimal Match Score*

The final parameters needed to create a complete scoring rule are setting an optimal match score. The value assigned to the MatchScore attribute in the <ScoringRule> term determines the linking performance of the system.

As the match score increases, the scoring rule will link fewer references resulting in higher precision and lower recall. Conversely, decreasing the match score will link more references resulting in lower precision and higher recall. At the extremes, if the match score is set so high that no pair of records can be linked, the precision will be 100% but the recall will be 0%. Conversely, if the match score is set so low that all records are linked, the precision will be 0% and the recall will be 100%. In both cases, the F-measure will be 0.0.

The balance between the precision and recall is the F-measure calculated as the harmonic mean of the precision and recall values. For some applications, the precision may be more important than recall, and the optimal threshold may not be the value producing the highest F-measure, but instead, the necessary level of precision. Other applications may favor recall over precision. For research described here, the threshold value producing the highest F-measure is considered optimal.

The review score and Clerical Review Report can be important tools in adjusting the value of the match score. It allows the user to examine pairs of records scoring just below the match score and possibly provide evidence for decreasing or increasing the match score value.

There is no formula for setting the Match Score. It is typically adjusted by trial and error. To start, review the output of the Boolean rule process and select a cluster containing several records, then extract these records from the source file to create a new test file containing only the records from one cluster. As described earlier, set the match score high and the review score low. Also, omit the Index so that every pair of records in the input will be compared, scored, and written to the Clerical Review Report.

The scores from the report can be used as the starting point for setting the match score. The reasoning is as follows: If the Boolean rules linked these records into a single cluster, then the scoring rule should link these records as well. Estimate the starting match score as the average of all the scores generated by the cluster. To improve the starting estimate, repeat the same process for other clusters linked by the Boolean rules, and include these scores in the average.

The reason for starting with the average score and not the minimum score is that not every pair of records in a cluster necessarily matches. Some records come into the cluster through transitive closure. So even if a cluster contains records A, B, and C. It may be because A matched with B, and B matched with C, while it may not be true that A matched with C. Setting the match score low enough for A and C match in the scoring rule may create a large number of false positive links. A better estimate is the average score.

G. *Indexing the Scoring Rule*

The scoring rule must have an index (inverted index blocking) when the reference source contains more than a few hundred records. Otherwise, the runtime will be unacceptable. The problem is that aligning the indices to a scoring rule is more complicated than aligning indices to a Boolean Rule. For Boolean rules, you can inspect each rule (AND clause), and as long as the rule uses at least one hash function comparator, then it is possible to design an index in alignment with the rule.

In general, indices for Scoring Rules are less precise than for Boolean Rules. Rather than a single index on First-Name + Last-Name+ Street-Number, you might instead have three separate indices, one for First-Name, another index for Last-Name, and another for Street-Number. The downside to the strategy of indexing each attribute separately is that the list of match candidates will be larger for each input reference, causing more comparisons, and slower performance.

Another method for designing the index is a more complex analysis. If the scoring rule uses N attributes, then generate all possible combinations of agreement and disagreement weights as N-dimensional vectors of ordered pairs. The ordered pair for each dimension would represent a combination of Agree or Disagree together with the corresponding weight. For example, if there are three attributes, the vectors would look like [(A, 120), (D, -30), (A, 50)] where (A, 120) represents agreement on the first attribute with a score of 120, (D, -30) represents a disagreement on the second attribute with a score of -30, and (A, 50) agreement on the third attribute with a score of 50.

As you can see, there can be a great many of these vectors when there are value-level weights. For example, if the first <Term> has a table of 20 value agreement weights, plus the default agreement, and default disagreement. This would be a total of 22 possible weights for the first attribute. If these second attribute has 32 weights, and the third attribute has 3 weights, then there are 22 x 32 x 3 = 2,112 possible weight vectors.

Next, compute the total score (sum of the weights) for each vector and classify the vectors into 2N categories according to agreement/disagreement patterns similar to those used by Fellegi and Sunter [1]. As in the previous example where N=3, then there would be 8 (23) pattern categories: “DDD”, “DDA”, “DAD”, “DAA”, “ADD”, “ADA”, “AAD”, and “AAA”. Next, create an analysis file by generating an ordered pair for each vector where the first element is the agreement pattern and the second element the total score. Using the previous example for the vector [(A, 120), (D, -30), (A, 50)] the total score is 140. This vector would produce the pair (ADA, 140).

Finally, sort all of the ordered pairs according to their match scores. For any given choice of a match score, such as the starting estimate, you can now see all of the agreement patterns above and below that score. Suppose in this example that for a selected match score, all of the patterns above this score are “AAD”, “DAA”, or “AAA”. From this information you can design two indices to support the scoring rule for this match score. One index is a combination of the first and second attributes and the second index a combination of the second and third attributes. From an indexing standpoint, the pattern “AAA” is redundant to the first two patterns of “AAD” and “DAA”, i.e. any candidate returned by indexing on all three attributes would also be returned by indexing on two elements. However, the index solution just described will only provide alignment when all of the attributes a hash function for comparison, such as SOUNDEX or SCAN(), which can also be used for indexing.

H. *Assessing Performance and Adjusting Match Score*

You can use the Review Score setting to generate the Clerical Review Report. This will let you examine “near” matches and decide if you need to lower the Match Score. However, this approach does not show you possible false positive linking of pairs above the Match Score.

IV. EXAMPLE

The following example gives more detail on the calculation of the weights used for the Scoring rule given in Figure 1 as well as the performance of its linking with various MatchScore settings as shown in Figure 4. The weight calculations and performance measures are based on a test file of 141,745 synthetic person references known as ListB. The ListB reference were created in previous research [24]-[28] in a way that the correct linking is known. This allows the precision and recall to be immediately calculated for any given ER process.

Even though the true linking for ListB is known, it was not used to calculate the probabilistic weights for this example. The true linking was only used to evaluate the precision and recall of the linking results. The weights were calculated by bootstrapping the scoring rule from Boolean linking as described earlier. Given a set of entity references in the real world, exactly which pairs of references are equivalent is not known. The process starts by selecting the identity attributes and linking the references with a reasonable set of Boolean (deterministic) rules to serve as a “surrogate truth set” for generating probabilistic weights. The OYSTER Attributes script used to define the starting Boolean rules is shown in Figure 2.

```
<OysterAttributes System = "IndexTestWith">
        <Attribute Algo="none" Item="Fname"/>
        <Attribute Algo="none" Item="Lname"/>
        <Attribute Algo="none" Item="Nbr"/>
   <Attribute Algo="none" Item="DOB"/>
<!-- -->
<Indices>
          <Index Ident="X1">
                    &lt;Segment                              Item="Lname"
Hash="SCAN(LR,LETTER,0,ToUpper,SameOrder)"/>
                    &lt;Segment                                Item="Nbr"
Hash="SCAN(LR,DIGIT,0,ToUpper,SameOrder)"/>
          </Index>
     <Index Ident="X2">
       &lt;Segment                                           Item="Lname"
Hash="SCAN(LR,LETTER,0,ToUpper,SameOrder)"/>
       &lt;Segment                                             Item="DOB"
Hash="SCAN(LR,DIGIT,0,KeepCase,SameOrder)"/>
     </Index>
</Indices>
<IdentityRules>
          <Rule Ident="1">
                    <Term                                 Item="Fname"
SIMILARITY="SOUNDEXORNICKNAME"/>
                    <Term                                 Item="Lname"
SIMILARITY="SCAN(LR,LETTER,0,ToUpper,SameOrder)"/>
                    <Term                                   Item="Nbr"
SIMILARITY="SCAN(LR,DIGIT,0,ToUpper,SameOrder)"/>
          </Rule>
          <Rule Ident="2">
                    <Term                                 Item="Fname"
SIMILARITY="SOUNDEXORNICKNAME"/>
                    <Term                                 Item="Lname"
SIMILARITY="SCAN(LR,LETTER,0,ToUpper,SameOrder)"/>
                  <Term                                     Item="DOB"
SIMILARITY="SCAN(LR,DIGIT,0,ToUpper,SameOrder)"/>
          </Rule>
        </IdentityRules>
</OysterAttributes
```

Fig. 2. Boolean rule configuration for processing ListB.

The Boolean rules shown in Figure 2 uses First Name, Last Name, Street Number, and Date-of-Birth. The rules as configured here produced a linking results with a Precision of 0.9800, Recall of 0.5064, and F-Measure of 0.6677.

Figure 2 also shows the inverted index blocking used for this rule set. Note that the first index (X1) is in alignment with both Boolean rules (1 and 2). Any pair of references matching by either rule would also produce the same match key by this index. The two terms of the index correspond exactly to the second and third terms of both the rules. This is possible because hash functions such as SCAN () can be used as both a comparator (True if both hash values agree) and a match key generator (hashing one string into another string). On the other hand, the comparator function NICKNAME used in the second Boolean rule (Ident=2) is a true similarity function requiring two strings as input while producing a Boolean True or False output. Because similarity functions do not produce a hash output, they cannot be used for inverted indexing using hash keys.

Following the weight calculation steps outlined in the previous section, ListB was profiled and individual agreement weights were calculated for the most frequently occurring Fname and Lname values. In this case, the top 10% of the high frequency names were selected. Figure 3 shows some of the entries in the weight table for Lname. Note that many of the names were only given as initials as illustrated by the first entry “A”, and the file also contained many typographical errors, e.g. “ADASM” a typing transpose of “ADAMS”.

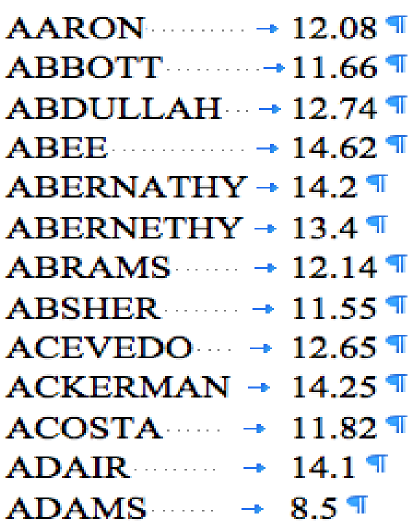

AARON → 12.08 ¶
ABBOTT → 11.66 ¶
ABDULLAH → 12.74 ¶
ABEE → 14.62 ¶
ABERNATHY → 14.2 ¶
ABERNETHY → 13.4 ¶
ABRAMS → 12.14 ¶
ABSHER → 11.55 ¶
ACEVEDO → 12.65 ¶
ACKERMAN → 14.25 ¶
ACOSTA → 11.82 ¶
ADAIR → 14.1 ¶
ADAMS → 8.5 ¶

Fig. 3. Lname agreement weights.

Figure 4 also shows the performance of its linking with various MatchScore settings. After removing the top 10% of the most frequent first name values, the agreement weight for the remaining names was calculated as 11.27. The disagreement weight for all first name values was -8.30. Similarly, after removing the top 10% of last name values, the agreement weight for the remaining last names was calculated as 4.20, and the disagreement for all Last Name values was -19.93. Because the street number values and date-of-birth values were somewhat evenly distributed, only the overall agreement and disagreement weights were calculated these attributes. For street number the values were 10.42 and -2.96, respectively, and for date-of-birth the values were 5.94 and -1.31, respectively.

```
<OysterAttributes System="School">
  <Attribute Algo="none" Item="Fname"/>
  <Attribute Algo="none" Item="Lname"/>
  <Attribute Algo="none" Item="Nbr"/>
  <Attribute Algo="none" Item="DOB"/>
<!-- -->
<ScoringRule Ident="Example1" MatchScore ="12"
ReviewScore="11">
    <Term Item = "Fname" SIMILARITY="SoundexOrNickName"
AgreeWgt="11.27"
WgtTable="/Users/Alassad/Desktop/Oyster/LScoringIndex/Input/Wg
tTableFirst.txt" DisagreeWgt="-8.30" />
    <Term Item ="Lname" SIMILARITY="Exact"
DataPrep="Scan(LR, Letter, 0, ToUpper, SameOrder)"
AgreeWgt="4.20"
WgtTable="/Users/Alassad/Desktop/Oyster/LScoringIndex/Input/Wg
tTableLast.txt" DisagreeWgt="-19.93"/>
    <Term Item="Nbr" SIMILARITY="Exact"
DataPrep="Scan(LR, DIGIT, 0, KeepCase, SameOrder)"
AgreeWgt="10.42" DisagreeWgt="-2.96"/>
    <Term Item="DOB" SIMILARITY="Exact"
DataPrep="Scan(LR, DIGIT, 0, KeepCase, SameOrder)"
AgreeWgt="5.94" DisagreeWgt="-1.31"/>
</ScoringRule> <Indices>
    <Index Ident="X1">
      &lt;Segment Item="Fname" Hash="SOUNDEX"/>
      &lt;Segment Item="Lname"
Hash="SCAN(LR,LETTER,0,ToUpper,SameOrder)"/>
      &lt;Segment Item="Nbr"
Hash="SCAN(LR,DIGIT,0,KeepCase,SameOrder)"/>
    </Index>
    <Index Ident="X2">
      &lt;Segment Item="Fname" Hash="SOUNDEX"/>
      &lt;Segment Item="Lname"
Hash="SCAN(LR,LETTER,0,ToUpper,SameOrder)"/>
      &lt;Segment Item="DOB"
Hash="SCAN(LR,DIGIT,0,KeepCase,SameOrder)"/>
    </Index>
    <Index Ident="X3">
      &lt;Segment Item="Fname" Hash="SOUNDEX"/>
      &lt;Segment Item="Nbr"
Hash="SCAN(LR,DIGIT,0,KeepCase,SameOrder)"/>
      &lt;Segment Item="DOB"
Hash="SCAN(LR,DIGIT,0,KeepCase,SameOrder)"/>
    </Index>
    <Index Ident="X4">
      &lt;Segment Item="Lname"
Hash="SCAN(LR,LETTER,0,ToUpper,SameOrder)"/>
      &lt;Segment Item="Nbr"
Hash="SCAN(LR,DIGIT,0,KeepCase,SameOrder)"/>
      &lt;Segment Item="DOB"
Hash="SCAN(LR,DIGIT,0,KeepCase,SameOrder)"/>
    </Index>
  </Indices>
</OysterAttributes>
```

Fig. 4. Final scoring rule with blocking.

V. Testing and Results

Table 1 shows the details of precision, recall and F-measure comparison between scoring rule results. Figure 5 is the line chart specifically compares the F-measure performance.

TABLE I. COMPARE THRESHOLD SCORE RULE RESULTS

	Score Rule		
Threshold	Precision	Recall	F-Measure
-20	0.79	0.72	0.75
-15	0.9	0.72	0.8
-10	0.91	0.72	0.8
-5	0.91	0.72	0.8
0	0.93	0.71	0.81
5	0.93	0.64	0.76
10	0.95	0.59	0.73
15	0.98	0.56	0.71
20	0.98	0.48	0.65

VI. COMPARISON OF RESULTS

Table 2 shows the scoring rule using weights computing from Boolean rule linking can outperform the initial set of Boolean rules. In this example, a 20% improvement in F-measure with significantly higher precision

TABLE II. SUMMARY OF RESULTS

	Rule- Generation		
	Precision	Recall	F-Measure
Boolean Rule	0.6677	0.5064	0.6677
Scoring Rule	0.93	0.71	0.8014

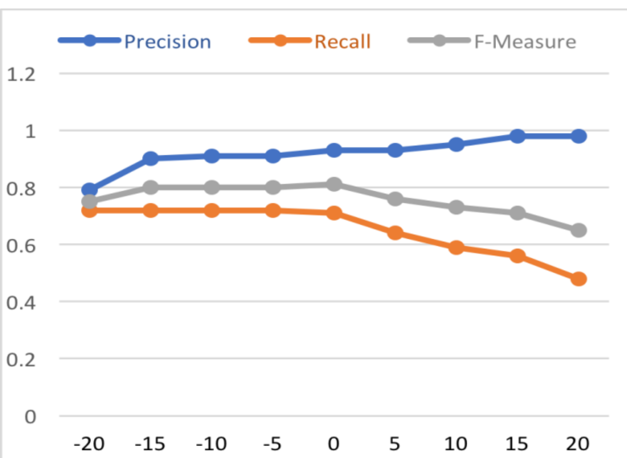


Fig. 5. Scoring rule performance for different thresholds.

VII. CONCLUSION AND FUTURE WORK

The bootstrap process described in this paper allows researchers to configure a probabilistic entity resolution engine using OYSTER by taking advantage of its ability execute both Boolean rules and the scoring rule with frequency-base weights. It also shows the amount of tuning required to get good results from probabilistic matching due to the large number of variable including the

- Choice of identity attributes
- Comparators used for the identity attributes
- Choice of identity attributes for frequency-based value weights
- Number of values selected for individual weights
- Match score selection
- Impact and treatment of missing values
- Impact and treatment of inconsistent values

At the same time, these variables create the opportunity for additional research including the use of both, agreement and disagreement weights at the value level, analysis tools for setting optimal match scores, the treatments to mitigate the effect of missing values, and the optimal frequency cutoff for frequency-based value weights.

Another avenue of research is based on repeating the bootstrapping process. Just as the Boolean linking results can be used as a proxy for equivalence to generate weights for a scoring rule, the linking of the scoring rule itself can be used to generate a second set of weights, and so on. Will the weight generated from on-scoring rules output allow the construction of another scoring rule that will always outperform its predecessor? [14], [17], [18].

REFERENCES

[1] Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183-1210.

[2] Zhou, Y., Talburt, J.R., Kobayashi F., and Nelson E. (2012). Implementing Boolean matching rules in an entity resolution system using XML scripts. *The 2012 International Conference on Information and Knowledge Engineering (IKE'12)*, Las Vegas, Nevada, July 16-29, 2012 (pp. 332-337).

[3] Agichtein, E., & Ganti, V. (2004). Mining reference tables for automatic text segmentation. *Proceedings of the Tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 20-29). ACM.

[4] Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media.

[5] Benjelloun, O., Garcia-Molina, H., Kawai, H., Larson, T. E., Menestrina, D., Su, Q., ... & Widom, J. (2006). Generic entity resolution in the serf project. Stanford InfoLab.

[6] Kobayashi, F., Eram, A., & Talburt, J. (2018). Entity resolution using logistic regression as an extension to the rule-based oyster system. *Proceedings 2018 IEEE Conference on Multimedia Information Processing and Retrieval* (MIPR) (pp. 146-151).

[7] Wang, P., Pullen, D., Talburt, J. R., & Wu, N. (2014). Probabilistic Matching Compared to Deterministic Matching for Student Enrollment Records. *Proceedings Information Technology: New Generations* (ITNG), (pp. 355-359).

[8] Kobayashi, F. & Talburt, J.R. (2013) Probabilistic Scoring Methods to Assist Entity Resolution Systems Using Boolean Rules, *The 2013 International Conference on Information and Knowledge Engineering (IKE'13)*, Las Vegas, Nevada, July 22-25, 2013, CSREA Press (pp. 101-107).

[9] Whang, S. E., & Garcia-Molina, H. (2010). Entity resolution with evolving rules. *Proceedings of the VLDB Endowment*, 3(1-2), 1326-1337.

[10] Bilenko, M., Kamath, B., & Mooney, R. J. (2006). Adaptive blocking: Learning to scale up record linkage. In Data Mining, 2006. *Proceedings Sixth International Conference on Data Mining* ICDM'06. (pp. 87-96).

[11] Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 1-16.

[12] Koybayashi, F. & Talburt, J.R., Oyster Version 3.6 Reference Guide, https://bitbucket.org/oysterer/oyster/downloads/

[13] Talburt, J. R., Zhou, Y., & Shivaiah, S. Y. (2009). SOG: A Synthetic Occupancy Generator to Support Entity Resolution Instruction and Research. *Proceedings International Conference on Data Quality* (ICIQ-09), pp. 91-105.

[14] Wang, P., Pullen, D., Talburt, J., & Wu, N. (2014) Iterative approach to weight calculation in probabilistic entity resolution. *19th MIT International Conference of Information Quality*, August 1-3, 2014, Xi'an, China, (pp. 244-258).

[15] Zhou, Y. & Talburt, J. (2011) The role of asserted resolution in entity identity information management, *Proceedings: 2011 Information and Knowledge Engineering Conference (IKE 2011)*, Las Vegas, NV, July 18-20, 2011 (pp. 291-296).

[16] Hashemi, R., Ford, C., Vanprooyan, T., & Talburt, J. (2002). Extraction of features with unstructured representation from HTML documents. In P. Isaias (Ed.), *International Association for Development of Information Society (IADIS) International Conference on WWW/Internet 2002* (pp. 47- 53). Lisbon, Portugal.

[17] Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer Science & Business Media.

[18] Hernández, M. A., & Stolfo, S. J. (1995). The merge/purge problem for large databases. *ACM Sigmod Record* (Vol. 24, No. 2, pp. 127-138).

[19] Pushkarev, V., Neumann, H., Varol, C., & Talburt, J. R. (2010). An Overview of Open Source Data Quality Tools. *Proceedings of the International Conference on Information and Knowledge Engineering* IKE (pp. 370-376).

[20] Talburt, J., Wang, R., Hess, K., & Kuo, E. (2007). An algebraic approach to data quality metrics for entity resolution over large datasets. In *Information quality management: Theory and applications* (pp. 1-22). IGI Global.

[21] Dong, X., Halevy, A., & Madhavan, J. (2005). Reference reconciliation in complex information spaces. *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (pp. 85-96).

[22] J. Talburt, "Oyster v3.3 Reference Guide 2013 03 18: http://downloads.sourceforge.net/project/oysterer/OYSTER_3.3/. [Accessed 16 July 2015].

[23] Talburt, J. R. (2011). *Entity resolution and information quality*. Morgan Kaufmann.

[24] F. Kobayashi and J. R. Talburt, Decoupling Identity Resolution from the Maintenance of Identity Information. *Proceedings: International Conference on Information Technology* (ITNG), Las Vegas, NV, 2014. (pp. 349-354).

[25] Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science*, 130(3381), 954-959.

[26] Zhou, Y. (2012). Modeling and design of entity identity information in entity resolution systems. *Doctral Dissertation*. University of Arkansas at Little Rock.

[27] Syed, H., Talburt, J., Liu, F., Pullen, D., & Wu, N. (2012, January). Developing and refining matching rules for entity resolution. *Proceedings of the International Conference on Information and Knowledge Engineering* (IKE) (p. 1).

[28] Talburt, J. R., & Zhou, Y. (2015). Entity information life cycle for big data: Master data management and information integration. Morgan Kaufmann.

[29] Pullen, D., Wang, P., Talburt, J. & Wu, N. (2014) Mitigating data quality impairment on entity resolution errors in student enrollment data. *11th International Conference on Information and Knowledge Engineering*, July 21-24, 2014, Las Vegas, NV, (pp. 96-100).

[30] Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4), 5-33.

[31] Talburt, J., Wu, N., Pierce, E., & Hashemi, R. (2007). Entity identification using indexed entity catalogs. In H.R. Arabnia & R.R. Hashemi (Eds.), *The 2007 International Conference on Information and Knowledge Engineering* (pp. 338-342). Las Vegas, NV: CSREA Press.

[32] Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). Swoosh: a generic approach to entity resolution. *The VLDB Journal—The International Journal on Very Large Data Bases*, 18(1), 255-276.

[33] Zhou, Y., Talburt, J. R., Su, Y., & Yin, L. (2010). OYSTER: A tool for entity resolution in health information exchange. *Proceedings of the 5th International Conference on Cooperation and Promotion of Information Resources in Science and Technology* (pp. 358-364).

[34] Wang, R.Y., 1998. A product perspective on total data quality management. *Communications of the ACM* 41 (2), 58-65.

Intellectual Paradigm of Artificial Vision: From Video-Intelligence to Strong Artificial Intelligence

E. M. Yarichin
Laboratory of brain-like technologies
"NeocorTek Lab" (LLC)
Krasnoyarsk,
Russian Federation

V. M. Gruznov
Trofimuk Institute of Petroleum Geology and Geophysics
Siberian Branch of Russian Academy of Sciences
Novosibirsk, Russian Federation

G. F. Yarichina
Institute for Business Processes and Economics Management
Siberian Federal University, SFU
Krasnoyarsk, Russian Federation

***Abstract*—A new (post-Shannon) informational approach is suggested in this paper, which allows to make deep analysis of nature of the information. It was found that information could be presented as an aggregate of quantitative (physical) and qualitative (structural) components to be considered together. It turned out that such full information theory could be efficiently used as the guiding theory at modeling of video-information recognition, perception and understanding. These hierarchical processes are solving the intellectual tasks step-by-step for formation of the corresponding video-information evaluation and also represent a strong interactions-measurements video-information's ensuring adequacy of these assessments. That is why there is a need to build corresponding video information macro-objects (video-thesauruses) on every level of hierarchy of artificial vision system, which are formed by training (self-training) and form together an upward hierarchy of qualitative measuring scales. The top of this hierarchy is video-intelligence. Information theory of artificial intelligence is a logical development of new information approach from analysis to synthesis. Further "analysis through synthesis" allows establishing the informational nature and structure of not only video-intelligence, but also strong artificial intelligence, which for video-intelligence constitute as intellectual suprasystem.**

***Keywords*—Gauge approach; fibration space; informational hypergraph; video-intelligence; strong artificial intelligence**

I. Introduction

Intuitive representation of intelligence as some informational macro-object that a perform information processing for the purpose of her understanding led to deep revision of classical (by Shannon) informational approach. As this approach was quite heuristic one, the goal was to build an overall view of surrounding material world properties using the actual advances in fundamental physics and mathematics without intuitive and heuristic aspects. Actually, the quantitative classical approach to information did not meet the needs of the researchers who deals with the tasks which qualitative nature became more obvious. Such tasks as recognition, perception and understanding can be example. If of usage of the quantitative approach to information at recognition tasks is somehow applicable, then in more complicated informational tasks of perception and understanding it is totally unhelpful. Any attempts to improve the classical informational approach by concept of information's value and other intuitive and heuristic aspects did not lead to anything.

That is why, post Shannon approach the information was suggested by one of the authors (E.M. Yarichin) of current article. This practical physics-mathematics approach let to clearly realize all the depth of information paradigm, to obtain the correct definition and to find out the structure of information. Information is turned out to be presented as a system of physical (quantitative) and structural (qualitative) components strongly interacted with each other that are always to be considered together. Quanta of physical and structural informational components are fundamentally differing from each other. Quanta of physical component are unstructured (zero-dimensional, similar to the point). In its turn, structural component quanta or, otherwise, local qualts could be considered as infinitesimal univariate quantities determined in one-dimensional space and due to this fact have a certain orientation. In small dimension world's spaces local qualts could be considered as the most natural elements to synthesize one-dimensional global 1-qualts, then two-dimensional global 2-qualts based on the previous ones, as well as global 3-qualts and 4-ones further.

Thus, as for full information (both physical and structural), known as video-information, its analysis for artificial vision is turned out to be a very important and obligatory stage, but not a sufficient one. Actually, if analysis is presented as investigation method of dividing unit on its parts, than synthesis as investigation method to merge divided parts and terms which were obtained during the analysis is more important when forming video informational evaluations. In other words, if post Shannon approach to information is an analysis, it takes to establish a dual approach to information, i.e. a synthesis to provide a forming upward hierarchy for of evaluations video-information in artificial vision.

II. The Gauge Approach to Video-Information Processes in The Fibration Formalism

Video-information is inherent in any material system formed by a strong interaction (interaction-measurement) of its own (unary) or mutual (binary) components [1]-[4]. These components of video-information must be comparable physically and structurally and the video-information itself could be interpreted in two ways. On the one hand, video-information is the only keep geometrophysical quantity (in the form of a second-order tensor) in the corresponding material system and, as a result, is taken as a source of its video-informational (geometrophysical) field (active, passive or activated). On the other hand, video-information is considered as geometrophysicsal measure of interaction-measurement of the physical and structural components of the corresponding material system. This dynamic aspect of video-information is a consequence of "gauge principle" based on the idea of gauge invariance.

In practice, the method of video informational field is carried out by the weak field approximation only, what is explaining by natural reduction of video-informational field during its paraxial space-time propagation from the source (active and/or activated video-information) to video sensor. Moreover, the weak video informational field could be considered as a gauge field [2], [3]. Due to weak informational field's ambiguity (gauge arbitrariness) the calibration of the video-receiver of the vision system is necessary. Only then will this geometrophysics measuring instrument uniquely form an estimate of weak video information in the form of an image (Fig. 1).

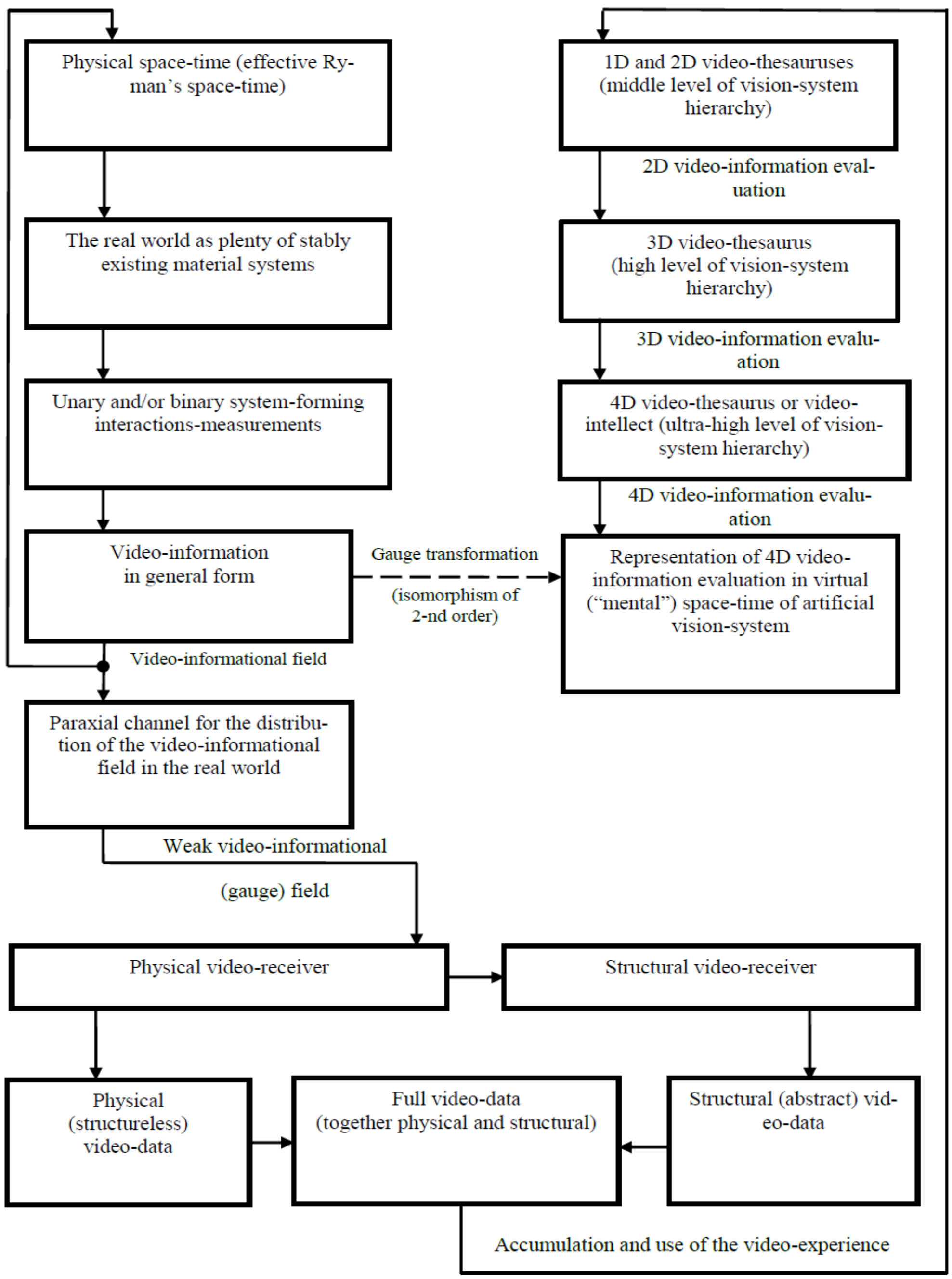


Fig. 1. General structural and functional scheme of artificial vision processes: the formation of video-information, the weakening of the video-information field, the restoration of video-information and the representation of its evaluation in the "mental" space of the artificial vision system.

Weak informational field taken by video receiver is determined from its aperture's dimensions and initiates the process of artificial vision (Fig. 1). Besides, physical and structural components of video-receiver provide the analysis of physical and structural components of weak video-information respectively which is collated to this weak field so it has a weak (coarse) topology. Therefore, the weak video informational evaluation obtained is presented as obvious two-dimensional image (in the form of a 2-manifold) and also is the base to form a more advanced representation about this video-information at virtual ("mental") space with large number of dimensions. Corresponding abstract (e.g. 3D) image has a stronger (thin) topology, which could be presented by bundle on two-dimensional items, each one of them is an image of a studied object in a certain orientation concerning video-receiver.

The process of Images formation could be considered as weak video-information evaluation obtained by weak field analysis as a gauge field. From this point of view, images forming could be considered as gauge transformation according to gauge theory. As is known, such transformation does not affect on physical object's characteristics that could be studied and measured so it provides the adequacy of weak video-information and of evaluation the weak video-information.

As gauge transformations are supra-coordinated transformations they don't act on vectors and tensors of space-time but act on the spaces of internal characteristics of video-informational fields [5]-[7]. Therefore, the geometry determined by the gauge fields is not space-time geometry, but a more General kind – a fibration space.

Thus, gauge approach to video-information could be formed due to local invariance principle corresponded to compensation circuit of gauge theory as well as within the framework of common description for video informational processes in terms of fibration. The fibration is presented as powerful tool for going out of traditional gauge theory. This could be the method formalization modeling of systems with the help of fibrations what is especially important for those systems where usual methods cannot be applied [7]-[9].

III. Video-Informational Approach to Artificial Vision

The theory of information in its modern form is dedicated to the problems of information processing and transmission with redundancy, noise, losses and coding taken into account. Classical theory of information hasn't consider the fundamental problem of information creation and distinguish not take into consideration what could be referred to information and what to senseless and noise [10]-[14]. In addition, the information's forming it is assumed the prerogative of the third party system or suprasystem. This approach, allowable at communication theory, makes it hard to analyze and research processing systems of full information when not only quantitative (physicals) but also qualitative (structural: geometrical, logical) information's characteristics determine the end processing result and therefore should be considered together.

Necessity of understanding video-informational processes specificity from the point of modern science (not heuristics), made forced to revise the fundamentals of information theory. That's why appear a new paradigm of information's which based on fundamental analysis of information occurrence in the real World. Such new informational paradigm is a logical result of modern scientific knowledge development and allows switching from heuristic aspects to modern study the problem of information's quantity and quality joint evaluation. While studying this problem it was realized that artificial understanding the information of surrounding world is constitute as intellectual process. However, this process also is not to be intuitive and/or heuristic but it must be correct from the point of view of modern physical and mathematical fundamental science.

Thus, new approach to information based on informational paradigm of the World generalizes and supplements K. Shannon approach that based on heuristics of information's quantitative measurability, being stated long before the era of high information technologies even began. Traditional (by Shannon) informational approach has shown its efficiency at adequate informational tasks on information's quantity evaluation, but at the same time it turned out to be totally unhelpful when used for other purposes, i.e. at tasks where quality of information (geometry and logical properties of surrounding world) it the crucial factor. In other words, K. Shannon informational approach not allow finding out and/or taking into account quantitative and qualitative information's attributes that could be determined by corresponding physical (quantitative) and structural (qualitative) measurements.

In this connection, it is necessary to clarify the complicated issues of the real World's information content and respectively world's understanding is based on this information. The absence of answers on these issues makes it difficult to deep understanding of notions of information and its evaluation. In turn this circumstance prevents an effective development of high and ultra-high information technologies which includes an intellectual artificial (technical, computer, machine) vision.

In artificial vision (IV) the weak informational field is the carrier of weak video-information in the avail-able (in the hardware sense) video-information channel of one or another physical nature (Fig. 2) [2]. Therefore, the weak video informational field is so important for artificial vision. Exactly this physical field interacts with the system of vision and initiates the artificial vision where process the following interrelated tasks could be pointed out:

1. IV's direct problem assumes of evaluating of weak video-information by short-range interaction-measurement with vision system. The weak physical video-informational field is this interaction's intermediary to the physical component of video-receiver. The physical component of weak video-information's evaluation is a measure of such interaction, i.e. the weak video informational field's paraxial propagation and small aperture's of physical video-receiver allow to form the evaluation of weak video-information's physical component, traditionally considered as image.

2. IV's inverse problem provides a recovery for structure component of video-information which is weakened naturally as a result of transformation of video-information of common type to the weak one at the video informational field's propagation through the layer of space-time. Actually, the "object-video-receiver" space layer extent is traditionally considered at

paraxial approximation [15]-[17] under conditions of material objects observation. It allows to register (in the near-field or far-field diffraction region) only the weak video informational fields (primary and/or secondary), which source is the corresponding weak video-information. Thus, being formed as a result of corresponding IV direct problem's solving the weak video-information evaluation (physical picture) initiates IV's inverse problem, which could be solved only on the base of long-range interaction-measurement evaluation of the weak video-information with specially formed video-information resources (video-thesaurus and video-intelligence) of vision system. These resources act as video-informational macro-objects that could be considered as measuring scales presented by a plenty of corresponding measuring video-structures. Long-range interaction is used without any intermediary's as physical field, what is the evidence of formal mathematical nature of such interaction and therefore is unlimited for quick speed, so in fact it is instantaneous. In practice, it could correspond to wide use of special tools to make the video-information's evaluation by formal methods of present-day mathematics, which is, as well known, represent an immense nonlinear structure with very strong interactions [18]-[20].

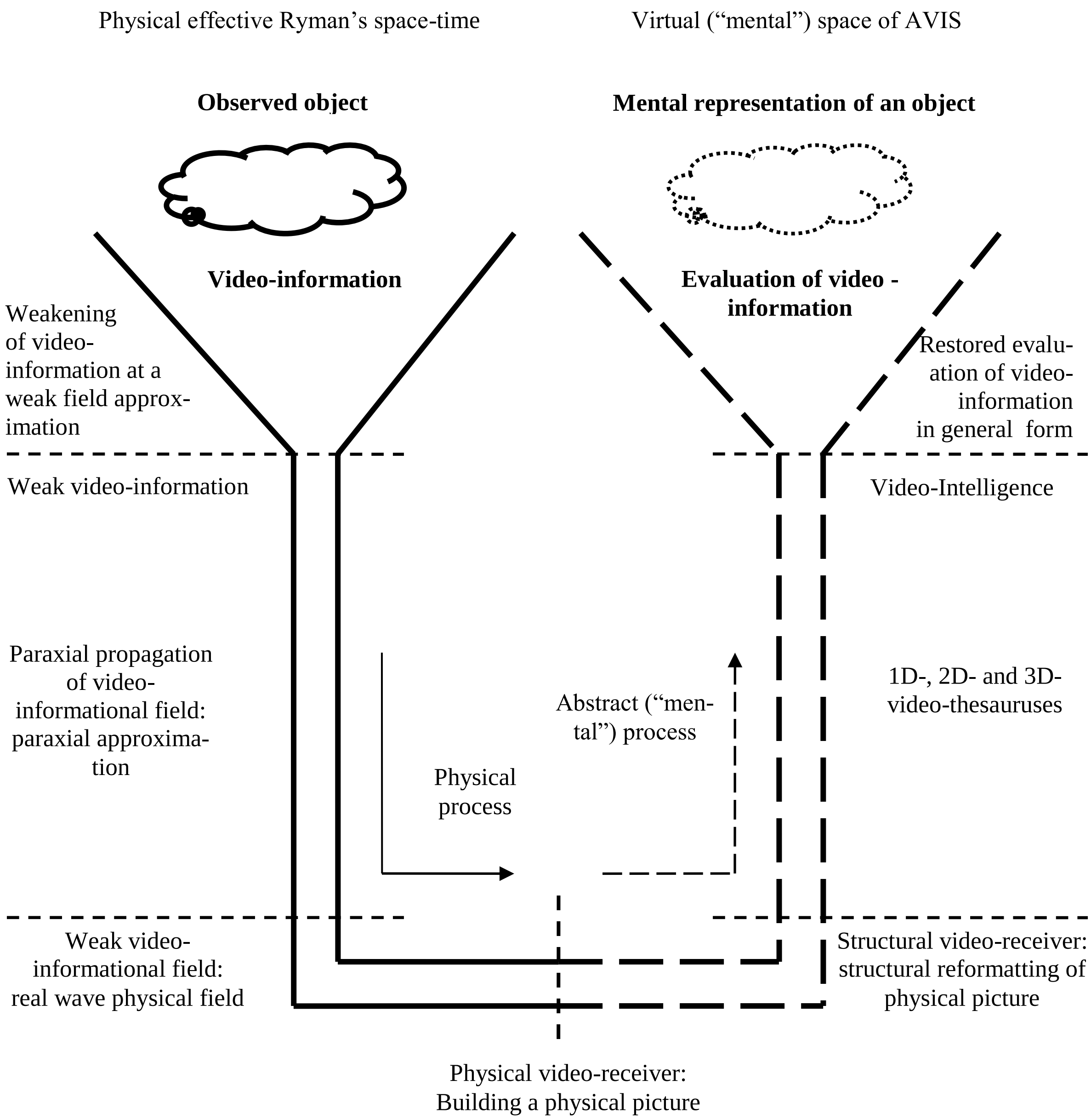


Fig. 2. The conditional symbolic scheme of video-information weakening by paraxial physical channel of propagation, physical component's forming of the evaluation weak video-information and subsequent restoration of video-information in virtual ("mental") space of artificial vision intellectual system (AVIS).

An evaluation of the weak video-information physical component in the form of physical picture should be considered at quantitative (physical) and qualitative (structural) aspects together, i.e. as partially physicalized. In this case a discrete physical reflection could be considered as plenty of physical samples with the properties of physical quants of video-information, which structure collapse to a point. Geometrical point as structure, as known, has a zero-dimensional order and thus it doesn't allow forming unambiguous higher order structures from the points. That is why, all geometrical elements (physical samples, pixels) reduced anyway to a point are unstructured. Using these elements, for example, it is possible to make a direct sample of field physical video-data on practice in the multi-element video-receiver mouth located in the lens's or reflector's focus (optical, thermal, acoustic, radio wave) [15]-[17], [21]-[23]. The example of indirect registration of physical video-data is a joint sample of amplitudes and phases of radio- or acoustic holograms provided the further reconstruction of corresponding physical picture [24]-[26].

Structural (totally geometrized) component of the weak video-information evaluation or, in other words, structural reflection could be collated at qualitative (topology) aspect to the structure of the observed two-dimensional surface of visually controlled material object. As is known from topology, this surface could be divided into arbitrarily small parts ("be measured" with any high accuracy) by plenty of one-dimensional geometrical objects (lines) only [27]. Such division of two-dimensional surface into the one-dimensional geometrical structures could be presented as its one-dimensional fibration.

In the surrounding macro world exactly are interactions-measurements (internal or external) of material systems generate the properties of any nature [1]. There is a single informational pattern in accordance which the interaction-measurement takes place only if the features of some dual interaction's components are comparable and one component's properties are macroscopic for the proper-ties of another one. The result of measuring, made by compare of micro-component (measured value) to macro-component (measuring scale) of corresponding dual interaction - measurement, could be a ratio and/or the system of ratio (structure that also includes video-structure) of this or that dimension. Meanwhile interaction–measurement macro-component could be presented as quantitative scale of ratios (dimensional physical ratios) and/or qualitative scale of spatial ratios systems (dimensional abstract video-structures).

There could be two types of measurements realized this way: physicals (quantitative) and structural (qualitative) [28]-[30]. Physical measurements mean the existence of quantitative (extensive) scales, e.g. absolute ones and ratios scales, expressed by numbers and allowed for calculations. Structural changes in its turn imply the existence of qualitative (intensive) scales, e.g., title and category scales dealing with the qualitative differences, when properties (attributes) ordering is possible, but calculations are not allowed, only comparisons. This is due to the fact that computing operations are applicable to spatial numbers, not to properties [30]. Moreover, in case of structural measurements there is no ability to set calibration and deviation (accuracy) for qualitative measuring scale, as the last one could be formed, e.g., as a list.

Physical quantities defined above as intensity measures of interaction-measurement of forms of matter (e.g., substances with appropriate field) naturally exist in nature. Therefore measuring of physical quantities assumes the existence of some physical device that made the interaction-measuring with the physical quantity and provides the result of measurement as some ratio (of number) of this or that physical dimension.

Structures in a natural way meetings in form temporary, spatial and space-time kinds of distribution (of laws of changes) of physical quantities. Structural interactions-measurements, aimed to obtain the assessment (of measuring result) of qualitative information, also assume the existence of some abstract (formally mathematical) instrument for structures "measuring". Such "measuring instrument" should be able to realize the interaction-measurement between the structures (ratios systems), providing the result of qualitative measurement in qualitative measure units, among which there could be simpler structures (in particular video-structures), which interpreted as dimensional structures.

It should be pointed out that AV process presented as a sequence of interactions–measurements (Fig. 2) is significantly oriented on the level of video-informational ratios systems (of video-structures), therefore concerning vision system it takes availability an advanced structural video-information resources on every level of structural video-information processing hierarchy. On the middle hierarchy's level (which is psychologically corresponding to perception and recognition of the weak video-structures), the support must be provided for the assessments (measurements) of those image's physical structures that took place in the composition of the physical picture. This is provided by 1D and 2D video-thesauruses. On the high and ultra-high levels of vision system hierarchy, psychologically corresponded to "objectification" (perception) and video-information's understanding, 2D video-information assessment initiates a video-information resource of high level (3D video-thesaurus), which in turn initiates a video-information resource of ultra-high level of hierarchy (4D video-thesaurus or, in other words, video-intelligence) through 3D video-information assessment. All these video-information resources are presented as a video-experience, formed artificially and memorized by vision system as a result of her training (self-training) for 1D, 2D, 3D и 4D video-data, collated with the real world's objects that could be used as examples of possible objects for video-observation.

As is known, thesaurus is presented as organized information having a number of various states, in this regard similar to entropy (although thesaurus and entropy have a different meaning) [31]. There-fore, thesaurus is measureable and measured in units of information. Thesaurus may be considered as internal informational resource of corresponding system, which could have a multilevel structure, e.g., hierarchically repeat at various level of detail, i.e. with the use of various types of morphism on every level of the hierarchy. Finally, in terms of systemology, thesaurus for the thesaurus itself could be considered as meta-thesaurus or, in other words, as consciousness or intelligence [31].

Thesaurus of vision system (video-thesaurus) is not observed in reality, as it is presented by passive video-informational macro-object (video-resource), which is able to make interactions-measurements with video-information on the high level of vision system hierarchy. For this, the video-thesaurus has having in its composition advanced set of dimensional video-structures. Actually, video-thesaurus is a dynamic form of video-experience. In particular, video-experience of vision system in the conditions of cognitive interaction with real world gives a support to qualitative measurements on every level of hierarchy, which sequentially form of real world visual representation.

The forming and keeping of such video-informational macro-objects as video-thesaurus or video-intelligence justifies the existence of virtual space which is the "mental" space of vision system [32]. Exactly here is possible to get an abstract (formally mathematical) generation and association of video-information's evaluations, obtained on the all hierarchical levels of video-information's processing (evaluation). Virtual space of vision system is induced by information's evaluations in a manner similar to how video-information induces curved surrounding space-time. Thus, the virtual space of vision system could be presented as full and adequate reflection of the real world, as formal analogue of physical space-time. Such kind of virtual spaces could be considered as a type of "modeling reasoning", the essence of which are concluded in virtual reality modeling as phenomena in the field of machine psychology.

The concept of video-sensor space generates the concept of vision as of a "mental" vision for the real world's space-time content with help of video-experience definitely organized at video-sensor's memory. As J. Piaget wrote in this regard, «Looking at the object is an intellectual act already». Thus, natural and artificial vision could be collated psychical and intellectual mechanism respectively which will allow seeing the world as it is.

As is known, isomorphism of 2nd order between virtual ("mental") space-time and physical space-time is presented as psychologically reasonable assumption for explanation of the visual representations of the human. [33]. It is assumed that visual information contained at human's memory is in relation isomorphism of 2nd order to the corresponding real video-data. In other words, the natural vision or artificial emulation "mental" images, formed in a mental space of corresponding vision system are resemblance to the real objects' images. Besides, a feature of this resemblance is that the ratio systems (video-structures), formed by mental images elements in the memory of vision system are the same as the ones between the elements of the video-informational images, formed by the real world. Thus, isomorphism of 2nd order differs from isomorphism in general (mathematical concept for designation of relationship for two objects basically identical to each other) in that it implies something closer to resemblance.

In this connection, one can put forward a hypothesis that artificial vision process, in case of being adequate its result, could be referred to video-information transformations, that don't change the physical condition of material object and don't make any influence on its observed characteristics. It is as a result of this is reaching effect of isomorphism of 2nd order between video-information and its evaluation formed by vision (Fig.2). As is known, such transformations are called gauge ones [8], [34]. Gauge transformation of some quantity corresponds to arbitrary (scale) change of her value with constant gradients of this quantity being maintained. Absent change for the quantities observed at this transformation means gauge invariance of the real world observed [8]. In the human world gauge invariance is perceived as imperceptible game with numbers, which does not make visible changes to the observing world, therefore, people don't notice this transformation [34].

Based on video-information paradigm, can be implemented the natural decomposition of artificial vision as process of video-information evaluation, to some stages (sub processes) to simplify the solution of this definitely complex problem. For example, at the initial stage of artificial vision the interaction-measurement of the weak video informational field with the video-receiver should be considered. It allows getting image which could be considered as evaluation weak video-information. This evaluation also has both physical and structural components for which could be designated according to their meaning where physical component of the image could be considered as physical picture, while the image's structural component present as structural picture [35].

On the further steps of artificial vision there is the processing of the image is realized, which is presented as 2D manifold (2-manifold) in terms of traditional geometrical intuition. At artificial vision the image processing aims to reconstruct the image from the weak (coarse) topology to the level of stronger (fine) topology typical for, e.g., 3D video-information evaluation. Such evaluation of video-information in view of the fact that with regard when 3D manifolds (3-manifolds) geometrical intuition doesn't work out will have a form of some abstract 3D image. The necessity of abstract 3-manifolds representation forces to abstractly consider the images that generate them (2-manifolds). In case when space-time, which embraces solving artificial vision issue, accepted as four-dimensional one, abstract images in the form of 2-manifolds must totally give way to the ones in the form of 3-manifolds.

Even greater dimension of artificial vision task appears in case when controlled object is four dimensional, e.g. it takes form of space-time scene. In this case, the enveloping space is five-dimensional; therefore this circumstance can take one aside to the side from solving the problems of the real world's visual perception. Actually, additional (fifth) dimension is compactified and is enclosed with very small period (10^{-31} cm), therefore 5-th coordinate is not observed [36], [37]. However, it should be taken in mind that artificial vision is generated by the effort to emulate the natural vision which is the sense organ; so, it is intended to fix a kind of projection of four-dimensional (space-time) processes on the really perceptible 3D (spatial) world. Modern researches in the area of small dimensions topology provide every reason to hope that the structure of four-dimensional space-time is defined by its 3D part's structure [18]-[20].

The forms of real objects in surrounding world could be rather complex so they are difficult to be described analytically and be modeled. The obvious solution, totally meeting the famous quotation "To understand is means to get used to and learn how to use" [R. Feynman], is in memorization of a plenty of real space-time ratio and systems of space-time ratios (video-structures). This is carried out to use their as dimensional ratios and video-structures, of which as a result of their association (during training and self-learning) the video informational macro-objects (qualitative measuring scales) of vision system could be formed.

IV. Structural and Functional Features of Intelligent Systems of Artificial Vision

Artificial vision intelligent system (AVIS) solves the following problems:

1) Evaluation of weak video-information about of the visual observable condition of the real world's objects as some of 1D and 2D images. These images correspond to relatively "coarse" topologies of 1- and 2-manifolds, which could be collated with the intuitive understandable images of lines and surfaces.

2) Evaluation of "strong" video-information in the form of 3D and 4D images of observed objects of surrounding world, those are carriers of more "thin" topology. These images are matched 3- and 4-manifolds which are representing intuitionally non-understandable images.

Generally speaking, these two global video-information processing (evaluation) stages assume a significantly advanced set of complex and non-trivial tasks making hard for perception of important de-tails of artificial vision process. Therefore, it is necessary to develop a process decomposition of the global stages to sub-stages with shallower local subtasks. Their understood on easier when considering the process of video-information evaluation at AVIS as an intellectual multistep technology of video-information evaluation.

Structurally-functional scheme of AVIS stages is presented on Fig. 1. The stages of video-information's and video informational field's forming were studied at works [1]-[4], so it is not necessary to add any comments. Then by processing of weak video- informational field physical video-receiver of vision system formed physical video-data is as a plenty structureless samples of physical picture. Structural video-receiver of vision system form a structural video-data in the form of plenty of elementary structural samples (local qualts) by physical video-data's reformatting.

Physical picture is formed by video-receiver's physical component (television camera, thermal imaging, multielement radio- or hydro acoustic antenna) that provides a physical image's registration as a space-time sample of physical video-data (optical, thermal, radio wave and ultrasound). The sizes and number of these samples are defined by resolution of video-receiver as well as transverse and longitudinal (according to depth) dimensions of its field of vision. The importance to provide high resolution for the depth on the initial stage of artificial vision is confirmed by known experiments with random-pointed stereograms [21], [38]. As can be seen from these experiences, volumetric (stereoscopic, holographic) vision could not only precede the shape's perception, but also carry out without her recognition. In other words, perception of outlines and forms is not a prerequisite for achieving a volume effect, i.e. "not so much the contours are push [us] on the thought of depth, how much a depth - on the thought of contours" [J. R. Shiffman].

V. Informational Hypergraphs of Video-Intelligence and Strong Artificial Intelligence

In present time, the gauge theory at the solving some field problems successfully goes beyond a compensatory wording that which has proved itself well at in case of internal symmetries. Therefore, modern gauge approach could be formed not only proceeding to local invariance principle but also within the framework of general description classical fields on the base of fibration formalism which is commonly used at the transition from the local description of solved problem to its global one [7]-[9], [39].

The role of geometrical methods in various areas of science and technology increased significantly now. As is known, mathematical analysis methods provide a qualitative phenomena's characteristic in small values of changes of these or that parameters. Geometrical methods, in its turn, provide to a qualitative description of the same phenomena at large values of changes of its parameters. The other equally important peculiarity of geometrical methods is the convenience of geometric language for various regularities explaining. In this regard along with mathematical analysis of manifolds (with the use of modern differential geometry) a fibrations theory which is turned out to be the most convenient one at many applications has been developed greatly.

Already at the first stages gauge interpretation of video-information faces difficulties as generally gauge fields is connectivity as well as video informational field is a metric field (In this case two conditions are imposed on the connectivity: of symmetry and metricity [7], [8], [25], [40], [41]). The gauge status of video-informational field remains unclear but is sufficient productive in compensatory scheme of gauge theory in case internal symmetries. This allows to in particular correctly identify the peculiarities of video-information interaction generating the video-information and her internal structure's which from the help of coordinate transformations only intuitionally are define. However, the existing experience forming of gauge theories for field systems indicates that for the space-time symmetries the compensatory wording of gauge theory is turned out to be not effective enough so it is necessary to use fibrations mathematics [7]-[9].

At the initial stage of artificial vision equations forming, the traditional compensatory wording of gauge theory is quite enough. That is why, the hypothesis advanced earlier, asserting that the objects in effective Riemannian space-time and virtual ("mental") space of AVIS are connected by gauge transformation, should be used. This hypothesis is correct due to the fact that, as is known, gauge transformation not reflected on the observed state (physical and structural) of the a real world's objects [8], [34]. In this case, there is every reason to assume that video-information about a spatial content of the real world and its evaluation in mental space of AVIS to be connected with each other by isomorphism of 2nd order, which implies something closer to the similarity than isomorphism [33]. In other words, the video-information evaluation process is a vision process, in case there a relations between the natural video-information in a physical space-time and its evaluation in "mental" space of AVIS is close to of equality relations.

Limited opportunities of classical informational approach hardly met science and technology needs already at the end of 20th century. However currently (as established and shown in the works [1, 2]) quantitative (physical) and qualitative (structural) components of information should be considered together. Wherein, in the expression for a full video-information t_m^n, considering at coordinates of Minkowski space it's physical $t_{m(E)}^n$ and structural $t_{m(S)}^n$ components enter additively making it's share in a full video-information. In general it could be presented as

$$t_m^n = t_{m(E)}^n + t_{m(S)}^n \quad (1)$$

However, the surrounding material world due to non-linearity of corresponding sensors having usually quadratic characteristic of sensitivity's is usually percepted on full information intensity (power). In doing so, the expression for full video-information intensity I_m^n could be quite obviously presented as

$$I_m^n = \left(t_m^n\right)^2 = \left(t_{m(E)}^n + t_{m(S)}^n\right)^2 = \left(t_{m(E)}^n\right)^2 + 2t_{m(E)}^n t_{m(S)}^n + \left(t_{m(S)}^n\right)^2 \quad (2)$$

As shown from the latest expression, the one for video-information intensity in common form allows to explicitly identify its components interaction carried out by their multiplying. This confirms the assertion developed in works [1, 2] that the strong interaction of video-information components is possible only when each of them is dual. Namely, physical component of video-information is only partly physicalized and structural one is only partly geometrized respectively. Precisely because of duality for strongly interacted components a video-information becomes a system of physical and structural components considered together.

Video-information components interaction is particularly evident in the expression for weak video-information intensity

$$\overset{(0)}{I_m^n} = \left(\overset{(0)}{t_m^n}\right)^2 = \left(\overset{(0)}{t_{m(E)}^n} + \overset{(0)}{t_{m(S)}^n}\right)^2 = \left(\overset{(0)}{t_{m(E)}^n}\right)^2_{\to 0} + 2\,\overset{(0)}{t_{m(E)}^n}\,\overset{(0)}{t_{m(S)}^n} + \left(\overset{(0)}{t_{m(S)}^n}\right)^2_{\to 0} \to 2\,\overset{(0)}{t_{m(E)}^n}\,\overset{(0)}{t_{m(S)}^n}, \quad (3)$$

Where $\overset{(0)}{t_m^n}$ − weak video-information.

The meaning of the expression (3) is that in case of weak video-information the interaction of its physical and structural components is realize by its multiplying.

Let us consider the main aspects of the idea on which video-information evaluation in AVIS is based. As video-information evaluation is a video-information also, it is obvious that the results obtained above about video-information structure are also right for video-information evaluation $\hat{t}_m^n$, which is generally presented as

$$\hat{t}_m^n = \hat{t}_{m(S)}^n + \hat{t}_{m(E)}^n \quad (4)$$

The surrounding material world due to video-sensors non-linearity is commonly percepted by video-information intensity (power). Therefore, the expression for the intensity of video-information evaluation could be presented as

$$\hat{I}_m^n = \left(\hat{t}_m^n\right)^2 = \left(\hat{t}_{m(S)}^n + \hat{t}_{m(E)}^n\right)^2 = \hat{I}_{m(=)}^n + \hat{I}_{m(\times)}^n, \quad (5)$$

$$\hat{I}_{m(=)}^n = \left(\hat{t}_{m(S)}^n\right)^2 + \left(\hat{t}_{m(E)}^n\right)^2, \quad \hat{I}_{m(\times)}^n = 2\hat{t}_{m(S)}^n\,\hat{t}_{m(E)}^n, \quad (6)$$

Where

$\hat{t}_m^n$, $\hat{t}_{m(S)}^n$, $\hat{t}_{m(E)}^n$ − accordingly, video-information evaluation, its structural and physical components;

$\hat{I}_m^n$, $\hat{I}_{m(=)}^n$, $\hat{I}_{m(\times)}^n$ − accordingly, intensity of video-information evaluation, its background component and the wanted component, which considered as multiplication of physical and structural components of video-information assessment.

From the expressions (5) and (6) it follows that video-information evaluation intensity includes its physical and structural components interaction carried out by their multiplying.

Multiplicative character of components interaction of evaluation weak video-information is particularly demonstrable in the expression for the intensity of weak video-information evaluation

$$\overset{(0)}{\hat{I}_m^n} = \left(\overset{(0)}{\hat{t}_m^n}\right)^2 = \left(\overset{(0)}{\hat{t}_{m(E)}^n} + \overset{(0)}{\hat{t}_{m(S)}^n}\right)^2 = \left(\overset{(0)}{\hat{t}_{m(E)}^n}\right)^2_{\to 0} + 2\overset{(0)}{\hat{t}_{m(E)}^n}\,\overset{(0)}{\hat{t}_{m(S)}^n} + \left(\overset{(0)}{\hat{t}_{m(S)}^n}\right)^2_{\to 0} \to 2\overset{(0)}{\hat{t}_{m(E)}^n}\,\overset{(0)}{\hat{t}_{m(S)}^n} \quad (7)$$

or

$$\overset{(0)}{\hat{I}_m^n} = \left(\overset{(0)}{\hat{t}_{m(E)}^n} + \overset{(0)}{\hat{t}_{m(S)}^n}\right)^2 \to \overset{(0)}{\hat{I}_{m(\times)}^n} = 2\overset{(0)}{\hat{t}_{m(E)}^n}\,\overset{(0)}{\hat{t}_{m(S)}^n} \quad (8)$$

Currently when researching artificial vision systems there are the attempts to intuitionally circumvent the laws of nature, e.g. with the help of all kinds of neural networks, which is able to react only to the physical component's intensity of evaluation weak video-information collated to traditional physical image. By “deep” training or some other heuristic method on the base of neural networks, the researchers try to solve the problem of artificial vision with the structural component not being taken into consideration. However, according to expression (8) in this case inadequate evaluation weak video-information is formed.

In the process of solving the tasks of intellectual artificial vision a set of hierarchical video-information evaluations could be considered as a result. Its psychological analogues are recognition, perception and understanding of video-information. Such complex evaluation of video-information allows for every image of the real world to collate “mentally” the corresponding abstract image depending on the depth of solving video-information processing (evaluation) problem.

The system-forming factor of AVIS is upward video-thesauruses hierarchy, including ultra-high level’s video-thesaurus or otherwise video-intelligence. The whole this hierarchy is functioning in full compliance with post Shannon information paradigm as internal video-informational macro-objects and provide the strong video-informational interactions on every hierarchy’s level.

On the initial stage of artificial vision due to the relatively small apertures dimensions of video-receiver the space-time layer of «observable object – video-receiver» relative to video-informational field acts as a narrowband spatial filter. Therefore, the strong initial video informational field (secondary or primary), whose source present video-information with strong (thin) topology transforms into weak video-informational field obtained by narrowband filtering by specified spatial layer (Fig. 2). Weak video-information can be considered as the source of such weak video informational field, which formed from general view of video-information by its relative topology weakening (coarsening). Weak video informational field initiates artificial vision process by activating of passive video informational resource of AVIS. In doing so, physical and structural video-receivers of the system both provide a full sample of video-data, presented as physical quanta (structureless samples) and structural quanta (local qualts) arrays is considered together.

Fibration space as is known could be considered as some $k-$dimensional space (base), which points are changed to $l-$dimensional spaces (layers) with its own geometry [8]. Therefore, 3-manifold evaluation in the form of fibration could be considered as set-theoretic sum of two-dimensional layers. At computer representation this set theoretically sum (taking into account idempotency law where logical structures coefficients and exponents are eliminated from consideration) could always be reduced to the form of disjunctive union of layers parameterized by the database and “bonded” together by fibration space topology [8].

The majority of the real world’s objects have a video-information, which due to its complexity and unpredictability in general case could not be described formally in details. Therefore, there is a necessary to act according to the modern information template instead of trying to describe the real world formally. According to this template, any process that generates information at this or that form or providing its evaluation should be organized as interaction-measurement. This implies the presence in corresponding video-information system (physical or abstract) having at least two collated and strongly interacted micro- and macro-component. Micro-component of interaction is as measurable value while macro-component acts like a measuring scale “calibrated” at corresponding dimensional values or dimensional structures in order to quantitatively and/or qualitatively be evaluation results this interaction-measurement.

At such approach the process of artificial vision could be considered as sequential empirical process of hierarchical measuring scale forming for video-structures (similar to multi-level Bourbaki scale) and its further “mental” use for video-information evaluation about the spatial content of the real world. By the way, it was introduction of the Bourbaki scale formal concept has allowed N. Bourbaki formally introduce the concept of structure.

Let us consider the features of sequential forming for some video-thesaurus as video informational macro-object ordered by the principle of accounting for all paired relations between “dimensional” video-information’s that thesaurus consists of. This forms a certain systems of relations (structure) of video-thesaurus. Actually, for the composition of two video-information in general form, when there is no loss of video-information, their pairing is the ratio of union (of logical sum) is the simplest way. As is known, ratio could be represented generally as some operation not only between the numbers but also between the objects of any nature including the one between the structures. For the number of video-information’s more than two (three, four, etc.) one could speak number of elements (r_1) of video-thesaurus which at the increase of their number $(r_1 = 3, 4, 5, \dots K)$ is bound together in as much paired relations as different combinations from r_1 elements for two could be formed. It allows to present video-thesaurus forming scheme as $r_1(r_1-1)/2$ combinations paired relations of «dimensional» video-information (Fig. 3). This scheme provides video-thesaurus forming as a set-theoretic sum of informational components of the same dimension that could be transformed in accordance to idempotency law to more convenient (for computer analysis and synthesis) form of disjunctive union of these components. Thus, informational meaning of fibration and disjunctive union for of the components of having the same dimension mean that in geometrical structure is introduced structure of logic and this so it allows logically analyzing and synthesizing the relevant geometry.

The expression (9) as applied to 3D video-thesaurus demonstrates the transition from set-theoretical form of video-thesaurus to its representation in accordance with idempotency law in the form of disjunctive union of its video-informational components of the same dimension which appropriate to the high level of AVIS hierarchy

$$\theta_m^n = (K-1)\sum_{\eta=1}^{K} \hat{t}_{m(\eta)}^{n} \overset{\text{idempotency law}}{\Rightarrow} \bigcup_{\eta} \hat{t}_{m(\eta)}^{n}, \quad \eta = \overline{1,K}, \qquad (9)$$

where

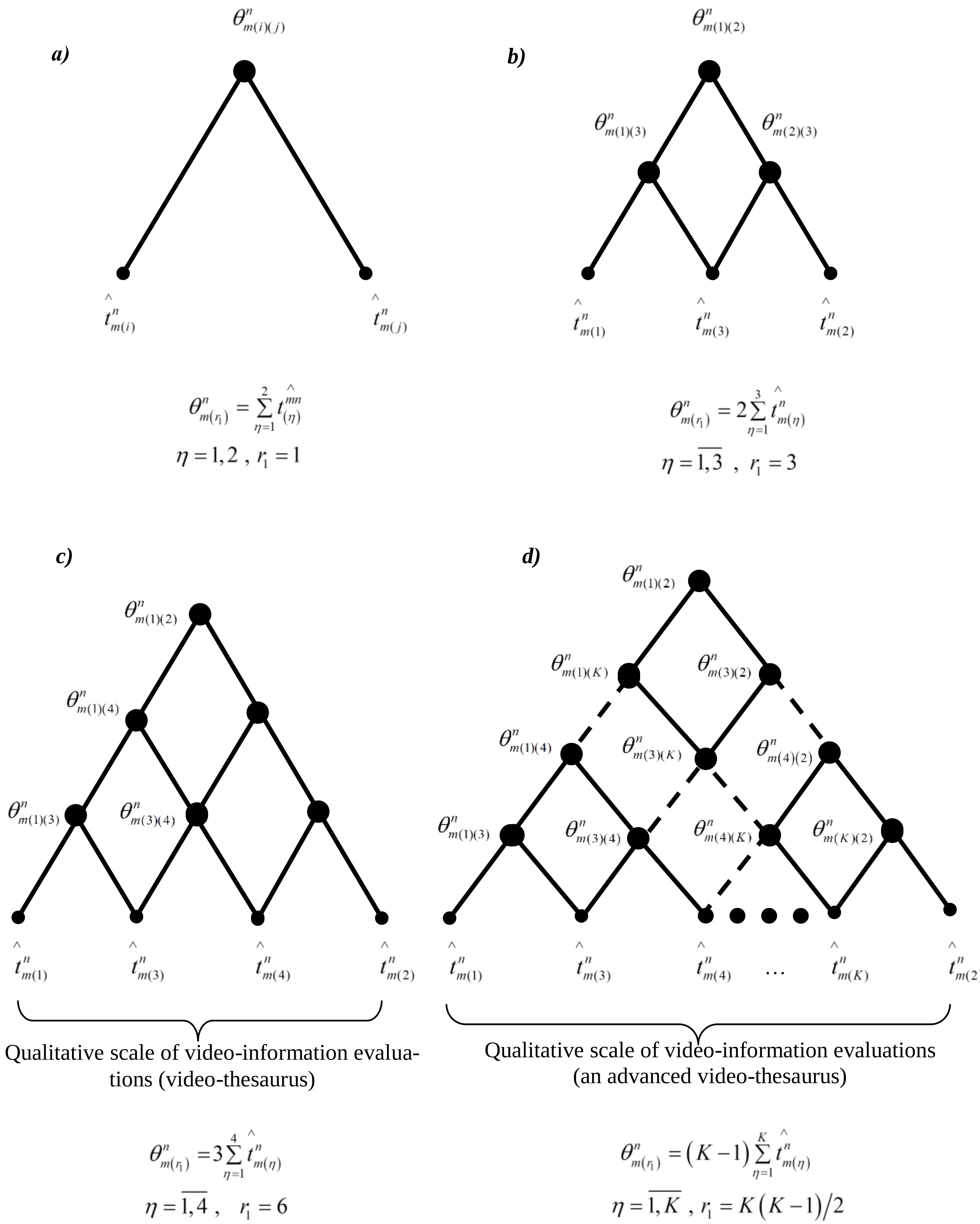


Fig. 3. Informational hypergraphs of video-thesauruses θ^n_m of 3-th (*b*), 4-th (*c*) and K-th (d) ranks with a heterarchical organizational structure on the base of paired relations of $\theta^n_{m(i)(j)}$ (*a*) video-information evaluations.

θ^n_m – 3D video-thesaurus: considered as 3D video informational macro-object or qualitative measuring scale representing the union of "dimensional" video-information of the same spatial dimension;

$\hat{t}_m^n = \bigcup_{\rho} \overset{(0)}{\hat{t}_{m(\rho)}^n}$, $\rho = \overline{1,R}$ – current "dimensional" video-information in the form of abstract 3D image (of a three-dimensional manifold), which fibration into two-dimensional images of the observing object's;

K – video-thesaurus rank; R – rank of a three-dimensional manifold.

Video-thesaurus development schemes are shown in Fig. 3, presenting hypergraphs, where assigned relations don't have fixed parity. The edges of hypergraph connect not the single pairs of vertex, but rather their subsets. It is important to emphasize that information contained at hypergraph is non-metric information and could be characterized only topologically. Hypergraphs, shown on Fig.3 correspond to the process of video-thesaurus sequential development as of qualitative "measuring" scale.

Generally, any video-thesaurus be formed by aplenty of video-experiences as sequentially increasing array of video-information evaluations. This array of video-information evaluations units all the paired relations between them and present itself as video-informational macro-object of some K-th rank. The rank of such macro-object corresponds quantity to its partial components (of "dimensional" video-informations) covered by the heterarchy of paired relations. This heterarchical organization structure of video-thesaurus that must form during AVIS training, generally speaking, very difficult at implement (Fig. 3). However, in accordance to idempotency law such integration of "dimensional" video-information, pairwise organized combinatorically, degenerates into much more simplier empirical qualitative measuring scale of video-structures in the form generalized informational hypergraph, which quite easy to be realized (Fig. 4).

The theory of physical structures currently proves that there does not exist nontrivial theories of ternary, tetradic, etc. systems of relations. In other words, nature limited itself only to cases of binary and unary systems of relations [36], [42]. Thus, the considered variant of video-thesaurus development at video-experience accumulation by combinatorial union of all paired additive compositions of video-information assessments is turns out to be the only possible one.

It should be noted, that in this case rather important additive law of composition acquires geometrophysical interpretation. Actually, unlike elementary algebra, additive representation of law of composition does not necessarily mean addition [43].

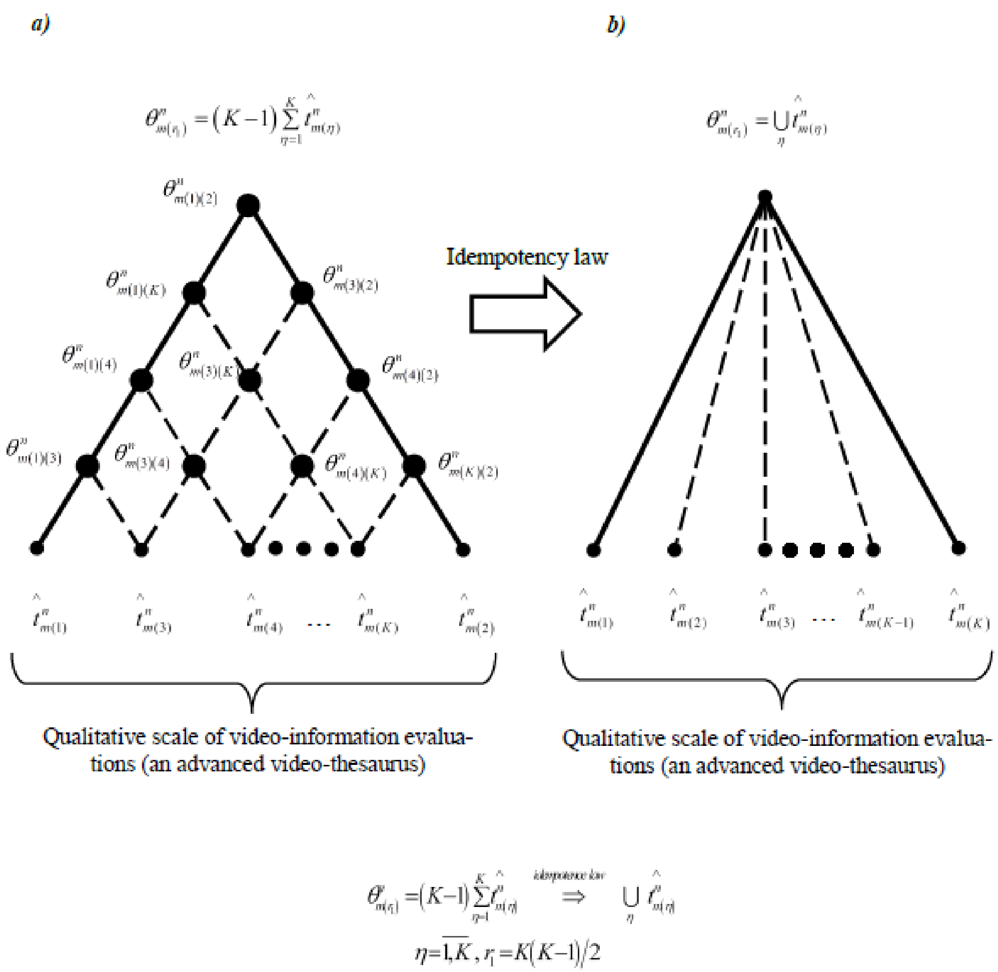


Fig. 4. Transformation of informational hypergraph video-thesaurus as set-theoretic sum of video-information evaluations (a) in to much more simple and available structure (b) of generalized informational hypergraph.

This operation only seemingly looks like an ordinary addition operation, but essentially, it is a disjunctive union of video-informations evaluations on corresponding (low, middle, high or ultra-high) hierarchy levels in ATIS.

From the expression (9) follows what 3D video-thesaurus in accordance with idempotency law and 3D generic hypergraphs structure becomes of qualitative measuring "scale" for 3D video-informatios. These "dimensional" video-informations are presented by abstract images $\hat{t}_m^n$, collated to the 3D video-information assessments and fibered into congruent two-dimensional images $\overset{(0)}{\hat{t}}{}_{m(p)}^n$, which correspond to weak video-information evaluations. Here it should be taken into account that dimension of some set (manifold) means its decomposition on as small as possible by the subsets (fibers) of dimension $n-1$ and the impossibility of such split by subsets (fibers) of dimension $n-2$ [27]. Therefore, the 3D video-information evaluation formally could be presented as

$$\hat{t}_m^n = (P-1)\sum_{p=1}^{P} \overset{(0)}{\hat{t}}{}_{m(p)}^n \Rightarrow \bigcup_p \overset{(0)}{\hat{t}}{}_{m(p)}^n, \quad p=\overline{1,P} \tag{10}$$

Where

$\hat{t}_m^n$ – 3D video-information evaluation presented by fibration on evaluations congruent 2D video-information;

$\overset{(0)}{\hat{t}}{}_{m(p)}^n = \overset{(0)}{\hat{t}}{}_{(p)}^{mn}$ – current two-dimensional layer as one of evaluations congruent 2D video-information;

P – number of two-dimensional layers (rank) fibration space 3D video-information .

According to the expression (10) video-information evaluation is presented as manifold in the form of disjunctive union of fibers $\overset{(0)}{\hat{t}}{}_{m(p)}^n = \overset{(0)}{\hat{t}}{}_{(p)}^{mn}$ (in case of the evaluation weak video-information there is no need to distinguish co- and contravariant tensor components), forming its atlas [8], [9], [13], [27], [39].

Taking into account this fact, the current 2D fiber in the form of 2D evaluation weak video-information $\overset{(0)}{\hat{t}}{}_{(p)}^{mn}$ could be presented splitting to 1D evaluation video-information and therefore could take the form of

$$\overset{(0)}{\hat{t}}{}_{(p)}^{mn} = (Q-1)\sum_{q=1}^{Q} \overset{(00)}{\hat{t}}{}_{(q)}^{mn} \Rightarrow \bigcup_q \overset{(00)}{\hat{t}}{}_{(q)}^{mn}, \quad q=\overline{1,Q} \tag{11}$$

Where

$\overset{(00)}{\hat{t}}{}_{(q)}^{mn}$ – evaluation weak 1D video-information;

Q – quantity of one-dimensional layers (rank) of 2D video-information evaluation.

In its turn, each 1D video-information evaluation present as one-dimensional geometrical object (one-manifold), which could be presented as linear combination of elementary samples (quanta) of physical and structural video-data, considered together

$$\overset{(00)}{\hat{t}}{}_{(q)}^{mn} = (L-1)\sum_{l=1}^{L} \overset{(000)}{\hat{t}}{}_{\times(l)}^{mn} \Rightarrow \bigcup_l \overset{(000)}{\hat{t}}{}_{\times(l)}^{mn}, \quad l=\overline{1,L} \tag{12}$$

Where

$\overset{(000)}{\hat{t}}{}_{\times(l)}^{mn} = \bigcup_r d\left\{\overset{(0000)}{\hat{t}}{}_{E(r)}^{mn}\right\} \overset{(0000)}{\hat{t}}{}_{E(r)}^{mn}$, $r=\overline{1,R}$ – full video-data in the form of elementary physical and structural samples considered together;

L – quantity of components (rank) of 1D evaluation video-information;

$\overset{(0000)}{\hat{t}}{}_{E(r)}^{mn}$, $r=\overline{1,R}$ − structureless sample of physical video-data;

d − operator of differentiation.

Full sample of elementary video-data is presented as indivisible joint physical and structural video-information quantum underlying the base of video-information evaluations hierarchy (Fig. 5). This element is fundamental to provide an intellectual process of video-information evaluations forming and on the importance can be compared only with the string that is considered by modern physics as fundamental element of matter [9], [44].

Actually, at intellectual artificial vision based on AVIS video-informational processes are the processes of sequential formation of video-information evaluations upward hierarchy. In these processes evaluations of structural components of video-information are dominated, as their nature provide the technology of intellectual assessment for video-information. The physical components evaluations play an important supportive role to avoid video-information loss during its evaluation.

It should be noted, that local qualts (structure quanta) be formed action by differentiation operator's on the structureless sample of physical picture. Only in case of weak video-information outer differentiation operator acts like a common differentiation operator in the vicinity of the point coinciding with spatial coordinates of corresponding structureless sample on the physical picture

$$\overset{(000)}{\hat{t}}{}_{S(l)}^{mn} = d\left\{\overset{(0000)}{\hat{t}}{}_{E(r)}^{mn}\right\}, r=\overline{1,R} \tag{13}$$

Where

$\hat{t}^{mn\,(0000)}_{E(r)}$ − structureless sample of physical picture;

R − rank of the differentiation operator's "window" of physical picture.

The expression (13) refers to the low level of AVIS hierarchy and corresponds to the reformatting of structureless samples of physical picture into the local qualts.

Using the above formalism as a guiding principle can be built generic hypergraph of strong artificial intelligence in the form of a composition of interacted artificial intelligences of sensor and non-sensor nature (Fig. 5). Strong artificial intelligence of sensor nature can be considered in the form of a fibration on video-intelligence (artificial vision), audio-intelligence (artificial hearing), tactile intelligence (artificial touch), chemo intelligence (artificial smell) and many other artificial intelligences that use the available variety of sensors of physical nature. The strong artificial intelligence of non-sensor (structural, abstract) nature is a much more numerous ("broad"). It can be considered as a union of mathematical, physical, musical, engineering, driving, household and many other kinds of intelligences constructed (trained) on the basis of relevant structural (abstract) information. All these numerous partial intelligences are considered together, what allows to present strong artificial intelligence as

$$\chi^n_m = \alpha^n_m \circ \beta^n_m \tag{14}$$

Where

$\chi^n_m = \alpha^n_m \circ \beta^n_m$ – strong artificial intelligence in whole is formed as a composition partial strong artificial intelligences (of sensor and non-sensor nature considered together) of the brain-like system;

$\circ$ − mathematical symbol of composition;

α^n_m − partial strong artificial intelligence of sensor nature of rank N (including video-intelligence $\alpha^n_{m(3)}$, see Fig.5)

$$\alpha^n_m = \bigcup_z \alpha^n_{m(z)}, \quad z = \overline{1, N} \tag{15}$$

β^n_m − partial strong artificial intelligence of non-sensor nature of rank M

$$\beta^m_m = \bigcup_s \beta^n_{m(s)}, \; s = \overline{1, M} \;. \tag{16}$$

Informational approach to the strong artificial intelligence as to an informational category allows to consider uniformly its numerous partial intelligences of any nature united by its common attribute – information, as the information is a measure of strong interaction (interaction-measurement) comparable forms of matter. In doing so, the interacted forms of matter, e.g., physical (chemical) and structural (abstract) ones or, in a more general form, quantitative and qualitative ones should always be considered together. An abstract (image) thinking needed at complex and dynamically changing environment to assess the current and prospective state of the surrounding world underlies the strong artificial intelligence [45 – 48]. Strong artificial intelligence in a brain-like system acts as a backbone system-forming informational macro-object in the form of some complex qualitative measuring scale to obtain qualitative measurements on the ultra-high level of the system's hierarchy. This provides building a brain-like system for a numerous qualitative joint evaluations for informations of any nature which as a result with the need formed an understanding of the surrounding real World

On Fig. 5 is shown a generic hypergraph of strong artificial intelligence, in which one of artificial intelligences, namely, the video-intelligence, is shown with the unfolded internal structure of the ascending hierarchy of its evaluations of video-information. The initial fundamental feature of information, that the quantitative and qualitative information components are always considered together, is used when forming the hypergraph. It corresponds to the concept of a new (post Shannon) information approach under which the paradigm of information is united and dos not depend on of the nature of components for corresponding interactions-measurements which rank could be extremely large. Thus, not only natural advanced biological systems, but also technologically developed artificial ones (that could be formed as intelligent systems of different purposes) can have intelligence. Currently such systems most often implement intuition ways intellectual processing (evaluation) of information that have a heuristic nature and could be formed, e.g., on the base of various neural networks. However, times of "intelligent systems", which work could not often be explained even by its researcher, are passing. They must be replaced by truly intellectual brain-like systems adequate to intellect's complex nature that could be understood only by the use of current advances in basic science. This circumstance tightens the requirements for the educational level for of modern intelligent systems researchers, who should use fundamental science not as a philosophy, but as a necessary working tool.

Natural and artificial intellects the highest achievements of biological evolution and technological progress in the field of data processing systems. Moreover, in both cases it is the vision system, provided an adequate abstract world's reflection in virtual (mental) space of system, which generates an abstract thinking as high-effective data processing technology. An abstract thinking becomes the determining factor of natural intellect's evolution, that generated its pursuit of self-knowledge, by the creation of emulator an artificial intelligence including. However, an abstract thinking also underlies the artificial intellect technology. Therefore, an artificial intelligence could have unlimited evolutionary possibilities by preliminary training and a further self-training that would generate the new knowledge for all or separate areas of science. Furthermore, an advanced artificial intelligence is certainly able to effectively support a natural intellect (e.g. as a brain-like supercomputer) when solving critical tasks for humanity. In doing so, its simplified problem-driven versions could provide on a competitive basis an effective alternative to natural intellect when solving numerous secondary tasks from the society, industrial production and the surrounding material world.

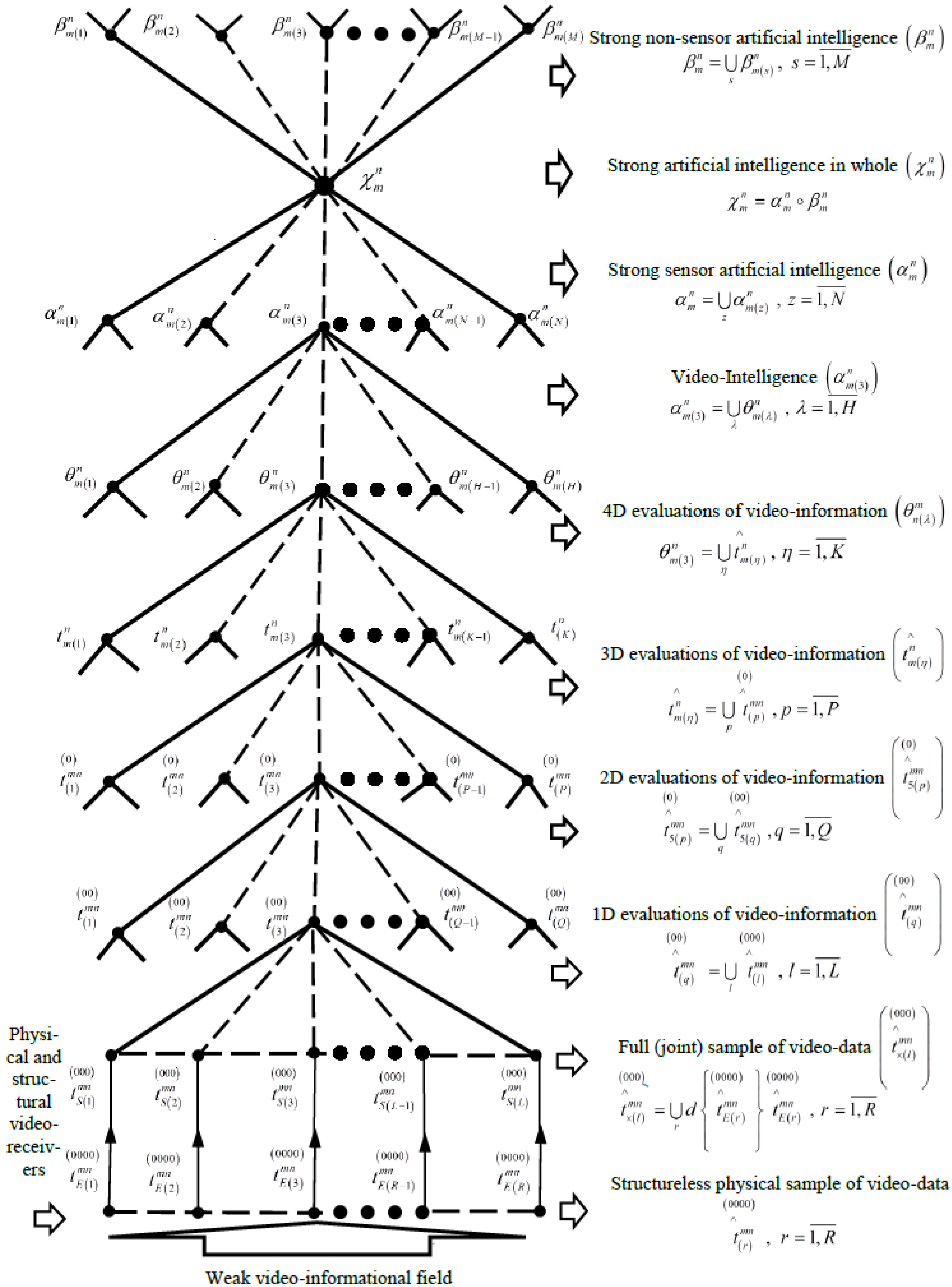


Fig. 5. Strong artificial intelligence's hypergraph in whole are regarded a composition the generalized hypergraphs of it non-sensory and sensory components in which as an example of a deeper structures being considered video-intelligence as hierarchy generalized hypergraphs of evaluations video-informations and video-data.

VI. Discussion

This article discusses the development of post-Shannon informational approach, created at "NeocorTek Lab" where this approaches was additionally adapted to the problem of video-information evaluation in artificial vision. NeocorTek Lab suggested of the productive informational approach-synthesis for creation of applications on the basis of artificial intelligence systems (including strong artificial one). Created approach (at the absolute abstraction from the concepts and technologies of neural networks) is self-contained and is able to provide an effective support to the processes of forming the truly intelligent systems for various applications. Such intelligent systems have structure that comparable to the structure of the human's neocortex and therefore they can be considered as brain-like systems.

Artificial intelligence, in our view, could be considered as an intelligence systems ability (e.g. of artificial vision similar to human's vision) to the forming of virtual "mental" space-time where all the information about the surrounding world could be adequately presented in the form of comparable and interacted "mental" formally images. Such approach allows to correctly setting and adequately solving the tasks of understanding concerning information.

Currently the systems of socalled "weak" artificial intelligence mainly aimed to solve the recognition tasks and based on neural networks technologies (deep and many others) are already quite widespread. According to the current informational approach-synthesis, these heuristic technologies do not correspond to the informational essence of intellectual mechanisms of recognition, perception and understanding of information. Therefore, informational approach at these systems almost does not work, especially as much critically important information is irretrievably lost at the entrance area of neural network already. The further compensation these fundamental losses of information by any heuristic method are impossible.

The main achievement of NeocorTek Lab which located in Russia is the creation of productive information theory of artificial intelligence which allows creating intelligence systems meaningfully and purposefully. Initial elements of this theory were demonstrated in the current article. In the process of in-depth study of artificial vision managed to open completely new principle of informational mechanisms of the artificial intelligence. It turned out that In the process of its synthesis at macro-level of consideration (the structural and functional level) objectively is forming structure similar the known structure neocortex of the human. In addition, it turned out that at the micro-level of consideration (the level of the element base); there is absolutely no need to emulate the neurobiological element base with the help of neural networks. This allows us to solve the problem of constructing an artificial neocortex using the traditional computer element base, strictly preserving only the structural and functional similarity with the human neocortex.

In Russia there are practically no players is actively investing in artificial intelligence as a market segment. Therefore, NeocorTek Lab appeals to representatives of the international business community who are interested in the development of technologies for strong artificial intelligence, with the proposal of partnership, the options of which can be considered and discussed.

VII. Conclusion

1) Currently, a significant increase in the role of geometrical methods in various fields of natural and applied sciences could be explained by the fact that mathematical analysis methods provide a quantitative characteristic of phenomena "in a small" using the infinitesimal changes their parameters. Geometrical methods, in its turn, allow obtaining the similar description of these phenomena "in general", using the large values changes of the same parameters.

2) The image forming process could be considered according to gauge theory as gauge transformation which is realized by means of the analysis of a weak video-information field representing a gauge field. Such transformation does not reflect on the observed (measurable) characteristics of physical object and ensures the adequacy of the image and of the physical component evaluation of weak video-information.

3) Gauge transformation is an over-coordinate one as does it not act on vectors and tensors of internal characteristics spaces of video-informational fields. Therefore, the geometry defined by gauge transformations is the geometry is not of space-time but of a more general kind of space - of a fibration space.

4) There are some analogies with the psychology of human's visual processes that could be used in artificial vision. In particular, the work result of artificial vision system could be considered in the form of topical visual field in the internal (virtual, "mental") space-time of this system. This internal space and physical space-times are in the second-order isomorphism relation. At this "mental" space the representation of video-information evaluation is a reconstructed evaluation of video-information of general view, adequately collated to space-time content of objects observed in the corresponding spatial scene of the surrounding real world

5) At the artificial vision process form the hierarchy of video-information assessments as an upward hierarchy of 1D, 2D, 3D and 4D abstract images (representations) of the real world, which providing in the aggregate the harmoniousness of visual process, which physical and structural properties are considered together from the first stage of this hierarchical process of "mental" reconstruction of video-information.

6) The main function of the hierarchy of video-thesauruses, at an highest level of the hierarchy of which the video-intelligence is located, is phased "strengthening" of the evaluations of video-information from a weak (rough) topology to a strong (thin) topology, which is carried out by the ascending hierarchy of AVIS.

7) At the "mental" (psychological) AVIS hierarchy level, video-intelligence is a system-forming macro-object which is considering as an ultra-high level of AVIS informational hierarchy. In set-theoretical form, video-intelligence could be considered as an aggregation of pairwise interacting 4D video-information evaluations or taking into account the law of idempotency as union these evaluations of video-information.

When using the common template of informational approach, video-intelligence is presented as video-informational 4D macro-object (a measuring scale for hierarchy's ultra-high level), which used to support the interactions-measurements in "mental" 4D space of AVIS. The process is building of corresponding abstract images forming by video-intelligence could be interpreted as image thinking and it result will be an understanding of video information about the spatial content of the surrounding real world

8) Video-thesauruses (1D, 2D, 3D) in the forms of corresponding "dimensional" scales of AVIS up-ward hierarchy are formed with vision system by sequential accumulation of 1D, 2D, 3D video- experience.

9) Video-thesaurus development scheme is formed as a sequence of hypergraphs corresponded to video-thesaurus sequential development as a qualitative measuring scale. These heterarchical organizational of the video-thesaurus structures, should be formed during AVIS training process and her are rather difficult to implement. However, taking into account the law of idempotency such union of combinatorically paired organized "dimensional" video-data could degenerate into much simplier empirical qualitative measuring scale of video-structures. This scale takes the form of a fairly simple realizable generalized information hypergraph of video-thesaurus. Video-thesaurus has the structure of combinatorically organized binary system of relations between video-thesaurus elements. The theory of physical structures currently proves absent informative nontrivial theories of ternary, tetradic and other systems of relations (structures). There are only binary and unary systems of relations occurring in the nature. Thus, the considered way of development of video-intelligence by accumulating a video- experiment is the only possible.

*10)*The formalism of generalized information hypergraphs allows us to construct generalized hyper-graphs of video intelligence and strong intelligence, which in turn can be represented by a composition of generalized information hypergraphs of strong artificial intellects of sensory and non-sensory natures.

*11)*As an example, on Fig. 5 shows the structure of a simplified hypergraph of strong artificial intelligence. Here one of the partial sensory intellects, namely, video-intellect, is shown with the unfolded internal structure of the upward hierarchy of hypergraphs of the video-informations evaluations of small spatial dimensions. This hypergraph demonstrates the possibilities of the process of developing strong artificial intelligence, but not him by "depth", but by "width." Indeed, the process of developing a strong artificial intelligence is possible only on the basis of training and self-training processes, which necessarily lead to increase in the number of elements in all corresponding thesauruses and the artificial intellects (sensory and non-sensory). It is the quantitative expansion of the element composition of all these information macro-objects that is the only way of qualitative development of strong artificial intelligence.

References

[1] E.M. Yarichin, "Informational Paradigm of Technical Vision, Pattern Recognition and Image Analysis", Vol.18, № 1, 2008, pp. 163 – 176.

[2] E.M. Yarichin, "Theory of Full (Video) Information: Bridging of the Gap between Quantity and Quality in Modern Paradigm of Information", International Journal of Information Science., vol. 3, no 3, 2013, pp.37 – 58.

[3] E.M. Yarichin, "Informational Mechanism of Intellectual Technical Vision", IEEE International Siberian Conference on Control and Communications, IEEE Xplore, Digital Library, 2011, pp.84 – 86.

[4] E.M. Yarichin, "Video-information: Internal Structure and the Features of the Discretization", Inter-national Journal of Engineering and Innovative Technology (IJEIT), vol. 3, no 5, 2013, pp.133 – 140.

[5] L.D. Landau, E.M. Liphshitz, Theoretical physics. Vol. II. Field theory. Moscow: Nauka, 1988 (in Russian).

[6] A.A. Logunov, Gravity field theory. Moscow: Nauka, 2001 (in Russian).

[7] D.D. Ivanenko, G.A. Sardanashvili, Gravity. 2nd ed. Moscow: Editorial URSS, 2004 (in Russian.

[8] N.P. Konoplyova, V.N. Popov, Calibrating fields. 3d ed. Moscow: Editorial URSS, 2000 (in Russian).

[9] R. Penrose, The Road to Reality. A Complete Guide to the Laws of the Universe, Jonatan Cape, London, 2004.

[10] A.I. Berg, et. al., Control, information, intellect. Moscow: Mysl, 1976 (in Russian).

[11] H. Haken, Information and Self-organization. A Macroscopic Approach to Complex Systems, Second Enlarged Edition, Springer-Verlag Berlin Heidelberg New York, 1999,

[12] F.P. Tarasenko, Introduction to the course of information theory. Tomsk: Izdatelstvo Tomskogo Universiteta, 1963 (in Russian).

[13] B.N. Petrov, I.I. Goldenblat, G.M. Ulanov, S.V. Ulyanov, Problems of relativistic and quantum dy-namical systems control (physical and informational aspects). Moscow: Nauka, 1982 (in Russian).

[14] I.M. Kogan, Applied information theory. Moscow: Radio i svyaz, 1981 (in Russian).

[15] I.Ya. Kremer, A.I.Kremer, V.M. Petrov, Space-time processing of signals. Moscow: Radio I svyaz, 1984 (in Russian).

[16] O.N. Litvinenko, Fundamentals of radio optics. Kiev: «Tekhnika», 1974 (in Russian).

[17] V.A. Zverev, Radio optics (Transformation of signals in radio- and optics). Moscow: «Sov.radio», 1975 (in Russian).

[18] W. Thurston, D. Wicks, "Mathematics of three-dimensional manifolds", «In the world of science», Scientific American. Russ. ed., Moscow: Mir, 1984, no.9, pp.74-88 (in Russian).

[19] P. Scott, The Geometries of 3-Manifolds, London Mathematical Society, 1983.

[20] W.P. Thurston, Three-Dimensional Geometry and Topology, Edited by Levy S., Princeton, New Jersey, 1997.

[21] D. Marr, VISION. A Computational Investigation into the Human Representation and Processing of Visual Information, W.H. Freeman and Company, New York, 1982.

[22] B.K.P. Horn, Robot Vision, the MIT Press Cambridge, Massachusetts London, England, 1986.

[23] J.M. Lloyd, Thermal Imaging Systems, Plenum Press New York and London, 1975.

[24] G.S. Safronov, A.P. Safronova, Introduction to radio holography. Moscow: «Sov. radio», 1973 (in Russian).

[25] L.M. Soroko, Holography and coherent optics basics, Moscow: Nauka, 1971 (in Russian).

[26] P. Greguss, Ultrasonic imaging. Seeng by Sound, Focal Press Limited, London, Focal Press Inc., New York, 1980.

[27] Yu.G. Borisovich, N.M. Bliznyakov, Ya.A. Izrailevich, T.N. Fomenko, Introduction to topology. Moscow: Vyssh. shkola, 1980 (in Russian).

[28] E.Yu. Kondrashina, L.V. Litvintseva, D.A. Pospelov, Knowledge of time and space in intelligent sys-tems. Moscow: Nauka, 1989 (in Russian).

[29] J. Pfanzagl, in cooperation V. Baunann, H. Huber, Theory of Measurement, 2nd revised edition, Physica-Verlag, Wurzburg – Wien, 1971.

[30] J. Celko's, Joe Celko's SQL Programming Style, Elsevier Inc., 2005.

[31] V.V. Druzhynin, D.S. Kontorov, Systematology problems (problems of complex systems theory). Moscow: «Sov. radio», 1976 (in Russian).

[32] M.A. Kholodnaya, Psychology of intellect. Paradoxes of research. 2nd ed. Saint Petersburg: Piter, 2002 (in Russian).

[33] R. Klatzy, Human Memory, W. H. Freeman and Company, San Francisco, 1975.

[34] R. Utiyama, What did physics come to (From the theory of relativity to calibrating fields theory). Russ. ed. Moscow: Znanie, 1986 (in Russian).

[35] S. - K. Chang, Principles of Pictorial Information Systems Design, Prentice-Hall International, Inc., 1989.

[36] Yu.S. Vladimirov, Metaphysics. Moscow: BINOM. Laboratory of knowledge, 2002 (in Russian).

[37] Yu.S. Vladimirov, Geometrophysics. Moscow: BINOM. Laboratory of knowledge, 2005 (in Russian).

[38] H.R. Schiffman Sensation and Perception, An Integrated Approach, Fifth Edition, John Wiley & Sons, Inc., 2001.

[39] B.A. Dubrovin, S.P. Novikov, A.T. Fomenko, Current geometry: methods and applications. 2nd ed. Moscow: Nauka, 1986 (in Russian).

[40] A.A. Logunov, M.A. Mestvirishvili, The relativistic theory of gravity. Moscow: Nauka, 1989 (in Russian).

[41] A.A. Logunov, Gravitational field theory, Moscow: Nauka, 2001 (in Russian).

[42] Yu.I. Kulakov, Theory of physical structures. Moscow, 2004 (in Russian).

[43] V.P. Sigorskiy, Mathematical tool of an engineer. Kiev: "Tekhnika", 1975 (in Russian).

[44] Sh.-T. Yau, S. Nadis, The Shape of Inner Space, String Theory and the Geometry of the Universe's Hidden Dimensions, A Member of the Perseus Books Group, New York, 2010.

[45] J. Hawkins, S. Blakeslee, On intelligence, Times Books Henry Ford and Company, New York, 2004.

[46] R. L. Gregory, The Intelligent Eve, Wiedenfeld and Nicolson, London, 1970.

[47] R. Penrose, The Emperor's New Mind. Concerning Computers, Mind and the Laws of Physics, Oxford University Press, 1989.

[48] R. Penrose, Shadows of the Mind. A Search for the Missing Science of Consciousness, Oxford University Press, 1994

Anomaly Detection with Machine Learning and Graph Databases in Fraud Management

Shamil Magomedov[1], Sergei Pavelyev[2], Irina Ivanova[3], Alexey Dobrotvorsky[4], Marina Khrestina[5]
Department «Automated control systems»
MIREA – Russian Technological University
Moscow, Russian Federation

Timur Yusubaliev[6]
Quality Software Solutions ltd Moscow, Russia

***Abstract*—In this paper, the task of fraud detection using the methods of data analysis and machine learning based on social and transaction graphs is considered. The algorithms for feature calculation, outlier detection and identifying specific sub-graph patterns are proposed. Software realization of the proposed algorithms is described and the results of experimental study of the algorithms on the sets of real and synthetic data are presented.**

Keywords—Data analysis; machine learning; graph database; fraud detection; anti-money laundering

I. INTRODUCTION

At present fraud is a major threat that is increasing every year. The global economic crime survey of 2018 carried out by PricewaterhouseCoopers [1] found that almost half (49%) of the 7,200 companies they surveyed had experienced fraud of some kind. Experts from HSN Consultants predict online credit card fraud to soar to $32 billion in 2020 [2]. Beside direct financial losses, fraud also affects customer loyalty and conversions in both digital and physical environments. For instance, 20% of customers change their banks after experiencing frauds. Meanwhile, manual review remains prevalent among the means of fraud detection. According to the annual Fraud Benchmark Report by CyberSource [3] 79% of North American businesses conduct manual reviews, and on average, these businesses manually review 25% of orders. At the same time, the survey found that these businesses accepted 89% of orders following manual review. This means that more orders are subject to manual reviews than might be necessary. Since manual review is usually the costly aspect of fraud management operations, automated screening could make fraud management processes more efficient by leaving only the most suspect orders to manual reviews.

Machine Learning technologies have shown their effectiveness in solving such tasks as spam detection, image recognition, product recommendation, predictive analytics etc. In fraud management, Machine Learning can be used to predict fraud in a large volume of transactions by applying cognitive computing technologies to raw data. The prediction problem can be further divided into two types of tasks: classification and regression. Regression analysis is a popular, longstanding statistical technique that measures the strength of cause-and-effect relationships in structured data sets. Regression analysis tends to become more sophisticated when applied to fraud detection due to the number of variables and size of the data sets. It can provide value by assessing the predictive power of individual variables or combinations of variables as part of a larger fraud strategy. According to this technique, the authentic transactions are compared with the fraud ones to create an algorithm, which will then predict whether a new transaction is fraudulent or not. Classification problem can be solved with the help of Decision Tree algorithms. They are essentially a set of rules that are trained using examples of fraud that clients are facing. The creation of a tree ignores irrelevant features and does not require extensive normalization of the data. By inspecting a tree, it is possible to understand why a decision was made by following the list of rules triggered by a certain customer. Random Forest technique uses a combination of multiple decision trees to improve the performance of the classification or regression. It allows smoothing the error that might exist in a single tree and increases the overall performance and accuracy of the model while maintaining the ability to interpret the results and provide explainable scores to the users. Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data. Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data and that they may over-fit data sets that are particularly noisy. Neural networks can be an excellent complement to other techniques, which improves with exposure to data. The neural network is a part of cognitive computing technology where the machine mimics how the human brain works and how it observes patterns. Neural networks can adapt to the change in the behavior of normal transactions and identify new patterns of fraud transactions. Data processing by neural networks is extremely fast which makes it possible to make decisions in real time.

Due to growing popularity of machine learning, many innovative enterprises are starting to implement these techniques in their fraud management processes. For example, PayPal uses a homegrown AI engine built with open-source tools to detect suspicious activity, and more importantly to separate false alarms from true fraud [4]. PayPal implements express assessment using linear models to separate uncertain transactions from ordinary ones. Then, all transactions that look suspicious are run through an ensemble of three models comprising a linear model, a neural network, and a deep neural network. The three models then vote to arrive at the result with the higher accuracy. With the help of this human and AI solution, PayPal has decreased its false alarm rate to half. MasterCard integrated machine learning and AI to track and process such variables as transaction size, location, time, device, and purchase data [5]. The system assesses account behavior in each operation and provides real-time judgment on whether a transaction is fraudulent. The project aims at reducing the number of false declines in merchant payments. Feedzai [6], a FinTech company, claims that a fine-tuned machine learning solution can detect up to 95% of all fraud and minimize the cost of manual reconciliations, which accounts now for 25% of fraud expenditures. Capgemini [7] claims that fraud detection systems using machine learning and analytics minimize fraud investigation time by 70% and improve detection accuracy by 90%. These facts prove the benefits of using machine learning in anti-fraud systems. On the other hand, banks have been slow to adopt machine learning and AI solution at a large scale. The reasons for this include high infrastructural costs, strict regulations and risk of replacing existing technology.

Machine Learning technologies also have their own limitations. One of such limitations is their blindness to connections in data when the initial data set is relatively small. Machine learning models work on actions, behavior, and activity. For example, the model can overlook a seemingly obvious connection such as a shared card between two accounts. To counter this machine learning models can be enhanced with Graph databases. Graph database addresses Gartner's fifth layer of fraud prevention: entity link analysis [8]. Graph database allows looking beyond the individual data points of discrete analysis to the connections that link them. Thus, graph technique can find multiple bogus actors for every single one prevented through scoring. Graph databases allow blocking suspect and bogus accounts before they have taken any fraudulent action. Another important trait that makes graph database a valuable addition to any fraud prevention solution is its inherent speed in calculating relationships. Since the relationships in graph database are treated with as much value as the database records themselves, the engine that navigates the connections between nodes can do so efficiently, enabling millions of connections per second. Graph database enables quick extraction of new insight from large and complex databases to help uncover unknown interactions and relationships.

II. Anomaly Detection Algorithms for Graph Structures

A. Local Outlier Factor (LOF) Algorithm based on Local-Sensitive Hashing (LSH) Method

LOF is an outlier detection algorithm that calculates certain numeric value for each point, which allows identifying the point as normal or anomaly. LOF value close to one corresponds to normal points; otherwise, the points are considered anomalies. Exact threshold for anomaly detection is set after conducting data analysis. An algorithm for calculation of LOF based on LSH method has been developed. Pseudo code for the developed algorithm is listed below:

```
Input: points // set of all points
Output: result // set of nearest neighbors for each point
result = 0
nv = StartNumVectors
// creation of hash-table
hash_table = 0
for 1..NumTables do
    // clearing hash-table for each iteration
    hash_table = 0
    hash_vecs = get_random_vecs(nv)
    forall p ∈ points do
        hash = get_hash(hash_vecs, p)
        hash_table[hash] += p
    foreach cell ∈ hash_table, cell.size < const do
        forall point ∈ cell do
            result[point] += all points ∈ cell without the point
    forall point ∈ result do
        result[point] = save_only_best_kNN_neighbors
    nv = nv + 1
foreach cell ∈ hash_table, cell.size > const do
    forall point ∈ cell do
        result[point] = save_any_kNN_neighbors ∈ cell
forall point ∈ result do
    result[point] = save_only_best_kNN_neighbors
foreach point : result[point].size < kNN do
    result[point] = brute_force[point]
```

LSH method was chosen due to necessity of fast identification of nearest neighbors for each point. The main principle of LSH method involves using special method of hashing when hash values are equal for the points close to each other. A set of hash-tables (*NumTables* = 100) has been generated with a set of vectors for each of those tables. For the first table a number of random vectors (*StartNumVectors* = 3) has been generated. For each subsequent table the number of random vectors increases by one with each iteration (number of iterations corresponds to the number of tables, one table is processed per iteration). Each vector consists of d random values generated according to normal distribution with mean value of zero and standard deviation of one. Hash is calculated for each point as a bit sequence where bit i equals one if scalar product of vector i of the processed table and the vector corresponding to the point in question (the point can be considered a vector) is equal or greater than 0. Otherwise bit i equals zero. Thus in each iteration of the algorithm all points are distributed among the cells of current hash-table.

Then the cells of current hash-table are considered with the size lower than *const* (4**kNN* in this example, where *kNN* is the number of nearest neighbors). For each point in such cells, all other points in this cell are added to the set of candidates for the point in question. Finally, for each point, duplicate candidates are removed and the nearest *kNN* neighbors are left.

At the last iteration, the cells with the size larger than *const* are considered and for each point in such cells *kNN* random points from the same cell are selected, duplicate candidates for each point are removed and the best *kNN* candidates are left. For each point, random *kNN* candidates are selected, because when the number of random vectors used for calculating hash is large enough each cell corresponds to a small part of n-dimensional space, which means all points from the same cell are close to each other. In the end for each point with the number of neighbors lower than *kNN* a naïve algorithm is used since the number of such points at the last iteration should be very small. In current realization, *kNN* was set equal to 10.

B. "Volcano" and "Black Hole" Patterns

So-called "volcano" and "black hole" patterns were described in [9]. "Black hole" refers to a sub-graph, which has only incoming edges from the vertices of the graph not included in this sub-graph. "Volcano" refers to a sub-graph which has only outgoing edges to the vertices of the graph not included in this sub-graph. The task of identifying "volcanoes" is inverse to the task of identifying "black holes".

An example of a "volcano" and a "black hole" is shown in Figure 1.

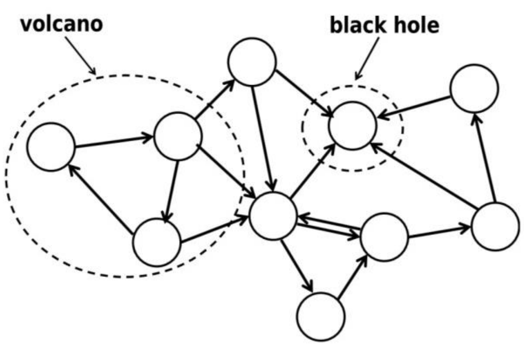


Fig. 1. An example of a "volcano" and a "black hole".

The task of identifying "volcanoes" and "black holes" is a combinatorial problem. In [9] two algorithms based on pruning schemes are proposed.

C. Algorithm for Identifying "Volcanoes" and "Black Holes"

An algorithm for identifying "black holes" has been developed. For identifying "volcanoes", the direction of all edges of the graph should be reversed and the same algorithm should be applied.

The set of *ancestors* for the vertex *v* is defined as a set of all vertices having at least one edge outgoing to *v*, as well as all ancestors of those vertices. The algorithm identifies "black holes":

- with diameter equal or lower than (*MaxIterCount*) and
- with the number of ancestors for each vertex in the sub-graph lower than *MaxSetSize* (considering only ancestors at the distance equal or lower than *MaxIterCount*).

The algorithm is described using vertex-centric [10] approach, but practical realization is carried out using resilient distributed dataset (RDD) API [11] without Pregel API in Apache Spark [12-15]. The algorithm is iterative with maximum number of iterations defined by *MaxIterCount*.

1) Description of Handler Function

Each vertex has a local buffer *send_buf*, which is cleared each time the handler is called. At the start of the algorithm, each vertex receives the message *init*, which is necessary for initialization of the algorithm. On receiving the message *init* the vertex puts its identifier *id* into the buffer *send_buf*.

Each vertex has its own set of ancestors – *ancestors*. Initially the set *ancestors* is empty for each vertex. Each vertex also has a flag *bad*, initially set to value *false*. This flag is set to value *true* if the vertex has too many ancestors. At each iteration, a vertex can acquire new ancestors in incoming messages. Acquired ancestors are added to the set *ancestors* if the size of resulting set does not exceed *MaxSetSize*, otherwise the set *ancestors* does not update and the flag *bad* is set to value *true*.

2) Description of message-sending operation

At each iteration of the algorithm after handler function finishes its work follows the message-sending step. At this step, all triplets of the graph corresponding to its edges are considered, and a message is sent along each edge.

For each triplet (edge):

- if the buffer *send_buf* of the initial vertex of the edge is empty or the flag *bad* of the terminal vertex equals *true*, then nothing should be sent along this edge;
- otherwise all elements from the buffer *send_buf* which are not present in the set *ancestors* of the terminal vertex are sent within a message along this edge.

3) Detection of "black holes"

In the end, each vertex owns a set including a number of its ancestors. A set of vertices *B* with a common ancestor *X* is considered a "black hole" if:

- *X* is not included in the buffer *send_buf* of any vertex from the set *B*. Otherwise some vertex could send *X* at the last iteration, which means there could be an outgoing edge from *B*, which contradicts the definition of a "black hole".
- *X* does not belong to the set ancestors of the initial vertex of each edge incoming to a vertex with the flag *bad* set to value *true*, and *X* is not the initial vertex of such edge. Otherwise (if such edge existed) some vertex (corresponding to such edge) could send *X* at the last iteration, but due to the defined limitations further transmission through the vertex with the flag *bad* set to value *true* would be impossible, and identifier *X* would not be able to reach any of the buffers *send_buf*.

Then the vertices should be grouped according to ancestors. In the Apache Spark realization, each pair (vertex, ancestor) should be mapped to a pair (ancestor, vertex) and then the pairs should be grouped by key (operations *map* and *groupBy*). Defined limitations guarantee that the sets of vertices acquired by the method described above does not include any outgoing edges.

An important feature of the developed algorithm is the ability to identify intersecting "black holes". Pseudo code for the developed algorithm is listed below:

MaxSetSize = 100 // constant limiting the size of sets ancestors and send_buf
MaxIterCount = 10 // constant defining the number of iterations
// vertex handler
(1) **procedure** handler(v: vertex, msgs: Vector[Long])
// message sending
(2) **procedure** sendMsg(triplet: EdgeTriplet)
v.send_buf = 0
// if initial message
if *msgs.size* = 1 *and msgs*(0) = -1 **then**

```
        vert.send_buf.add(vertex.id)
        return
if vert.ancestors.size + msgs.size ≤ MaxSetSize then
        // update set of ancestors
        vert.ancestors.add(msgs)
else
        vert.bad = true
        return
if msgs.size ≤ MaxSetSize then
        vert.send_buf.add(msgs)
else
        vert.bad = true
if !(triplet.srcAttr.send_buf.isEmpty or triplet.dstAttr.bad) then
        msgs = Vector[Long]()
        for x ← triplet.srcAttr.send_buf do
                if !triplet.dstAttr.ancestors.contains(x) then
                        msgs.add(x)
        send msgs along triplet
// merging messages
(3) function mergeMsg(a: Vector[Long], b: Vector[Long])
        // concatenation of two vectors
        return a + b
// main procedure
(4) procedure detect_blackholes(graph: Graph)
```

III. Experimental Study

Experimental study has been carried out on a set of data consisting of the database of all transactions (including 781 440 transactions and 15 034 710 involved entities) and the database of suspicious transactions (including 715 transactions and 349 involved entities). For evaluation of computational performance and scalability of the developed algorithms synthetic Erdos-Renyi [16] graphs of different sizes have also been used.

Software realization of the developed algorithms has been tested on a computational cluster consisting of 8 nodes connected with 1GBit Ethernet, each node running 8-core 2.2GHz E5-2660 processor, 64GB DDR3 memory, operation system SLES 11 SP4, Apache Spark 2.1.1 and Scala compiler sbt 0.13.13.

Software realization of the machine learning algorithm has been carried out using Scala, Spark 2.1.0 language and includes the following steps:

- Data input (database of all transactions, database of suspicious transactions and configuration data).
- Search for suspicious transaction in the database of all transactions (by comparing the fields DATA, SUME, ACC_B0 and Date, RealQty, AccClientOtpr accordingly).
- Creation of a graph with entities as its vertices and transactions as its edges (entities being taken from the database of all transactions).
- Selection of a set of edges for machine learning consisting of all suspicious transactions and the same number of normal transactions (selected randomly).
- Formation of egonets around the vertices for calculation of features.
- Calculation of features for each edge (transaction) (32 features total): amount and time of transaction; for sender and recipient egonets - minimum, maximum and average degrees, indegrees and outdegrees of vertices, number of vertices, “volcano” vertices, “black hole” vertices and other vertices; number of transactions and total transaction amount.
- Machine learning: method – random forest; MinMaxScaler method used to project each feature into [0, 1] line; random division of the learning set – 70% of samples for training (using cross-validation), 30% of samples for testing.

Graph created in the experimental study consists of 114 791 vertices (entities) and 781 440 edges. The learning set for machine learning includes 1430 transactions (50% normal and 50% suspicious). Training set consists of 993 transactions – 477 class 0 objects (normal transactions) and 516 class 1 objects (suspicious transactions). Testing set consists of 437 transactions – 238 class 0 objects (normal transactions) and 199 class 1 objects (suspicious transactions).

Resulting classification accuracy (share of correctly classified objects) reached 97.7%.

A measure of importance for each feature has been calculated. Top five most important features (listed with their respective “coefficient of importance”; the sum of coefficients for all features amounts to one) turned out to be the following:

- Amount of transaction – 0.38
- Degree of the sender vertex – 0.12
- Total amount of transactions (incoming and outgoing) corresponding to the sender – 0.09
- Number of vertices in the sender’s egonet – 0.07
- Outdegree of the sender vertex – 0.07

It is evident that amount of transaction has the greatest influence on classification results.

Other metrics of classification quality calculated during the experimental study (the closer to 1.0 the better):

- AUROC – 0.999
- Sensitivity (also known as “Recall”) – 1.0
- Specificity – 0.958
- Precision – 0.952
- NPV (negative predictive value = number of correctly classified normal transactions divided by total number of transactions classified as normal) – 1.0
- F1 score – 0.975

Sensitivity equaling 1.0 means that the algorithm did not miss any suspicious transactions within the testing set. This metric is especially important, since in anti-money laundering tasks it is necessary not to miss any fraudulent transactions.

False positive rate (FPR = (1-Specificity)*100%) on the testing set amounted to 4.2%, meaning that among 238 normal transactions 10 were falsely classified as suspicious.

Recent studies using a Random Forest Classifier in order to predict fraudulent transactions on raw data [17] have shown results for such metrics of classification quality as Precision, Recall, F1-score and AUROC all below 0.9 in most cases. Thus, it can be concluded that using the developed set of features allows improving the classification quality.

Additional experiment has been carried out using the database of all transactions excluding the suspicious ones as the testing set (780 725 transactions total). The share of transactions classified as suspicious equaled 4.9%. It is worth noting that the share of transactions with amount exceeding 100 000 classified as suspicious equaled 3.2% of the total number of transactions.

Quality of LOF algorithm based on LSH method is shown in Table I.

TABLE I. TABLE TYPE STYLES

THR	**FN / P**	**FP / N**
2.5	8.03%	1.2%
3	10.27%	0.9%
3.5	12.13%	0.8%

Here THR defines threshold value for anomaly detection, FN – number of false-negative objects, FP – number of false-positive objects, P – total number of anomalies, N – total number of normal objects. Results were calculated for 6000 random points corresponding to the edges of the real graph. Though the classification quality is lower in comparison with the trained random forest method applied above, the advantage of this approach is the ability to carry out fraud detection without prior knowledge of transaction history.

Using the algorithm for identifying “volcanoes” and “black holes” an edge was considered an anomaly if its initial vertex belonged to any of 5% largest “volcanoes” or “black holes”. Figure 2 shows the distribution of identified “volcanoes” and “black holes” for the real graph.

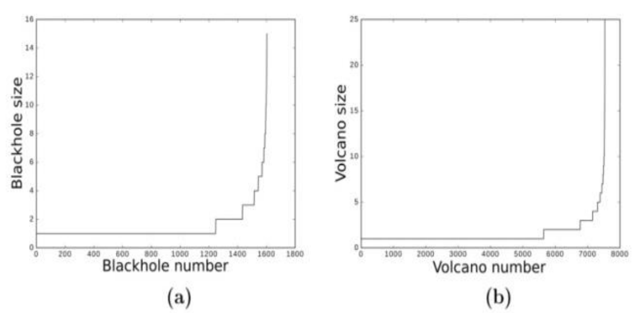


Fig. 2. Distribution of identified “volcanoes” and “black holes”.

Figure 3 illustrates strong scalability of the developed feature calculation algorithm, LOF algorithm based on LSH method and the algorithm for identifying “volcanoes” and “black holes”.

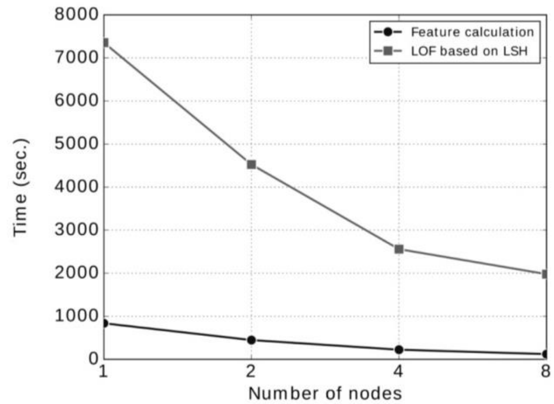


Fig. 3. Scalability of the developed algorithms.

Scalability of feature calculation algorithm was calculated for a synthetic Erdos-Renyi graph with 2^{19} vertices and 2^{22} edges. Scalability of LOF algorithm based on LSH method was calculated for the real graph. Scalability of the algorithm for identifying “volcanoes” and “black holes” was calculated for a synthetic Erdos-Renyi graph with 2^{21} vertices and 2^{24} edges.

Figure 4 illustrates the relative speedup of data processing with the developed algorithms based on the number of computational nodes involved in the processing of data.

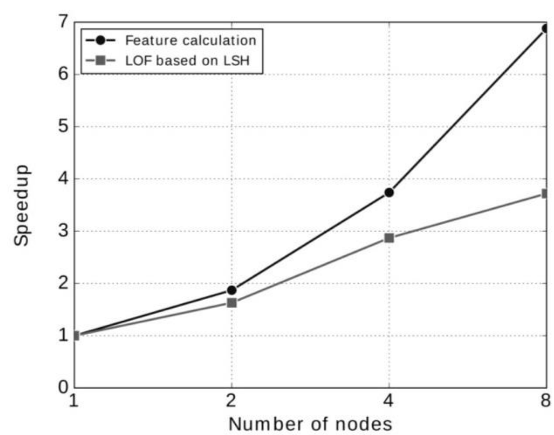


Fig. 4. Speedup of data processing according to the number of computational nodes.

IV. CONCLUSION

It can be concluded that the developed algorithm for feature calculation can be successfully implemented in conjunction with common machine learning methods to achieve high values of classification quality metrics. In particular, high value of sensitivity (and subsequently low False Negative Rate value) is important for anti-money laundering and other fraud management processes.

The results of experimental study have shown that the developed algorithms for anomaly detection demonstrate classification quality comparable with a trained random forest method when applied on a real transaction graph. This makes it possible to implement these algorithms in situations when prior knowledge of transaction history is not accessible, as well as for identification of new and unknown methods of fraud.

Software realization of the developed algorithms demonstrated high scalability, which makes it possible to significantly increase their performance using multi-processor computational clusters.

The developed algorithms are intended to be used in the framework of a software complex for automated fraud management in different areas of business.

Acknowledgment

The research is being conducted with the financial support of the Ministry of Education and Science of the Russian Federation (Contract №14.574.21.0142) Unique ID for Applied Scientific Research (project) RFMEFI57417X0142. The data presented, the statements made, and the views expressed are solely the responsibility of the authors.

References

[1] D. Lavion et al., PwC's Global Economic Crime and Fraud Survey, 2018, URL: https://www.pwc.com/gx/en/forensics/global-economic-crime-and-fraud-survey-2018.pdf (retrieved 28 August 2018).

[2] The Nilson Report, 2016, issue 1096, URL: https://nilsonreport.com/upload/content_promo/The_Nilson_Report_10-17-2016.pdf (retrieved 01 October 2018).

[3] 2017 North America Online Fraud Benchmark Report, URL: https://www.cybersource.com/content/dam/cybersource/2017_Fraud_Benchmark_Report.pdf (retrieved 01 October 2018).

[4] A. Chelsea, PayPal's history of fighting fraud, Fin Newsletter, URL: https://fin.plaid.com/articles/paypals-history-of-fighting-fraud (retrieved 01 October 2018).

[5] M. Cochrane, How MasterCard is using AI to improve the accuracy of its fraud protection, Business Insider, 2017, URL: https://www.businessinsider.com/mastercard-artificial-intelligence-fraud-protection-2017-1 (retrieved 01 October 2018).

[6] Demystifying machine learning for banking, URL: https://hollandfintech.com/wp-content/uploads/2018/01/Feedzai-Demystifying-Machine-Learning-for-Banking-2017.pdf (retrieved 01 October 2018).

[7] Next-generation fraud management solutions, URL: https://www.capgemini.com/wp-content/uploads/2017/07/next-generation_fraud_management_2017.pdf (retrieved 01 October 2018).

[8] C. Pettey and R. Van der Meulen, Gartner Says a Layered Fraud Prevention Approach Can Thwart Malicious Attacks, Gartner, 2009, URL: https://www.gartner.com/newsroom/id/1254413 (retrieved 01 October 2018).

[9] Z. Li, H. Xiong, and Y. Liu, Detecting Blackholes and Volcanoes in Directed Networks, CoRR, 2010, URL: https://arxiv.org/pdf/1005.2179.pdf (retrieved 23.11.2016).

[10] R.R. McCune, T. Weninger, and G.R. Madey, Thinking Like a Vertex: a Survey of Vertex-Centric Frameworks for Distributed Graph Processing, CoRR, 2015, URL: http://arxiv.org/abs/1507.04405 (retrieved 23.11.2016).

[11] M. Zaharia, M. Chowdhurr, M.J. Franklin et al., Spark: Cluster Computing with Working Set, HotCloud, 2010, vol. 10, p. 7.

[12] M. Armbrust, T. Das, A. Davidson et al., Scaling spark in the real world: performance and usability, Proceedings of the VLDB Endowment, 2015, issue 8, vol. 12, pp. 1840-1843.

[13] Voit A., Stankus A., Magomedov Sh., Ivanova I. Big data processing for full-text search and visualization with elasticsearch International Journal of Advanced Computer Science and Applications. 2017. T. 8. № 12. C. 76-83. DOI: 10.14569/IJACSA.2017.081211

[14] Magomedov Sh. Organization of secured data transfer in computers using sign-value notation. ITM Web of Conferences. 2017. T. 10. DOI: 10.1051/itmconf/20171004004

[15] N. Chaimov, A. Malony, S. Canon et al., Scaling Spark on HPC Systems, 2016, pp. 97-110.

[16] P. Erdos and A. Renyi, On random graphs, Publicationes Mathematicae Debrecen, 1959, vol. 6, pp. 290-297.

[17] R. Pierre, Detecting Financial Fraud Using Machine Learning: Winning the War Against Imbalanced Data, 2018, URL: https://towardsdatascience.com/detecting-financial-fraud-using-machine-learning-three-ways-of-winning-the-war-against-imbalanced-a03f8815cce9 (retrieved 20.11.2018).

A Hybrid Intelligent Model for Enhancing Healthcare Services on Cloud Environment

Ahmed Abdelaziz[1]
Department of Information Systems
Higher Technological Institute,
Cairo, Egypt

Ahmed S. Salama[2]
Department of Computer and Information Systems, Sadat Academy for Management Sciences,
Cairo, Egypt

A.M. Riad[3]
Department of Information Systems,
Faculty of Computers and Information,
Mansoura University, Egypt

***Abstract*—Cloud computing plays a major role in addressing the challenges of healthcare services such as diagnosis of diseases, telemedicine, maximize utilization of medical resources, etc. Early detection of chronic kidney disease on cloud environment is a big challenge that healthcare providers are facing. This paper concentrates on the using of intelligent techniques such as Decision Tree, Clustering, Linear Regression, Modular Neural Network, and Back Propagation Neural Network to address this challenge. In this paper, the researchers propose a hybrid intelligent model based on cloud computing for early revealing of chronic kidney disease. Two intelligent techniques were used: linear regression and neural network. Linear regression was used to define crucial factors that have an impact on chronic kidney disease. The proposed model for early revealing of chronic kidney disease was built using Neural Network. The accuracy of proposed model is 97.8%. This model outperforms on the other models existed in the previous works in terms of the accuracy and precision, recall and F1 score.**

***Keywords*—Chronic kidney disease; linear regression; neural network; cloud computing**

I. Introduction

Healthcare plays an important role in saving people's lives. Healthcare provides modern trends such as telemedicine, diagnosis of diseases and etc. Cloud computing plays an important role in supporting health care services. It is composed of three main services as follows:

- Medical Infrastructure.
- Medical Platform.
- Medical Software.

Cloud computing allows many patients to retrieve their medical information anywhere and anytime by using different devices such as smart phone, personal computer and etc. There are many advantages of cloud computing for healthcare services that include the following:

- Usability
- Speed
- Accessibility
- Disaster Recovery
- Cost Savings
- Reliability
- Manageability

Stakeholders (patients, doctors, nurses and etc.) can access the medical cloud to get the right medical services anytime and anywhere. There are many health care challenges that are facing cloud computing such as:

- Telemedicine Applications
- Diseases Diagnosis Applications
- Management of Electronic Health record
- Execution Time of Medical Tasks
- Maximize utilization of Medical Resources

Many diseases diagnosis applications are used on cloud computing environment to enable stakeholders for taking an appropriate medical decision. Some examples of these applications are shown below:

- Decision tree is used to reveal of cervical cancer [1].
- Support vector machine is used to reveal of hepatitis disease [2].
- Naïve Bayes is used to reveal of Heart disease [3].
- Fuzzy logic is used to reveal of breast cancer [4].
- Fuzzy logic is used to diagnosis of lung cancer [5].
- Support vector machine is used to diagnosis of diabetes disease [6].

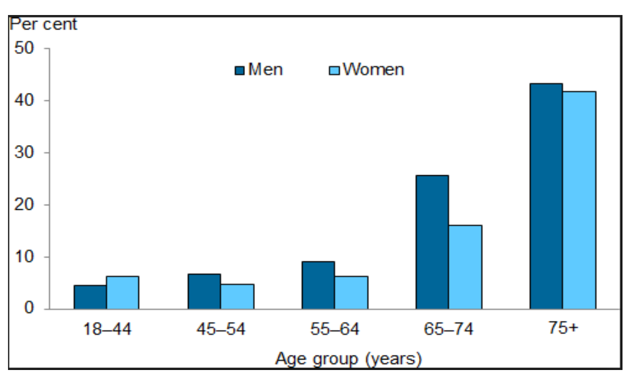


Fig. 1. Statistics of chronic kidney disease.

Early detection of diseases is one of the most important challenges for health care services on cloud computing [7]. This paper focuses on Chronic Kidney Disease because it is one of the most serious diseases facing humanity in modern times. Many statistics indicate the growing number of Chronic Kidney patients worldwide as shown in Figure 1.

This paper introduces a hybrid intelligent model for early detection of Chronic Kidney disease on cloud computing environment by using the machine learning tool on windows azure. The rest of the paper is organized as follows: section two introduces a background overview; section three presents the related work; section four introduces the proposed intelligent model; section five introduces the experimental results; and finally, section six presents the conclusion and future work.

II. Background and Overview

This section presents a review of Chronic Kidney disease, neural network and regression analysis as follows:

A. Overview of Chronic Kidney Disease

Chronic kidney disease is considered one of the most dangerous diseases that are facing the worldwide.

TABLE. I. Initial Factors that Influence on Chronic Kidney Disease

No	***Factor***	***Description***
1	Anemia	Nominal
2	Age	Numerical
3	Pedal edema	Nominal
4	Blood pressure	Numerical
5	Appetite	Nominal
6	Specific gravity	Nominal
7	Coronary artery disease	Nominal
8	Albumin	Nominal
9	Diabetes mellitus	Nominal
10	Sugar	Nominal
11	Hypertension	Nominal
12	Red blood cells	Nominal
13	Red blood cell count	Numerical
14	Pus cell	Nominal
15	White blood cell count	Numerical
16	Pus cell clumps	Nominal
17	Packed cell volume	Numerical
18	Bacteria	Nominal
19	Hemoglobin	Numerical
20	Blood glucose random	Numerical
21	Potassium	Numerical
22	Blood urea	Numerical
23	Sodium	Numerical
24	Serum Creatinine	Numerical
25	Class	Nominal

There are many factors that are influencing chronic kidney disease as shown in Table 1. A data set of chronic kidney disease from machine learning repository was used. The data set contains 800 instances.

B. Overview of Linear Regression

Linear regression has two types of regressions which are simple linear regression and multiple linear regressions. General equation of linear regression is formulated as follows:

$$M= \beta_0+ \beta_1 y_1+ \beta_2 y_2+\ldots\ldots.. B_N y_N+\epsilon \qquad (1)$$

Where:

- M is the dependent variable
- y_1, y_2, y_N are the independent variables
- β_i represents the regression coefficient
- ε represents the random error component.
- β_0 represents the y intercept

In linear regression, there are two main criteria's to measure the performance of the proposed model which are coefficient of determination and feature weight of independent variables as shown in Figure 2.

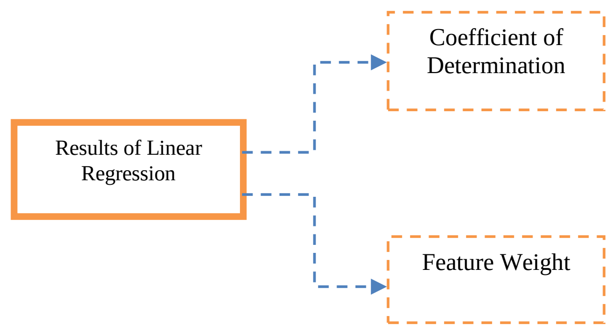


Fig. 2. Two main outputs in linear regression.

- Coefficient of determination indicates the level of performance of the linear regression model.
- Feature weight shows which variables are important. Whenever, feature weight is less than 0.05, and then the independent variable is considered significant variable.

C. Overview of Neural Network

Neural network is used to implement many intelligent applications such as diseases diagnosis and data classification and etc. There are some types of neural networks such as:

- Feed-forward network
- Back-propagation network
- Modular neural network
- Recurrent neural network

The neural network consists of input layers, hidden layers and output layers as shown in Figure 3.

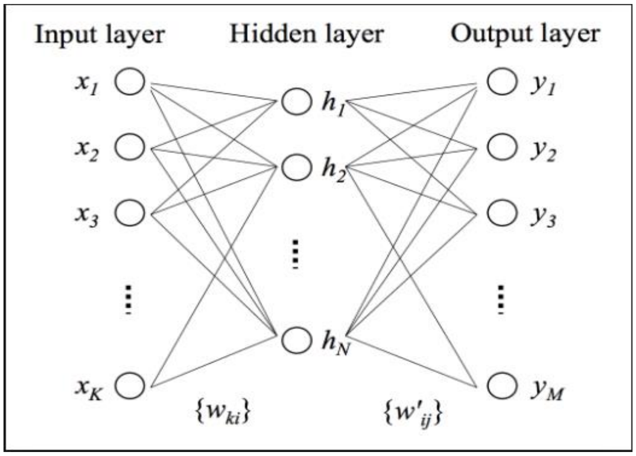


Fig. 3. General structure of neural network.

III. Related Work

This section introduces many researches that seek to detect chronic kidney disease as follows:

Lambodar Jena and et al, presented a novel model to detect chronic kidney disease by using data mining classification. This research is used WEKA tool. The accuracy of the proposed model is 95%. Naive Bayes outperforms on decision tree. The importance of the research is to predict chronic kidney disease by naive Bayes technique [8].

K. R. Anantha and et al, introduced a model which is decision tree technique for predicting chronic kidney disease on WEKA tool. This research shows that decision tree outperforms on clustering technique. The accuracy of the proposed model is 91%. The importance of the research is to predict chronic kidney disease by decision tree [9].

Basma B. and et al, presented a model for detecting chronic kidney disease by using support vector machine. The results show that support vector machine better than decision tree in terms of execution time and accuracy. The accuracy of the proposed model is 95%. The importance of the research is to detect chronic kidney disease by support vector machine [10].

Asif S. and et al, introduced intelligent model for revealing chronic kidney disease by using clustering technique on WEKA tool. This research shows clustering technique better than decision tree in order to accuracy. The accuracy of the clustering technique is 96%. The importance of the research is to reveal chronic kidney disease by clustering technique [11].

RUEY K. C. and et al, presented an approach for detecting of chronic kidney disease on cloud computing environment. The results show that back propagation neural network outperforms on modular neural network and feed forward neural network in order to accuracy and precision. The accuracy of the proposed system is 94.7%. The importance of the study is to detect chronic kidney disease on cloud environment [12].

Chien. z. W. and et al, introduced fuzzy expert system for revealing of chronic kidney disease on cloud computing environment. This research shows that fuzzy expert system better than neural network in order to accuracy and mean square error. The accuracy of the proposed model is 88.4%. The importance of the study is to reveal chronic kidney disease by fuzzy expert system on cloud environment [13].

Stuti N. and et al, introduced a new model based on Naive Bayes to predict chronic kidney disease on cloud environment (Google Application Engine). The results show that Naive Bayes outperforms on neural network in order to accuracy and precision. The accuracy of the proposed model is 97.1%. The importance of the research is to predict chronic kidney disease on Google Application Engine [14].

Anu B. and et al, introduced survey of many researches that seek to predict chronic kidney disease by using intelligent techniques. The importance of the research is to review many researches of chronic kidney disease [15].

Table 2 introduces summary of intelligent techniques of chronic kidney disease researches.

TABLE. II. Summary of Related Work

No	*Category*	*Some Approach*
1	intelligent techniques applied of chronic kidney disease	➢ quantitative approach ➢ Naive Bayes ➢ mean analysis ➢ mean time and modeling regression ➢ time series ➢ Clustering ➢ Feed Forward neural network ➢ Back Propagation neural network\\ ➢ Modular neural network ➢ Decision tree ➢ Support vector machine ➢ Fuzzy expert system

Through related work, intelligent applications are not accurate on cloud computing environment for revealing of chronic kidney disease. So, this paper introduces a new intelligent model for revealing chronic kidney disease on cloud environment.

IV. The Proposed Approach

This section introduces an intelligent model to reveal Chronic Kidney disease on cloud environment. As shown in Figure 4, this intelligent model has three steps:

1) Survey recent studies related to Chronic Kidney disease to develop a candidate list of factors that influence this disease.

2) Identify the critical factors form this candidate list using linear regression.

3) Develop an intelligent model for revealing chronic kidney disease based neural network.

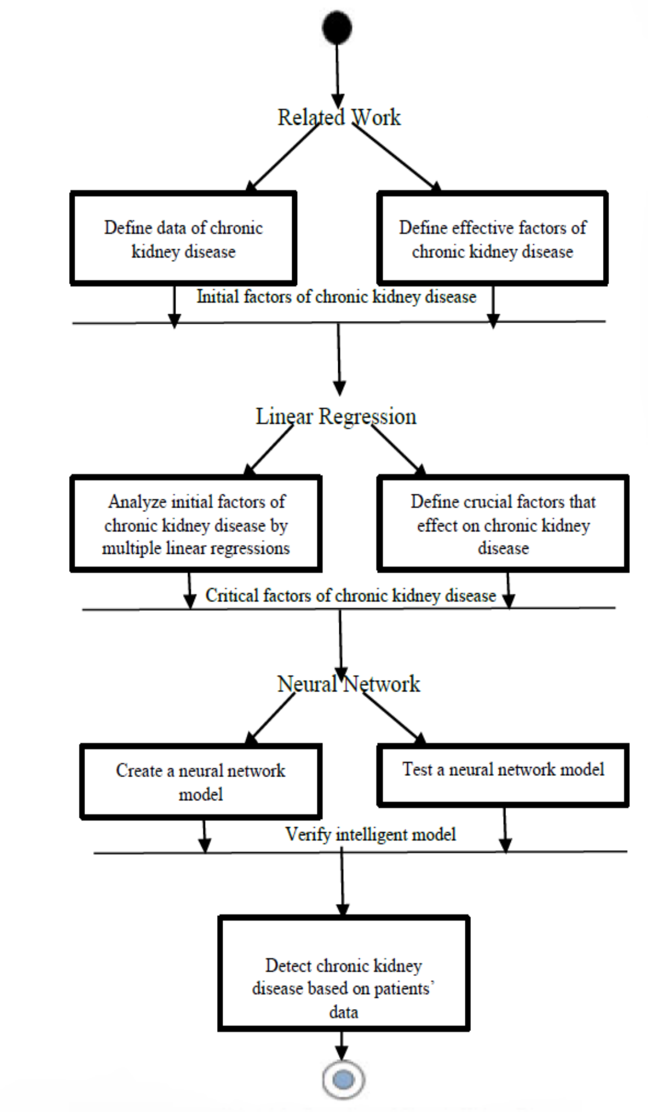


Fig. 4. The proposed model for revealing of chronic kidney disease.

V. The Hybrid Intelligent Model for Healthcare Services

This section presents the proposed hybrid model for healthcare services on cloud computing. It includes two intelligent techniques which are explained below.

A. Linear Regression

Linear regression is used to identify the critical factors of Chronic Kidney disease (CFCKD). It is composed of one dependent variable and 13 independent variables. It is formulated as follows:

$$M = \beta_0 + \beta_1 CFCKD_1 + \beta_2 CFCKD_2 + \ldots\ldots B_N CFCKD_{13} + \epsilon \quad (2)$$

Where:

M is the dependent variable (degree of influence of candidate factors on Chronic Kidney disease) and $CFCKD_1$, $CFCKD_2 \ldots$ $CFCKD_{13}$ are the independent variables (candidate factors).

This section also introduces linear regression algorithm to identify critical factors of Chronic Kidney disease.

Algorithm 1. The Algorithm of the Linear Regression to Identify Crucial Factors that influence on Chronic Kidney Disease.

1. **Input : ϱ ← (dependent variable** degree of influence proposed factors on chronic kidney disease)
2. **Output : £ ← (independent variables factors of chronic kidney disease)**
3. **ƛ = final list of crucial factors of chronic kidney disease**
4. **MLR = Multiple Linear Regression**
5. **FW = Feature Weight**
6. **SSR =** Sum of **Squares Regression**
7. **SST = Sum of Squares Total**
8. f_i **= the prediction**
9. y_i **= the true value.**
10. **SSR =** $\sum_i (f_i - \bar{y})^2$
11. **SST =** $\sum_i (y_i - \bar{y})^2$
12. **Build the MLR model based on the set of ϱ and £**
13. **Estimate the MLR model**
14. **Check coefficient of determination**
15. **coefficient of determination =** $\frac{SSR}{SST}$
16. **If (coefficient of determination < 0.5)**
17. **Change £**
18. **Go to step 13**
19. **Else**
20. **Accept the MLR model**
21. **End If**
22. **Check FW for each variable to define ƛ**
23. **If (FW < 0.05)**
24. **Approve ƛ**
25. **Else**
26. **Reject the other £**
27. **End If**
28. **Return ƛ**

B. Neural Network

Neural network has 13 critical factors as input layer, one hidden layer and two outputs (Chronic Kidney or No Chronic Kidney).

This section introduces neural network algorithm to reveal of Chronic Kidney disease.

Algorithm 2. The Proposed Algorithm of the Neural Network to Reveal of Chronic Kidney Disease.

1. **Input : £ ← (critical factors of Renal failure),**
2. **ϱ ← (hidden layers),**
3. **Output : ¥ ← (Detection of Renal failure)**
4. **NNM = Neural Network model**
5. **ƛ= Accuracy of the NNM**
6. **Create NNM with £, ϱ, and ¥**
7. **Verify NNM**
8. **Verify ƛ**
9. **If (ƛ > 0.95)**
10. **Accept NNM**
11. **Go to step 17**
12. **Else**
13. **Update the number of ϱ**
14. **Go to step 6**
15. **End If**
16. **Identify ƛ via validate NNM**
17. **Return ¥**

VI. Experimental Results

This section presents the execution of the proposed hybrid model on windows azure as shown in Figure 5.

TABLE. III. Summary of Critical Factors of Chronic Kidney Disease

No	Factor Name	FW
1	Anemia	-0.002
2	White Blood Cell Count	0.0007
3	Packed Cell Volume	-0.002
4	Hemoglobin	0.03
5	Potassium	-0.0006
6	Sodium	0.002
7	Serum Creatinine	-0.05
8	Blood Urea	0.003
9	Blood Glucose Random	0.005
10	Sugar	0.04
11	Specific Gravity	-0.002
12	Blood Pressure	-0.00005
13	Age	0.02

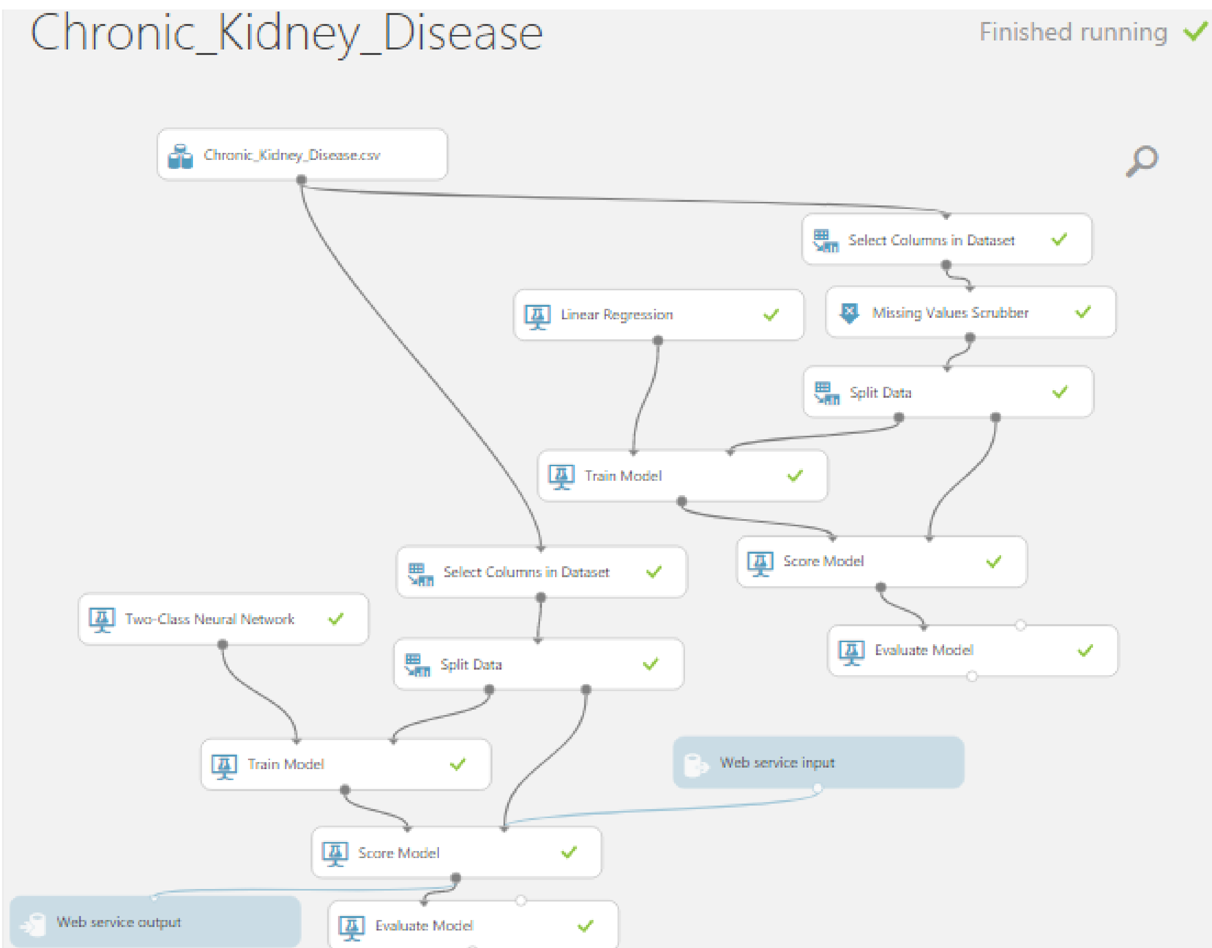


Fig. 5. Execution of hybrid model on windows azure.

A. Execution of Linear Regression on Windows Azure

This section introduces the execution of linear regression on windows azure. Linear regression was used to define critical factors of Chronic Kidney disease. Table 3 shows the critical factors of Chronic Kidney disease via the feature weights. Whenever, feature weight is less than 0.05, the factor is medically important. For example, the feature weight for Serum Creatinine (feature weight =- 0.04) is less than 0.05, then this factor is medically important as shown Chronic Kidney in figure 6.

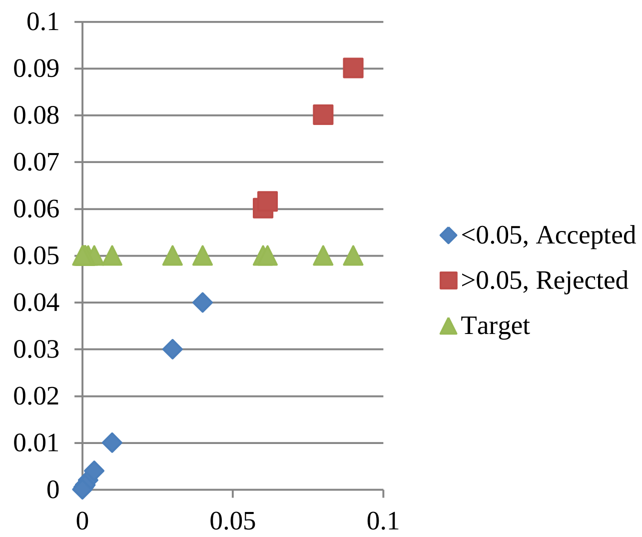


Fig. 6. Condition statement of factors of chronic kidney disease.

B. Execution of Neural Network on Windows Azure

This section introduces the execution of neural network on windows azure. NNM is composed of 13 neurons in input layer, one hidden layer and two outputs. The neural network type used in this paper is the two class neural network. Table 4 shows the performance of the neural network. It defines accuracy, precision, recall and f1 score. The accuracy of the proposed model is 97.8% as shown in Figure 7.

TABLE. IV. PERFORMANCE ATTRIBUTES OF NNM

True Positive	*False Negative*	*Accuracy*	*Precision*
51	0	0.978	0.962
False Positive	***True Negative***	***Recall***	***F1 Score***
2	39	1.000	0.981

C. Case Study

Three real cases of patients were conducted on windows azure to reveal of Chronic Kidney disease (CKD) as shown in Table 5 and Figure 8.

TABLE. V. DATA OF CKD PATIENTS

No	*Factor Name*	*Patient 1*	*Patient 2*	*Patient 3*
1	Age	49	59	29
2	Blood Pressure	71	81	101
3	Specific Gravity	2.007	1.026	1.2
4	Sugar	0	0	1
5	Blood Glucose Random	127	132	121
6	Blood Urea	57	19	21
7	Serum Creatinine	3.9	1.2	1.3
8	Sodium	113	142	131
9	Potassium	2.6	3.6	2.2
10	Hemoglobin	12.2	15.9	11
11	Packed Cell Volume	33	54	51
12	White Blood Cell Count	5700	6900	5100
13	Anemia	Yes	No	Yes
14	Class	CKD	NOCKD	NOCKD
15	Probability	0.00005	0.99993	0.99997

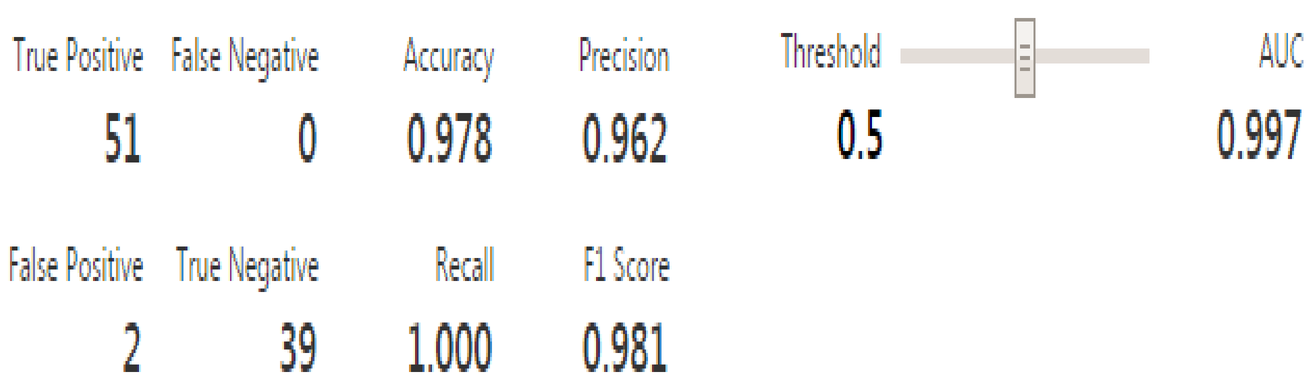


Fig. 7. Performance attributes of hybrid model on windows azure.

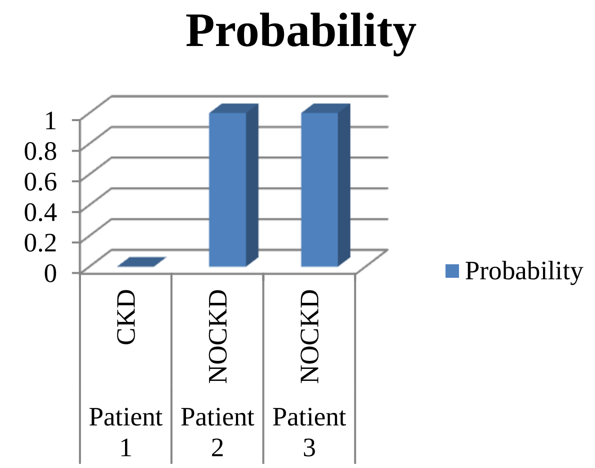


Fig. 8. Probabilities of CKD patients.

Figure 9 shows the accuracy of the neural network model on windows azure. This model outperforms on the other models existed in the previous works in terms of the accuracy and precision, recall and F1 score.

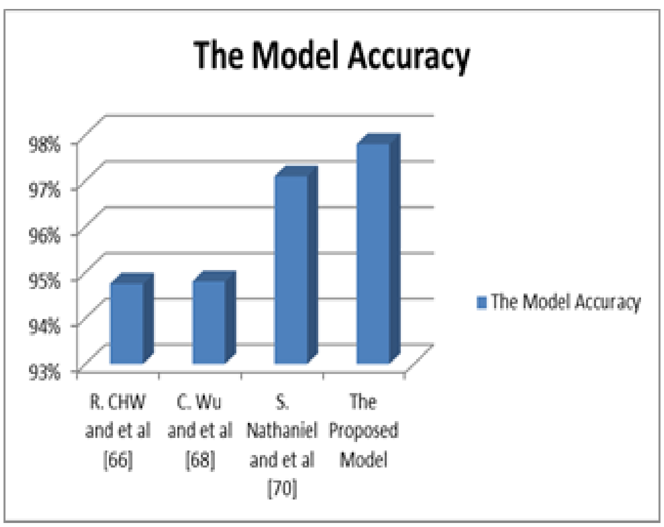


Fig. 9. Accuracy of the NNM compared to the rest of the models.

VII. Conclusion and Future Work

Cloud computing raises healthcare performance. Early detection of diseases is one of the most challenges that are facing healthcare providers. Chronic kidney disease is considered one of the dangerous diseases that are facing humanity worldwide. The accuracy of many traditional applications are not good enough for revealing chronic kidney disease. This paper tries to introduce intelligent model based on cloud computing for address this challenge. The proposed hyprid model includes two techniques: linear regression and neural network. The proposed model was implemented on windows azure. The accuracy of the proposed model is 97.8%. Three real cases of patients were conducted on windows azure to emperically validate the proposed model.

It is recommeded as future work:

- To use logistic regression for determining critical factors that affect Chronic Kidney disease.
- To uses deep learning for revealing Chronic Kidney disease on cloud computing environment.

References

[1] S. Sha "Cervical Cancer stage prediction using Decision Tree approach of Machine Learning", IJARCCE, Vol.5, No.4, 2016, pp. 345-348.

[2] C. Kum, M. V. Kumar, T. Gayathri and S. R. Kumar "Data Analysis and Prediction of Hepatitis Using Support Vector Machine (SVM)", IJCSIT, Vol.5, No.2, 2014, pp. 2235-2237.

[3] Prera. T , Shivaprakash .N and Swetha .N "Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS", IJSE, Vol.3, No.2, 2015, pp. 90-99.

[4] Tin. P and Paulin. R "Detect Breast Cancer using Fuzzy C means Techniques in Wisconsin Prognostic Breast Cancer (WPBC) Data Sets", IJCAT, Vol.2, No.5, 2013, pp. 614-617.

[5] M. HA "Lung Cancer Diagnosis by Using Fuzzy Logic", ijcsmc, Vol.5, No.3, 2016, pp. 32-41.

[6] Arjn and An. S "Diagnosis of Diabetes Using Support Vector Machine and Ensemble Learning Approach", IJEAS, Vol.2, No.11, 2015, pp. 68-72.

[7] M. Elhosen, A. Salama, Ahmed. A, A. M. Riad, "Intelligent Systems Based on Cloud Computing for Healthcare Services: A Survey", International Journal of Computational Intelligence Studies, Vol. 6, Nos. 2/3, 2017, pp. 157-188.

[8] L. Jen and N. Kam "Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease", IJERMT, Vol.4, No.11, 2015, pp. 110-118.

[9] K. Padmanab and G. Parthib "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease", ijst, Vol.9, No.29, 2016, pp. 1-5.

[10] B. Bouken, H. Mousan and A. Ha "Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease", IJDMS, Vol.8, No.3, 2016, pp. 1-9.

[11] A. Salek and J. Stanko "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes", ICHI, Vol.8, 2016, pp. 1-9.

[12] RUEY .C, R. Y. CHEN and S. JEN "INTELLIGENT SYSTEMS ON THE CLOUD FOR THE EARLY DETECTION OF CHRONIC KIDNEY DISEASE", International Conference on Machine Learning and Cybernetics, Vol.8, 2012, pp. 1737-1742.

[13] C. Wu, R. Chiu and S. Wang "A cloud-based fuzzy expert system for the risk assessment of chronic kidney diseases", Int. J. Business and Systems Research, Vol. 9, No. 4, 2015, pp. 315-333.

[14] S. Nathaniel, A. Motwani and A. Saxena "Cloud based Predictive Model for Detection of 'Chronic Kidney Disease' Risk", International Journal of Computer Sciences and Engineering, Vol. 6, No. 4, 2018, pp. 185-188.

[15] A. Bat and V. Sin "A Review to Predictive Methodology to Diagnosis Chronic Kidney Disease", International Conference on Computing for Sustainable Global Development, IEEE, Vol.4, 2016, pp. 2760-2763.

Implementation of a basic Sonar of Echolocation for Education in Telecommunications

Fredy Criollo-Sánchez[1], Rodríguez-Villarreal, Kevin[2], Mosquera-Sanchez, Cristian[3], Medina-Alvarez, Marco[4], Chavarry-Polanco, Danny[5]
Electronics Engineering Department Universidad de Ciencias y Humanidades
Lima, Peru

Alvarado-Díaz, Witman[6], Meneses-Claudio, Brian[7], Roman-Gonzalez, Avid[8]
Image Processing Research Laboratory (INTI-Lab), Universidad de Ciencias y Humanidades
Lima, Peru

***Abstract*—Currently, having a sonar of echolocation in an electronic lab is complicated due to the high cost of its implementation, which is why it is proposed that the implementation of a basic sonar, using agile technologies such as the Arduino, will be used to implement this paper. Also it has a servomotor and an ultrasonic sensor, which is responsible for detecting the distance where the objects are located. The Arduino will be in charge of controlling the servomotor movements which have to be between 15 and 165 degrees. In addition, it will send the information through the serial port to a computer, in which the data will be processed and displayed using the Processing software.**

***Keywords*—Arduino; processing software; sonar; ultrasonic sensor; servomotor**

I. Introduction

The sonar of echolocation isn't new, since it was used in the second world war with the purpose of locating the location or positioning of the enemy, it was also used in the navy for the measurement of distance for enemy targets; sonar is a complex system that uses several elements such as signal processing, data processing, parameter detection and estimation, among others, so having a sonar in the telecommunications laboratories is important since it can have great benefits in the education such as: studying the process of finding objects, tracking mobile objects in addition to image processing; however, having a sonar in the laboratories isn't possible due to its high cost, so the purpose of this research is to implement a basic sonar based on current technologies, to be used in the laboratories of the institutions that require it.

The paper consists of the implementation of a basic sonar based on agile technologies such as Arduino, for the laboratories of the universities and institutions that require it, for this purpose an Arduino, an ultrasonic sensor, a servomotor and a buzzer will be used; the system has the ability to detect objects through ultrasonic pulses, in addition to displaying them through a computer. Also, implement this technology, making a laboratory with low-cost resources for education at universities.

II. Methodology

The paper is based on the basic sonar principles, which is to emit an electromagnetic or acoustic wave and detect the reflected return from a distant surface; the signal of a sonar is generated in the transmitter, it radiated in the direction of an object where it is reflected towards the receiver, the same one that processes and obtains the information that can be the distance, speed, direction and characteristics of the object [1], [2], [3].

For this paper, the piezoelectric effect that is sensitive to the direction will be used, because the voltage produces a defined polarity in the voltage [4], while the compression produces an opposite one. So, if a voltage is applied to the piezoelectric transducer, the crystal changes in its dimensions, which causes a change of pressure in the surrounding medium (air), and vice versa, when the crystal is subjected to a change of pressure appear electrical charges at their ends, where a potential difference is created. So, this type of transducer can function as an ultrasonic emitter or receiver.

The main components of the system can be seen in Figure 1, where we see a general scheme of the connection of the servomotor and ultrasonic sensor to the Arduino, in addition to the serial communication with the computer.

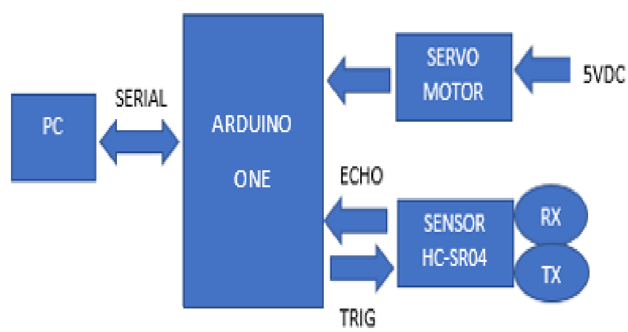


Fig. 1. Block diagram.

A. Description of Principal Components

1) Arduino one: The Arduino one (Figure 2) is the computer of the circuit which controls the operation of the servomotor and the transmission and reception of the pulses of shots towards the sensor [5], also intercommunicates with the PC thought the USB cable recognizing it, as a serial port, and thus of this way we can visualize graphs, executing it in the software Processing.

TABLE I. CHARACTERISTICS OF THE ARDUINO ONE

Parameter	Value
Input Voltage	+7 a + 12 V
GPIO Pins	14
Analog input pins	6
DC current per I / O pin	40 mA
Flash memory	32 KB
SRAM	2 KB
Communication Support	Serie IC
Clock frequency	16 Mhz

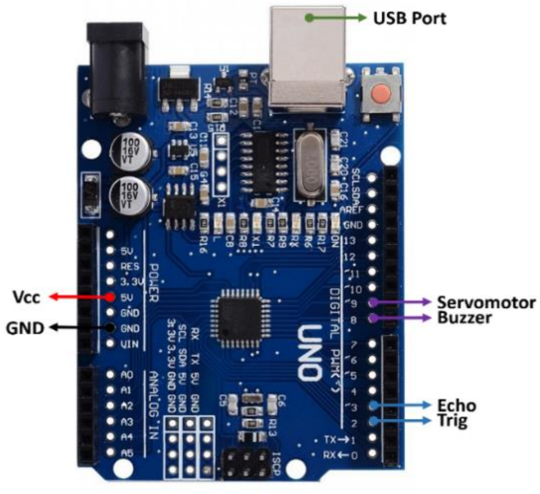


Fig. 2. Arduino one controller.

2) *Servomotor*: A Micro Servo 9g SG90 from Tower Pro as we can see in Figure 3 was used for this paper. These motors work with a PWM signal, with a working pulse between 1 ms and 2 ms and a period of 20 ms (50 Hz) [6]. This data tells us the maximum speed at which we can move the servomotor with Arduino. We can only change position every 20 ms. This will depend on the type and brand of our servomotor.

It is necessary to take into account, the angle of rotation of the servomotor, in this case, it allows us to make a sweep between -90º and 90º [3]. It means that is a 180º angle of rotation. Although the servomotor can move with a resolution of more than 1 degree, this is the maximum resolution that we will achieve due to the limitation of the PWM signal that is capable of generating Arduino One.

Fig. 3. Servomotor.

3) *Ultrasonic sensor*: We chose this device because it is a very cheap sensor, this sensor emits a pulse of sound at a high frequency not audible by the human ear, the sonar waves bounce back to the sensor. Since we know the speed of sound in air (343.2 m / s), we can calculate the distance with e = v * t. (space = speed per time). The measurement range of the sensors usually ranges from a few centimeters to several meters. The most common, like the HC-SR04, have 4 pins (see Figure 4). Two of them are for the connection (Vcc and GND) and the other two pins are the Trigger and the Echo [7].

TABLE II. CHARACTERISTICS OF THE HC-SR04 ULTRASONIC SENSOR

Parameter	Value
Working voltage	5 VDC
Working Current	15 mA
Working Frequency	40 Khz
Maximum range	4 m
Minimum range	2 cm
Measuring angle	15°
Trigger signal pulse	Pulse TTL 10 us
Dimension	45*20*15 mm

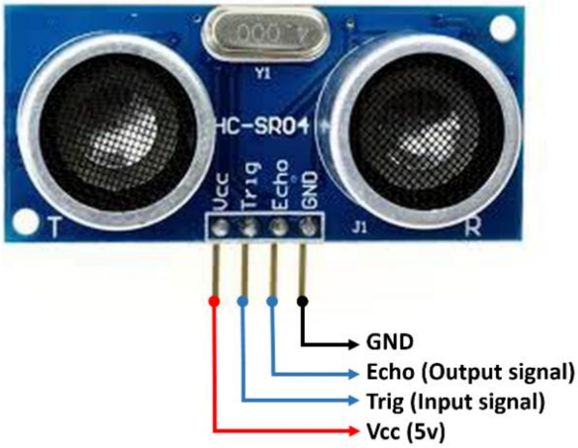


Fig. 4. Ultrasound sensor.

When a high level (HIGH, 5V) of at least 10 µs is provided to the "Trig" pin, the ultrasonic module emits a series of 8 pulses at a high frequency (40 KHz) and then sets its "Echo" output to HIGH. If the signal encounters an obstacle, it will be reflected and captured by the sensor module. At the moment of catching the rebound signal, the module will change its "Echo" output from HIGH to LOW. Therefore, in the "Echo" output, a high-level pulse (HIGH) has been maintained with a duration equal to the time used by the signal in going from the sensor to the obstacle and returning. The pulse width varies approximately between 150 µs and 25 ms. If there is no obstacle, the width of the echo pulse will be around 38 ms.

B. System Implementation

The system is implemented in a basic way, in order to demonstrate its operation, in addition to the aforementioned components, the necessary code is developed both for the Arduino and the coding in Processing, for the graphic visualization of the data. Flow diagrams that better explain the respective coding were constructed and one can see it in Figure 5.

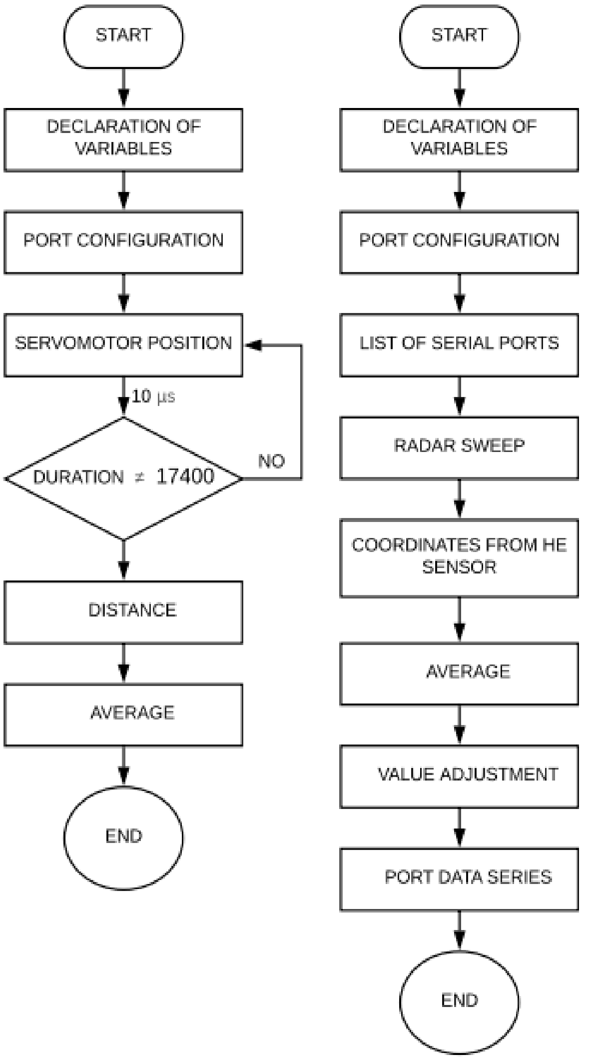


Fig. 5. Flow diagram.

For the Arduino, after declaring the variables, the servomotor must be positioned and then send the instructions for the ultrasonic sensor to perform the distance measurement, for which the following code was implemented, with which the servo could be positioned between 15 and 165 degrees [8], and in addition to calculating the distance, the data is sent through the serial port. The program is showing:

```
for (int t = 15; t <= 165; t++)
 {
  Servo.write(t);
  delay(60);
  dist = calcular_distancia();
  Serial.print(t);
  Serial.print(", ");
  Serial.print(dist);
  Serial.print(".");
```

A flow diagram of the system is going to be show next because we need to follow steps in order to make a paper reliable for future works.

In order to get the servomotor to return to a previous position, the parameters of the for loop must be changed (int t = 165; t> 15; t--), thereby achieving the aforementioned; the function calculate_distance () sends a high pulse of 10 us for the trigger pin, and receives the response pulse for the echo pin, measuring the duration of it, after knowing the time it takes the answer to apply the following formula to know the distance recorded at a given moment.

$$distance = time * \left(\frac{0.0343}{2}\right)$$

"time· is the time it takes the pulse and "distance" is the distance between the sensor and the test object, the number 0.0343 represents the speed of sound expressed in cm / us in the formula.

III. Results

The implemented circuit (Figure 6) was verified by us, the connection of the components before supplying power to the circuit, through the USB cable connected to the serial port of the PC. The ultrasonic sensor was mounted on the servomotor, a small cardboard was used and it was fixed with glue.

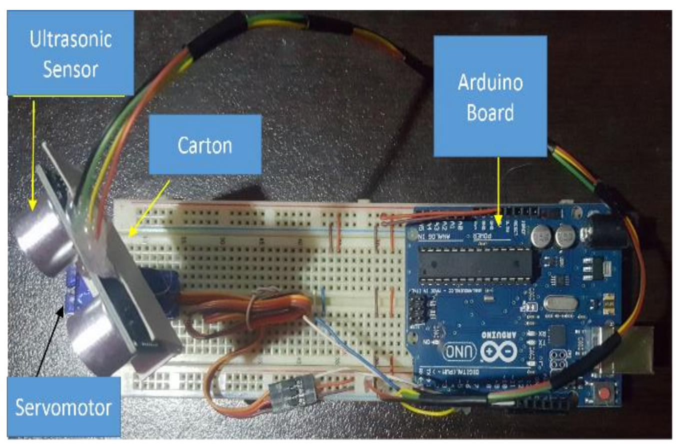


Fig. 6. System components.

By performing the tests, the graphic should look similar to the image in Figure 7.

As we can see in Figure 7, the graphic is shown in a bad location, so it was necessary to modify the code to center the image on the screen. With respect to the detection of the object is fulfilling the objective, the distance in which the object is located is shown on the screen.

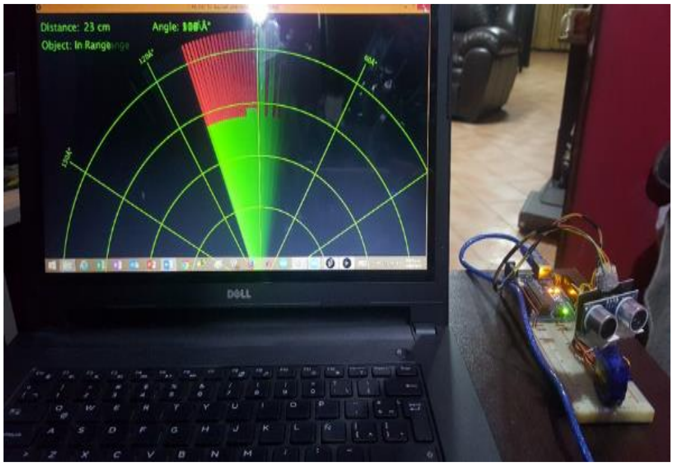


Fig. 7. Results of the Sonar.

IV. Discussion

The design of this device is very simple, the main purpose of detecting objects at different distances was achieved by visualizing it in a graph in a PC where the reader can see the distance and the location angle of the object, the Arduino card one that has different applications to perform all types of papers and the HC-SR04 ultrasonic sensor in charge of measuring the distance at the moment that receives the echo of a transmitted signal.

The use of Arduino one has an endless number of applications, several papers can be carried out, the Arduino and Processing codes are easy to program since the C ++ language is well known within the programs, besides the codes are free on the internet with some modifications to the program you can get. Based on the sonar principles and the development of the classes learned during my professional career, it was possible to finish this wonderful paper.

The recommendation is that amplifiers can be used to obtain greater distances and even other graphic visualization programs to improve and develop future work.

V. Conclusions

It was possible to build a system that can help to learn the basic concepts of the technology involved in sonars, with our system, we can teach students the concepts of object detection, signal processing, among other concepts that involve the sonars; This paper shows that it is possible to design low cost prototypes, which are possible to implement in the telecommunications laboratories of the interested institutions.

As a conclusion, the steps and also the process that we follow to make this paper was long. It’s really important for students to always follow the steps because they will improve their capabilities of understand the process in different ways.

References

[1] F. I. Vera Rosales and A. M. Mora Guijarro, “Diseño De Un Prototipo De Pruebas De Un Sonar Monoestático Utilizando Radio Definida Por Software”, UNIVERSIDAD POLITÉCNICA SALESIANA SEDE GUAYAQUIL, 2017.

[2] R. J. Álvarez González and S. I. Shagñay Vimos, “Sistema Sonar Para Análisis Y Reconstrucción De Perfiles Topográficos A Media Escala” ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO, 2015.

[3] M. Arrambí Díaz, “Modelado De Un Sonar Doppler De Pulsos (PDR),” INSTITUTO POLITÉCNICO NACIONAL, 2005.

[4] J. Ferrer Ferrer, “Caracterización De Sensores Piezoeléctricos Y Capas De Adaptación (Matching Layer) Para Detectores De Partículas Tipo Pico” UNIVERSIDAD POLITECNICA DE VALENCIA, 2015.

[5] M. J. Castillo Castillo, “Mis Proyectos con Arduino,” 2017.

[6] M. Á. Rubio Escudero, “Diseño detallado de las demostraciones de teoría con Arduino,” 2012.

[7] X. Cuamatzi Flores, M. A. Jiménez Muñiz, and F. J. Navarrete González, “SISTEMA DE PROXIMIDAD ULTRASÓNICO,” INSTITUTO POLTÉCNICO NACIONAL, 2010.

[8] Wang, Y., Zhao, Y., Bortoff, S. A., & Ueda, K. (2015). A real-time energy-optimal trajectory generation method for a servomotor system. IEEE Transactions on Industrial Electronics, 62(2), 1175-1188.

Experimental Study on an Efficient Dengue Disease Management System

Planning and Optimizing Hospital Staff Allocation

J.M.M.C.Jayasuriya, G.K.K.T.Galappaththi, M.A.Dilupa Sampath, H.N.Nipunika, W.H. Rankothge
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

***Abstract*—Dengue has become a serious health hazard in Sri Lanka with the increasing cases and loss of human lives. It is necessary to develop an efficient dengue disease management system which could predict the dengue outbreaks, plan the countermeasures accordingly and allocate resources for the countermeasures. We have proposed a platform for Dengue disease management with following modules: (1) a prediction module to predict the dengue outbreak and (2) an optimization algorithm module to optimize hospital staff according to the predictions made on future dengue patient counts. This paper focuses on the optimization algorithm module. It has been developed based on two approaches: (1) Genetic Algorithm (GA) and (2) Iterated Local Search (ILS). We are presenting the performances of our optimization algorithm module with a comparison of the two approaches. Our results show that the GA approach is much more efficient and faster than the ILS approach.**

Keywords—Optimization; genetic algorithm; iterated local search; algorithm comparison; nurse scheduling

I. Introduction

The number of Dengue fever cases has grown drastically in Asian countries like Sri Lanka, and hospitals are facing challenges when taking care of the patients due to the lack of resources: materials as well as human. Therefore, allocation of available hospital resources and hospital staff to take care of Dengue patients efficiently has become an important requirement.

According to a survey conducted in the context of Sri Lanka, Infectious Disease Hospital (IDH) [6] which is now known as National Institute of Infectious Diseases, faces the same challenges and they require an efficient Dengue Disease Management System that can predict the dengue out breaks, plan the countermeasures accordingly and allocate resources for the countermeasures.

We have proposed a platform for Dengue disease management with following modules: (1) a prediction module to predict the dengue outbreak and (2) an optimization algorithm module to optimize hospital staff according to the predictions made on future dengue patient counts.

This paper mainly focuses on the optimization algorithm module. The optimization algorithm module has been developed based on two approaches: (1) Genetic Algorithm (GA) and (2) Iterated Local Search (ILS). Furthermore, a web-based application was developed named “SmartScheduler”, which generates working schedules of each nurse as part of the optimization algorithm module. “SmartScheduler” tries to generate an optimal work plan for nurses’ staff who take care of the Dengue patients.

Experiments were conducted to measure the performances of the optimization algorithm module (GA based as well as ILS based) using the statistics collected from IDH. With the real data sets, the results show that, GA and ILS can decide the optimal allocations dynamically in the order of seconds. Also, the results show that the GA approach is much efficient and faster than the ILS approach.

The rest of the paper is organized as follows. Section II presents the related work. Section III introduces the optimization algorithm to decide nurse staff allocation. In Section IV, the result and discussions are presented. Under Section V final remarks are mentioned and finally references are mentioned at the end.

II. Related Work

To achieve the best utilization of resources, proper optimization is required. In this research, optimization is used to generate working schedules for the nurses of IDH hospital. Optimizing algorithm module is implemented using two approaches: GA and ILS. The GA and ILS approaches are well known methods in the planning and scheduling context [3, 4, 5]. In this section, the limited work on scheduling and planning in the hospital management context will be discussed.

The research work on [1] focused on optimizing healthcare staff in the Pediatric Department of Prince Sultan Military Medical City (PSMMC) using GA with cost bit matrix. The main goal of their optimization was to satisfy doctors scheduling problem as much as possible while fulfilling the employers’ requirements. Their results showed that the suggested method is very useful and able to give reasonable solutions to the problem fast compared to the traditional manual methods.

The authors in [2] worked on Hospital-Residents matching problem using a local search approach and a stable matching approach. The stable matching approach is a method where one finds a stable matching between two equally sized sets of elements. Agents are assigned to another sets of agents which consists of preference lists and capacities under certain constraints. The authors have evaluated both approaches on big artificial instances that are comparable with practical ones, which involve thousands of agents. Their experimental results show that the algorithm can return a solution in few seconds and with a high quality in terms of the size of the returned matching.

Our research work has been inspired by these works, and we are focusing on an optimization algorithm module to optimize IDH hospital staff according to the predictions made on future dengue patient counts.

III. Optimization Algorithm Module for Staff Scheduling

In this section the optimization algorithm that schedules nurse staff to take care of the dengue patients and generates work plan for each nurse is described.

A. Requirement Gathering

Before proposing an optimization model for scheduling the nurse staff, it was very important to understand the current context of dengue patients and their treatment process in Sri Lanka. Therefore, personal interviews and discussions with the IDH hospital staff were carried out to gather the monthly patient's data, nurse staff data, ward details etc. of IDH hospital [6]. Additionally, personal discussions and interviews with doctors and nurses outside of the IDH hospital were held to analyze the collected data for scheduling purposes [6].

B. Fitness Function

The optimization algorithm was developed using two approaches: (1) GA and (2) ILS. The goal of the algorithm is to find a solution that schedules nurse staff optimally. We have consulted IDH staff and general hospital administration bodies to identify the factors effecting staff scheduling [6]. They are as follows:

- Number of patients in a ward
- Number of wards and its priority
- Total number of patients
- Nurses requested to work overtime (OT)
- Maximum number of patients that could be allocated to a nurse

It is important to note that, according to the discussion and interviews, in the best case, the maximum number of patients that could be allocated to a nurse is 8 [6].

Therefore, considering these factors affecting staff scheduling, a fitness function was derived (a minimization function) to measure the quality of the solutions provided by the optimization algorithm as follows:

Patients per nurse as allocated by the algorithm > 8

fitness = fitness + x

else if Patients per nurse as allocated by the algorithm = 8

fitness = fitness + 0

else if Patients per nurse as allocated by the algorithm < 8

fitness = fitness + y

If the nurse already has worked (overtime)

fitness = fitness + x

C. Genetic Algorithm (GA) based Approach

Genetic algorithm is a search heuristic that is based on Charles Darwin's theory of natural evolution [9]. This algorithm reflects the procedure of natural selection where the fittest individuals are selected for reproduction to produce offspring of the next generation. Simply, it chooses individuals from the current population and uses them as parents to produce the children for the next generation. Over succeeding generations, the population evolves towards an optimal solution.

GA can be described by the following five key steps [9]:

1) *Generate an initial population F (0) with n solutions*
2) *Compute the fitness value u(f) for each individual solution f in the current population F(t)*
3) *Generate the next population F(t+1), by selecting i best solutions from F(t)*
4) *Produce offspring by applying the genetic operators to population F(t+1)*
5) *Repeat from Step 2 until a satisfying solution is achieved.*

In the research work, GA was used as one of the approaches to achieve optimization in nurse scheduling and the process is briefly explained below

The GA process begins with an initial population with n number of solutions, where in each solution, the nurses are scheduled to wards randomly, without considering any factors. Random initialization method is selected because it drives the population to optimality.

The encoding of a Solution (optimal number of nurses for wards) is demonstrated in Figure 1. S symbolizes the solution in a population. In a solution, each gene supplies two pieces of information: (i) the gene index number represents the ward index number; and (ii) the value in the gene signifies the optimum number of nurses for that ward. For example, the value of the first gene is 6, indicating ward number 1 has been allocated total of 6 nurses.

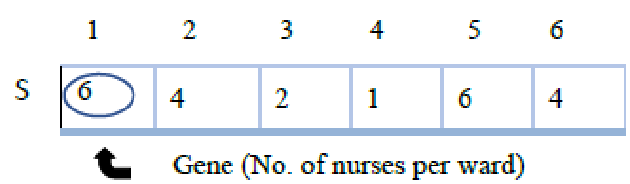


Fig. 1. Encoding of a Solution.

Two types of genetic operators to produce offspring were considered: (1) mutation and (2) crossover. The crossover is a convergence operation which is intended to pull the population towards a local min or max. The mutation is a divergence operation which is intended to occasionally break one or more members of a population out of a local min/max space and potentially discover a better space. Each generation of the GA approach goes through mutations and crossovers.

For the crossover operator, as shown in figure 2, the “One-Point crossover” approach was used. A random point of two solutions are selected as the crossover points and two new solutions are generated by exchanging the elements of parents among themselves up until the crossover point is reached.

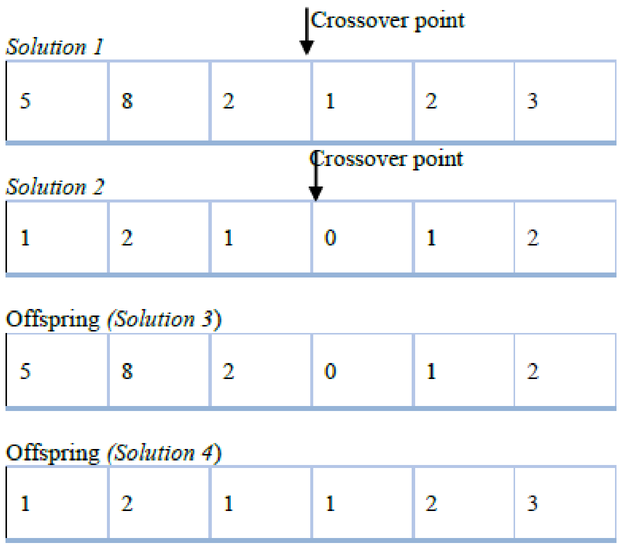


Fig. 2. One-Point Crossover.

For the mutation operator, as shown in figure 3, the “swap mutations” approach was used, where the over allocations and less allocation of nurses for a ward are swapped accordingly (added and removed within a solution).

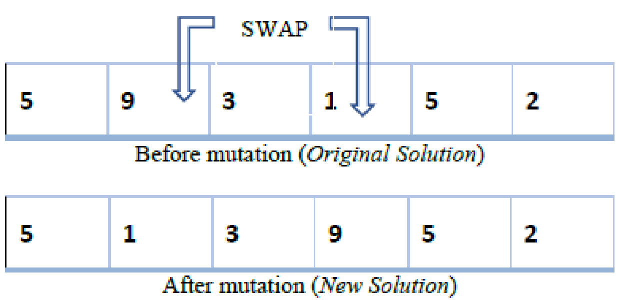


Fig. 3. Mutation.

The newly generated solutions are evaluated according to the fitness function derived in Section B. The process is continued until x number of generations are explored and the best solution (the solution with minimum fitness value) is selected as the optimal number of nurses for wards.

D. Iterated Local Search (ILS) based Approach

ILS is more effective and quickly explores the optimal solution. It can find a highly accurate optimum by using a few numbers of iterations [10]. This algorithm keeps track on its current state (current solution) and moves to neighboring states (neighborhood) with the results of current state. Then it repeats the same process if the neighboring states get better than the current state, until it gets the best solution.

ILS process can be described by the following key steps [10]:

1) Generate an initial solution f

2) Produce a new solution f’ by applying local search and perturbation to the solution f

3) Compute the fitness value for the new solution f’

4) If the fitness value of new solution f’ is better than fitness value of original solution f,

Accept the new solution f’

else accept the original solution f

5) Repeat from Step 2 until a satisfying solution is achieved.

The ILS starts with an initial solution. Two approaches to generate the initial solution are as follows: (1) best fit and (2) random fit. In the best fit approach, the nurses are allocated to wards according to the requirement of each ward (number of nurses required by each ward). Wards are considered sequentially (by the ward number), and nurses are allocated starting from first wards, second ward etc. Therefore, the last wards might not get nurses allocated at all. On the other hand, with the random fit approach, the nurses are allocated to wards randomly, without considering any factors.

Next, the initial solution goes through a perturbation process to generate a new solution. In the perturbation process, each nurse is selected sequentially and moved from the allocated ward to another ward. Each of this move generates a new solution. Each new solution will be evaluated using the heurist function. The previously mentioned fitness function under section B was used as the heuristic function for ILS. If the new solution gives a better fitness value, then the old solution is discarded, and new solution is considered as the current solution.

The process is continued until x number of iterations are explored and the best Solution (the solution with minimum fitness value) is selected as the number of tourists that can be accommodated for each location.

IV. Results and Discussion

A. Experimental Setup

The optimization algorithms were implemented using Spyder application under Anaconda Navigator which is a free and open source distribution of the python programming language. Experiments were performed on a laptop with the following specifications;

- CPU: Intel Core i7 (2.40GHz)
- RAM: 4GB
- OS: Windows 10 (64-bit Operating System)

The statistics collected from the IDH were used (number of dengue patients, number of specialized wards, number of nurses etc.) for the experiments [6].

B. GA based Approach for Nurse Allocation

In this section, the results of the GA approach are presented, which was one of the approaches used to allocate nurses optimally.

The solutions are generated assuming there are 25 nurses and 3 wards. There are 30 patients in Ward 1, 30 patients in Ward 2 and only 10 patients in Ward 3.

Experiment 1: As explained earlier, the GA process tries to improve the given initial solution by applying genetic operations over the generations. To explore how GA process improves the initial solution, 30 rounds of experiments were conducted [7]. First, find the initial solutions, and then improve the solution using GA process. As shown in the figure 4, the important observation was that most of the improvements in the fitness function happens early on (during first 250 generations) and after that improvements decrease significantly. In fact, there were very few improvements after 300 generations.

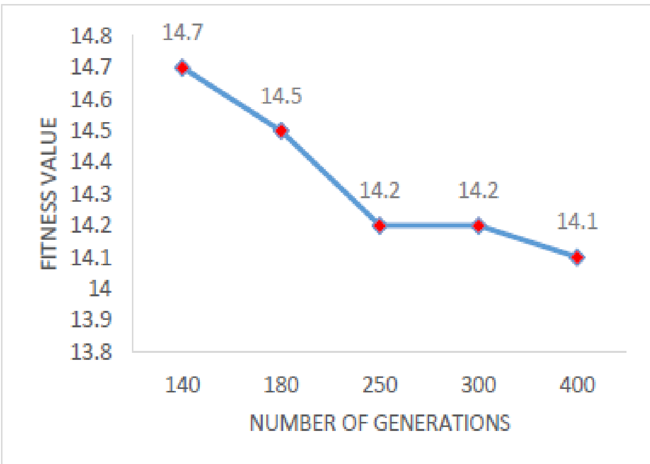


Fig. 4. Effect of Number of Generations on the Fitness Score.

Experiment 2: GA process begins with a set of solutions known as the initial population. To evaluate how GA process is affected by the size of the population (number of initial solutions), experiments were carried out for different population sizes (30 rounds of experiments for each population size) [7]. As shown in figure 5, the results showed that a better solution can be achieved when the population size is greater than 10.

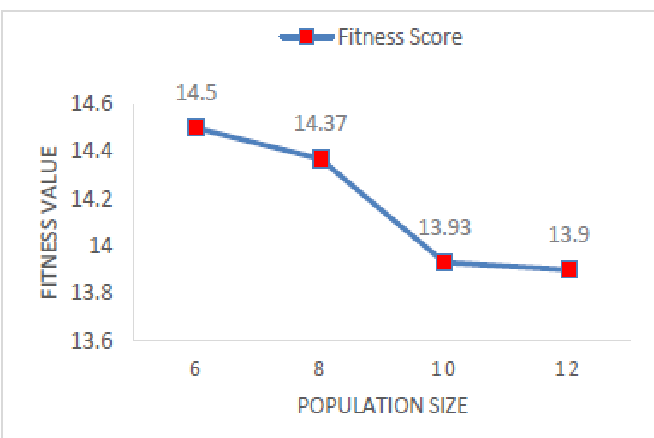


Fig. 5. Effect of Population Size on the Fitness Score.

C. ILS based Approach for Nurse Allocation

The same data set and same fitness function were used (as the heuristic function) for ILS approach, to compare ILS approach with GA approach.

Experiment 3: As explained in Section III, the ILS process generates the initial solution using two approaches: (1) best fit and (2) random fit. Two sets of experiments were conducted separately for the two approaches (30 rounds of experiments for each approach) [8].

In each experiment, revealed how the ILS tries to improve the given initial solution by applying perturbation and hill climbing over the iterations. First, find the initial solution (by best fit or random fit), and then improve the solution using ILS process.

As shown in the figure 6, the first observation was, starting with an initial solution generated by best fit approach was better than starting with an initial solution generated by random fit approach. As the best fit approach already provided an initial solution which was much better, the ILS process did not have to do much work to improve the solution in its iterations. Therefore, if started with an initial solution from the best fit approach, one can find a better solution fast.

The next observation is, in general, that most of the improvements in the heuristic function happens early on (during first 20 iterations) and after that improvements decrease significantly.

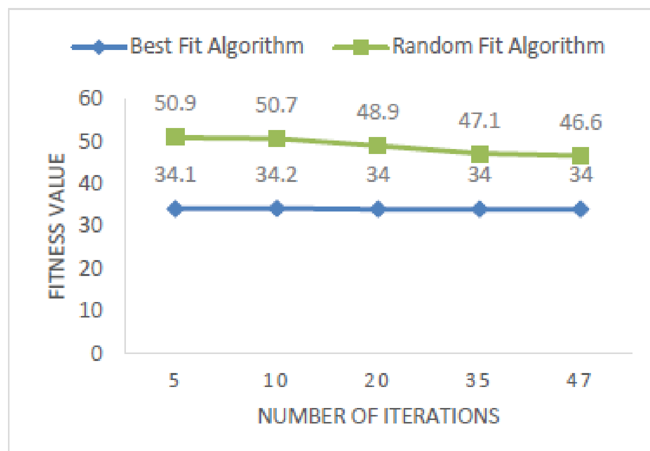


Fig. 6. Effect of Number of iterations on the Fitness Score.

D. Genetic Algorithm Approach vs. Iterated Local Search Approach

This section presents the results of GA approach vs. ILS approach comparison. For both approaches, the same initial solution was used and carried out GA process and ILS process separately.

Experiment 4: To evaluate the performances of GA and ILS, in terms of the fitness value, four scenarios were considered where the number of nurses were varied for each scenario. 30 rounds of experiments were conducted for each scenario, with GA and ILS separately. As shown in table 1 and figure 7 (averages of 30 experiment rounds), GA was able to find better solutions than ILS, in all four scenarios.

TABLE I. FITNESS VALUES COMPARISON (GA VS. ILS)

Number of nurses	Best Fitness Score	
	Genetic Algorithm	Iterated Local Search
15	49	67
20	22	25
25	12	16
30	9	10

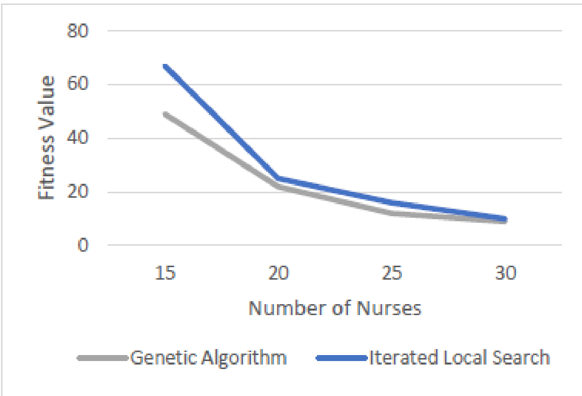


Fig. 7. Fitness Values Comparison (GA vs ILS).

TABLE II. TIME COMPARISON (GA VS. ILS)

Number of nurses	Time Taken (seconds)	
	Genetic Algorithm	Iterated Local Search
15	0.025986	0.027986
20	0.048973	0.049971
25	0.037977	0.051970
30	0.059965	0.070957

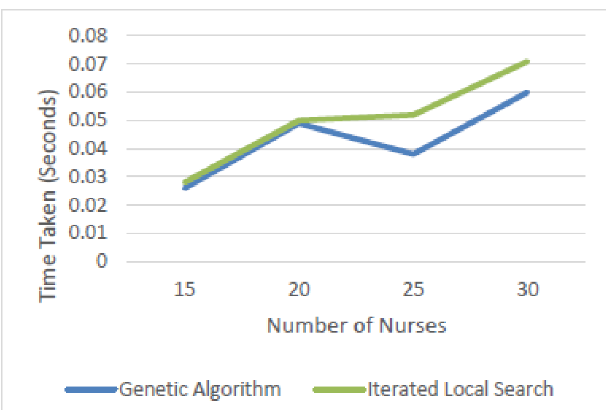


Fig. 8. Time Comparison (GA vs ILS).

Experiment 5: To evaluate the performances of GA and ILS, in terms of the time, once again, four scenarios were considered where the number of nurses were varied for each scenario. 30 rounds of experiments were conducted for each scenario, with GA and ILS separately. As shown in table 2 and figure 8 (averages of 30 experiment rounds), GA was able to find solutions faster than ILS, in all four scenarios.

V. FINAL REMARKS

We have proposed a platform for Dengue disease management with following modules: (1) a prediction module to predict the dengue outbreak and (2) an optimization algorithm module to optimize hospital staff according to the predictions made on future dengue patient counts.

This paper focuses on the optimization algorithm module to generate the hospital staff schedule, specially the nurse staff allocation, based on two approaches: (1) GA and (2) ILS. Experiments were conducted to measure the performances of the optimization algorithms using the statistics collected from IDH. With the real data sets, the results show that, GA and ILS can decide the optimal allocations dynamically in the order of seconds. Also, the results show that the GA approach is much efficient and faster than the ILS approach when considering the fitness value and the performance.

Comparing to similar surveys, this study depicts that the proposed optimization algorithm (GA) is not only fast but is also efficient and the same solution which we have developed using GA can be used to solve similar optimizing problems. As mentioned previously the experiments were conducted considering a limited dataset taking into consideration the situation of the IDH hospital. For further analysis, a larger dataset is essential to achieve improved results.

As the future work, we are planning to work on the prediction module to complete our proposed platform for Dengue disease management

REFERENCES

[1] AbirAlharbiand KholoodAlQahtani, A Genetic Algorithm Solution for the Doctor Scheduling Problem, ADVCOMP 2016.

[2] Matteo Sartori, Dott.ssa Maria Silvia Pini, a local search algorithm for matching hospitals to residents, Anno Accademico 2013/2014

[3] W. Rankothge, F. Le, A. Russo, J. Lobo, " Experimental results on the use of genetic algorithms for scaling virtualized network functions", Proc. IEEE SDN/NFV, pp. 47-53, 2015.

[4] W. Rankothge, J. Ma, F. Le, A. Russo, J. Lobo, "Towards making network function virtualization a cloud computing service", Proc. IEEE IM, pp. 89-97, 2015.

[5] W. Rankothge, F. Le, A. Russo, J. Lobo, " Optimizing Resource Allocation for Virtualized Network Functions in a Cloud Center Using Genetic Algorithms ", Proc. IEEE IM, pp. 89-97, 2015.

[6] N. Nipunika, "Requirement Gathering and Data Collection," smartScheduler, 13-Aug-2018. [Online]. Available: http://smartscheduler.blogspot.com/2018/08/blog-post.html. [Accessed: 14-Aug-2018].

[7] M. Jayasuriya, "Genetic Algorithm - Test Results," smartScheduler, 13-Aug-2018. [Online]. Available: http://smartscheduler.blogspot.com/p/genetic-algorithm.html. [Accessed: 14-Aug-2018].

[8] K. Galappaththi, "Iterated Local Search - Test Results," smartScheduler, 13-Aug-2018. [Online]. Available: http://smartscheduler.blogspot.com/p/iterated-local-search-test-results.html. [Accessed: 14-Aug-2018].

[9] M. Mitchell, An Introduction to Genetic Algorithms, Cambridge, MA, USA: MIT Press, 1998.

[10] A. Ishigaki and S. Takaki, "Iterated Local Search Algorithm for Flexible Job Shop Scheduling", 2017.

Industrial Internet of Things as a Challenge for Higher Education

Corneliu Octavian Turcu[1]
Computers, Electronics and Automation Department
Stefan cel Mare University of Suceava
Suceava, Romania

Cristina Elena Turcu[2]
Computers Department
Stefan cel Mare University of Suceava
Suceava, Romania

***Abstract*—This paper is aimed to examine the adoption of the Internet of Things (IoT) in industry (so-called Industrial Internet of Things, shortly IIoT) and the requirements for higher education in the times of the fourth industrial revolution. The addition of the fourth letter, "I" in front of the "IoT" coins the name of the new concept, "IIoT" in relation with another term, "Industry 4.0". Because these concepts have no precise and widely accepted definitions, we presented some considered relevant by scientific literature. The paper also highlights the most important similarities and differences between these concepts. IIoT is a very dynamic concept and it will constantly bring changes in digital technologies, requirements and markets, and will also transform industries and business practices. According to manifold studies, currently, there is a skill gap which may widen in the future if no action is taken. Higher education must adopt the latest related technologies and must adapt to the new ways in which people, machines, services and data can interact. Consequently, employees, students, graduates, etc. have to be equally dynamic in learning and acquiring new skills. The transition from higher education to employment is a challenge that could be more easily addressed through the efforts of all stakeholders, from individuals to organizations, and from businesses to governments. As changes in higher education take time, all stakeholders will now have to act in preparing for the Industrial Internet of Things.**

***Keywords*—Industry 4.0; Industrial Internet of Things; Internet of Things; higher education; skills gap**

I. Introduction

Industrial engineering is constantly bound to adapt to the many occurring changes, from progress in business models to the most advanced information and communications technologies; the purpose is to increase the overall quality of products and productivity, and also to reduce overall costs. Currently, we are witnessing the rise of a new digital industrial wave, namely the fourth industrial revolution (IR 4.0 or FIR), enabled by the widespread deployment of inexpensive smart sensors, processors, wireless sensor networks, embedded systems, but also by the advances in data storage, analytics, cloud infrastructure, and so on. Various worldwide surveys conducted in relation to the industry field reveal that the biggest current technological initiative for implementing this revolution is the Industrial Internet of Things (IIoT).

In a report [1], Accenture estimates that the Industrial Internet of Things could add $14.2 trillion to the global economy by 2030. The global Industrial Internet of Things market is anticipated to expand at a CAGR of +24% during the 2018-2022 [2]. A McKinsey report [3] anticipates that by 2025, the percentage of factories adopting IIoT will reach 65%-90% in advanced economies and 50%-70% in developing economies.

Researchers estimate that the IIoT development will impact different sectors, influencing both the industry and the labor market, leading to the creation of new jobs, but also to the elimination of some existing ones. Thus, various studies on technology uptake conducted in different countries reveal that the adoption of new technologies is expected to have a significant impact on the employment landscape. For example, according to [4], "in many industries and countries, the most in-demand occupations or specialties did not exist 10 or even five years ago, and the pace of change is set to accelerate". The same report estimates that "65% of children entering primary school today will ultimately end up working in completely new job types that don't yet exist". This is also the case of the Industrial Internet of Things. Various publications have focused their attention on this new concept, analyzing, among others, both the multiple possible benefits and challenges generated by the application of this paradigm across different economies. In addition to the technological barriers, the widespread and accelerated adoption of the Internet of Things (IoT) paradigm in the industrial field is hampered by the skills gap. According to manifold studies, currently, there is a skills gap which may widen in the future if no action is taken. For example, an analysis from Deloitte [5] reveals that over the next decade, there will be 3.5 million job openings in manufacturing, but only enough skilled labor to fill less than half of them. And as IIoT growth takes hold, the need for skilled workforce will intensify. One solution to reduce the skills gap lies in education, for example through effective skills, re- and up-skilling programmes. IIoT is a very dynamic concept and it will constantly bring changes in digital technologies, requirements and markets, and will also transform industries and business practices. Consequently, employees, students, graduates, etc. have to be equally dynamic in learning and acquiring new skills. Higher education must adopt the latest related technologies and must adapt to the new ways in which people, machines, services and data can interact.

This paper is concerned specifically with the importance of higher education in supporting the development of the skills and competencies required for the Industrial Internet of Things era. Currently, the pictures of Internet of Things, Industrial

Internet of Things and Industry 4.0 are still quite blurry. Although IoT, IIoT and Industry 4.0 are closely related concepts, they cannot be interchangeably used. So far, there is no generally accepted definition for each of these terms and in an attempt to understand these concepts, this paper tries to clarify them. The literature in the field proposes several definitions, some of them being presented in a section of this paper. Next section provides a short comparison of IIoT and IoT. The IIoT development is being enabled by the improved availability and affordability of inexpensive (smart) sensors, processors, embedded systems, etc., but also by the developments in data storage, analytics, cloud infrastructure, and so on. Standardization plays an import role for the further development and spread of IIoT. In addition to listing the important IIoT enablers, we also highlight the opportunities and challenges brought forth by IIoT. The next section outlines some possible directions for future research and the paper ends with some concluding remarks.

II. IoT, IIoT, Industry 4.0

A. *IoT, IIoT and Industry 4.0 Concept Disambiguation*

The concepts of Internet of Things, Industrial Internet of Things and Industry 4.0 do not have precise and widely accepted definitions. The literature in the field proposes several of them, some being presented in the following paragraphs.

Internet of Things:

- is "enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies" [6].
- "is not a single technology; rather it is an agglomeration of various technologies that work together in tandem" [7].

The Industrial Internet of Things, viewed as an industrial adaptation of the Internet of Things, has many naming variations— "Industrial Internet" as GE terms it, "Internet of Everything" term proposed by Cisco, Rockwell Automation's "IoT Industrial Revolution", IBM's "Smarter Planet" or the European "Industry 4.0" or "Industrie 4.0" (original German term), respectively French "Industrie du Futur" (Industry of the Future). The term Industrial Internet of Things (abbreviated as IIoT, I-IoT, I2OT or I2oT) is used to represent what is expected to revolutionize the industry, by merging the digital and real industry [8]. The Industrial Internet of Things is defined as follows:

- "a universe of intelligent industrial products, processes and services that communicate with each other and with people over a global network" [9].
- "is a distributed network of smart sensors that enables precise control and monitoring of complex processes over arbitrary distances" [10].

Industry 4.0 or Industrie 4.0:

- "focuses on the end-to-end digitisation of all physical assets and processes as well as integration into digital ecosystems with value chain partners" [11].

According to [12], the Industrial Internet of Things is related with Industry 4.0: "all Internet of Things applications in Industry 4.0 are forms of IIoT, but not all IIoT use cases are about the industries which are categorized as Industry 4.0".

In many cases, the terms Industry 4.0 and Industrial Internet of Things are used interchangeably [13], despite the differences between them. But, according to [14], "Industrial IoT and Industry 4.0 essentially have a cause-and-effect relationship. That is, the Industrial IoT is the basis for, and will result in, the fourth industrial revolution". Industry 4.0 or i-4.0 initiative is focused specifically on the manufacturing industry and ensuring its competitiveness in a dynamic global market. The IIoT concept is "more focused on enabling and accelerating the adoption of Internet-connected technologies across industries, both manufacturing and non-manufacturing" [15]. In fact, IIoT is viewed as a key enabler for the 4th industrial revolution (IR 4.0).

The term coined by the addition of the fourth letter, "I" in front of IoT describes the new concept, IIoT. There are some common points between these concepts, IoT and IIoT, but there are also notable differences, some of which are being presented in the next section of this paper.

B. *Industrial Internet of Things vs. Internet of Things*

Although the Industrial IoT follows the same core definition of the IoT and share some common characteristics (Table 1), these paradigms are also distinguished by different features (Table 2).

TABLE I. IoT vs. IIoT - Similar characteristics

Perspective	**Internet of Things and Industrial Internet of Things**
Availability of connected devices	availability of affordable and intelligent devices; connectivity brings the true value of IoT and IIoT, ensuring monitoring, control, etc.; although they are used to reach different goals
Architecture	[Collect | Store | Analyze | Share] architecture
Technologies involved	cloud computing, communication technologies, (smart) sensors technologies, big data, advanced analytics, machine learning, etc.
Challenges	the multitude of available standards and technologies, interoperability, security and privacy protection, skill shift for users, lack of experience and expertise for developers, acceptance, capital

TABLE II. IOT VS. IIOT - DIFFERENT CHARACTERISTICS

Perspective	Internet of Things	Industrial Internet of Things
Connected things	consumer-level devices, usually less expensive	critical machines, sensors, systems, usually with a high degree of complexity
Service model	human-centric	machine-centric
Applications	consumer-oriented applications	industry-oriented applications
Communication infrastructure	essentially wireless	wireless and wired
Communication capabilities	a small number of communication standards	a high number of connectivity standards and technologies
Amount of data	medium to high	high to very high
Criticality	not stringent	mission critical (timing, reliability, security, privacy)
Real-time requirement	usually no, dealing with less time-sensitive systems	most often has a key role

C. Industrial Internet of Things Enablers, Benefits and Challenges

There are various IIoT enablers: software and hardware technologies; business organization and culture; standardization; public policy; skilled human resources, detailed in Fig. 1. Thus, for example, in order to implement flexible, customizable, and efficient industrial systems, completely aligned with the IIoT concepts, all related enabling technologies (industrial networks, cloud computing, big data, artificial intelligence, data analytics, blockchain, etc.) and hardware systems (embedded systems, industrial robots, etc.) must be equally developed for being integrated into IIoT systems [16].

IIoT platforms have played and will play an important role in supporting the development and implementation of IIoT applications, by reducing the complexity of various phases of the application lifecycle like, for example, development, deployment, etc. Currently, there are more IIoT platforms available [17]. As IIoT becomes better defined and continues to grow and to be further developed, more impactful and powerful IIoT applications can and will be created and deployed; development of these applications may or may not use an IIoT platform.

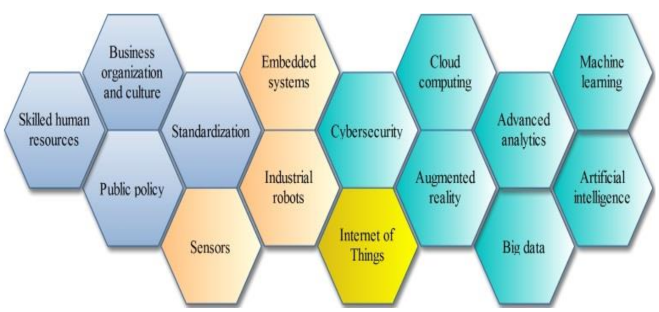


Fig. 1. IIoT Enablers.

IIoT could bring various benefits in an industrial context, some of the key ones being the following: monitoring production flow and inventory; enhancing automation, productivity, industrial safety, efficiency, security and quality control; enabling easy maintenance, inventory management, products tracking and tracing, development of new business models, services and/or products; optimization of packaging, logistics and supply chain; reduction of human errors and manual labor, and of costs (both in terms of time and money), etc.

Besides all the benefits, the large-scale deployment of IIoT is beset with various challenges. The key challenges "stem from the requirements in energy-efficient operation, real-time performance in dynamic environments, the need for coexistence and interoperability, maintaining the security of the applications and users' privacy" [18], establishing connectivity, lack of standardization, etc. Besides technical challenges, there are broader social, economic and political enabling factors that have to be overcome. Thus, the adoption of the Internet of Things paradigm in the industrial field is delayed by the so-called skills gap. The literature reports that currently, Industrial Internet of Things represents a cornerstone of many companies' business strategies. Also, many agreed that rapidly developing IIoT will have significant and far-reaching effects on the industrial and market sectors, transforming and redefining virtually all markets and industries; the effects will also generate changes in other fields, such as education.

III. IIOT HIGHER EDUCATION

In order to address higher education in the context of the Industrial Internet of Things, we have in view a holistic view of all IIoT-related higher education perspectives from:

- A scientific perspective – universities are conducting research both in the IIoT field and the IIoT-related fields, thus contributing to the expansion of knowledge horizons;
- An educational perspective - through their teaching, universities actively support the dissemination of IIoT know-how amongst students and graduates, thus improving the stock of skilled human capital;
- A technological perspective - universities support the transfer of their know-how to industry, by focusing on technology transfer.

Next, we are trying to present some aspects of IIoT higher education, by taking into account the above mentioned dimensions.

A. The Scientific Perspective

In recent years, there has been a growing interest in Industrial Internet of Things and IIoT, as Google Trend shows (Fig. 2). We can visualize the relative popularity of these keywords between 2008 and 2018.

Moreover, the number of publications that addressed Industrial Internet of Things is quickly growing. Some of these publications present the findings of the research carried out in universities. In order to summarize the current status of scientific publications on the use of IoT in industries, the authors of this paper carry out an extensive literature review by examining important articles from six relevant academic databases (Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, Scopus, ACM digital library). As a result, we found a large number of journal articles and conference papers, books and book chapters related to the existing IIoT research. For example, we found 453 IIoT-related scientific articles and books/chapters published from 2011 to 2018 just by searching the Web of Knowledge database. We also identified the publications where at least one author has an academic affiliation. Fig. 3 displays the number of scientific publications stored in the above-mentioned databases, by taking into consideration two categories for author's affiliation: academic and non-academic. The figure highlights the noticeably growing trend of IIoT research in the recent years and as well as the important role played by academic authors.

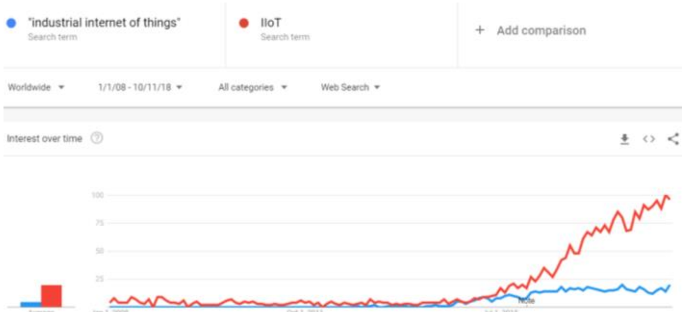


Fig. 2. Search Volume Index for the Data Provided by Google Trends Corresponding to the Industrial Internet of Things and IIoT Search Terms.

TABLE III. THE NUMBER OF SCIENTIFIC PUBLICATIONS ON INDUSTRIAL INTERNET OF THINGS

Search	Web of Science	IEEE	Science Direct	Springer Link	Scopus	ACM digital library
Industrial Internet of Things	453	461	63	442	725	23
Industrial Internet of Things and academic affiliation	187	316	50	397	509	16

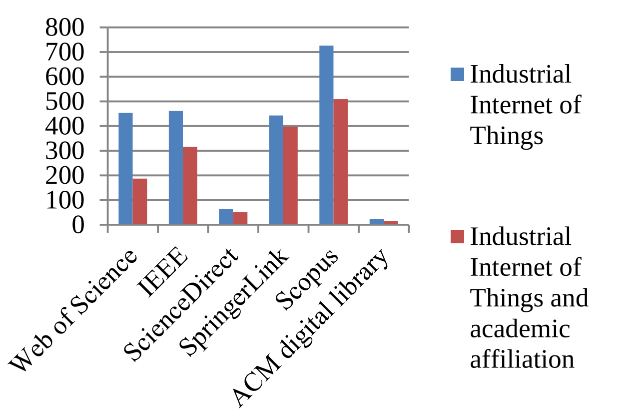


Fig. 3. Number of Scientific Publications Related to the IIoT.

B. The Educational Perspective

Secondly, we have in view the educational dimension of the IIoT Higher Education.

Various worldwide studies, reports and scientific papers reveal the fact that the biggest challenge of the Industrial Internet of Things supporters is not so much technology itself but the people. While digital technologies are rapidly becoming a commodity, success largely depends on an organization's Digital IQ [19]. Therefore, it is critical to improve the digital skills of the employees who need to roll out digital processes and service [20].

Based on the existing trends, experts predict that profound changes starting from the content to the delivery within the Internet of Things and Industrial Internet of Things are needed in regards to the main aspects of education. In order to meet the changing requirements and respond to the increasing demand for a future highly skilled workforce, new effective educational programmes will have to be improved or developed and/or the existing academic curricula will have to be re-structured. Furthermore, higher education institutions have to overcome the traditional way of learning, by embracing the latest technologies in order to innovate the learning process. According to [21], workers who are skilled at developing and deploying IIoT systems will find themselves in greater demand. Digital competences and innovation are widely considered to be some of the key drivers to boost the competitiveness of companies within the IIoT context. In this respect, disruptive technologies, such as those listed in the previous section, which have been leveraged for the development of IIoT, should be viewed as building blocks for the basis of a curriculum adjusted to meet the IIoT's requirements. Furthermore, according to the recommendations of various studies and reports, universities must equip students with attributes that will enable them to respond to an uncertain future – by "handling ambiguity, emotional intelligence, adaptability" [22]. It is a fact that the focus on the technical development of the individual is just a part of the IIoT's educational vision. According to [23], "the ongoing reorganization and realignment of education systems" will also be of a paramount importance.

The transition from higher education towards employment is a challenge that could be more easily addressed through the efforts of both universities and employers, as part of universities-employers partnerships. In fact, the importance of employer-university relationships is also supported by the graduate employment rate as a key performance indicator for universities. Consequently, universities have increased their focus both on supporting the employability rate of students and graduates, and on their partnership with the business and industry sectors. The purpose is twofold: to prepare students for the labor market as well as to increase the graduate employment rates. In fact, some universities have to work harder and in more innovative ways to attract potential employers. A solution that has highly been used in the recent years is to involve the private sector in developing the academic curricula, based on specific needs. Moreover, placements and internships in various companies give students a chance to gain work experience.

Yet, as experts point out, the current approaches regarding the Industrial Internet of Things education and training do not rise up to the challenge [24]. Thus, in order to shape up an enhanced environment for Industrial IoT education, the training needs in different domains have to be investigated in a more detailed manner. Academic and Higher Vocational Education and Training (HVET) programs, work-based education, virtual and remote laboratories for educational purposes, national and international educational policies and tools, extensions, etc. should be considered.

C. The Technological Perspective

No doubt that the progress made in the field of IIoT technologies has and certainly will have the potential to transform the industrial sector. Industrial Internet of Things is a concept with a high level of innovation. Over the past few years, we have seen many innovations that have been designed as solutions to the new challenges posed by the IIoT and Industry 4.0. Yet, technology transfer is essential for imparting the information and adapting these innovations to real life situations. Technology transfer is defined by the Association of University of Technology Managers (AUTM) as "the process of transferring scientific findings from one organization to another for the purpose of further development and commercialization". Worldwide, there is a large number of studies that tackle the university technology transfer process.

The active academic technology transfer could be of service to a university, the region and the country, industry partners, and the public in a numbers of ways. Some of them have been revealed in an extensive number of scientific papers and studies. Thus, for example, we could mention the article from the National Academy of Inventors that attempts to examine the importance of "technology transfer" for universities [25].

Without simplifying the overall technology transfer process from the point of view of the variety of activities that technology transfer process encompasses, the present study focuses solely on patenting in relation to the Industrial Internet of Things within the Fourth Industrial Revolution.

Based on the latest available information on patent applications on the Fourth Industrial Revolution that has been available until 2016, the European Patent Office (EPO) conducted a landscaping study on "Patents and the Fourth Industrial Revolution (4IR)" [26], by providing a first cartography of this dynamic technological field based on patent activity. According to this study, "European patent applications related to smart connected objects are rising rapidly, achieving a growth rate of 54% in the last three years" [26]. The EPO report reveals that despite the fact that the growth rate of patent applications in this field is higher than in any other field of technology and is increasing at a fast rate, the patent applications still represent a relatively modest percentage of all European patent applications (only 3,3%). Moreover, the 25 biggest 4IR applicants for the EPO in between 2011-2016 are large, traditionally ICT-focused companies.

In order to support IIoT innovation with quality patents, universities must focus their efforts in this direction and encourage sharing the results gathered by their own researchers through licensed patents. In addition, besides stimulating students' creativity, providing students with patenting skills is very important. Academic inventions and technology transfer could bring significant benefits to students, by providing them opportunities "to participate in real-world translational research, gain experience in the process of obtaining a patent, and work with industry, start-ups, and manufacturers" [25].

IV. Future Research

As changes in higher education take time, all stakeholders will now have to act in preparing for the Industrial Internet of Things. An analysis regarding the Industrial Internet of Things higher education should also take into consideration other aspects, which have not been presented in this paper, due to space limitations. A comprehensive study on how to develop an IIoT education system should consider and address existing approaches, the skills required and also the appropriate contexts. New digital workforce models could be developed. Also, the study should focus heavily on all the educational activities related to IIoT, including academic and Higher Vocational Education and Training (HVET) programs, work-based education, virtual and remote laboratories for didactic educational purposes, but also national and international educational policies and tools, etc. Alongside the realignment of university curricula, the main findings should focus on a comprehensive strategy, to develop good practices and recommendations to reform education so that it becomes more responsive to the requirements of the future workforce; the purpose is to further build and strengthen the collaboration between education and the labor market, in order to consolidate future employment trends.

V. Conclusions

This paper is aimed to examine the 'Industrial engineering education', focusing on the adoption of the Internet of Things in industry (Industrial Internet of Things) and the requirements for higher education in the times of the fourth industrial revolution.

The pictures of Internet of Things, Industrial Internet of Things and Industry 4.0 are still quite fuzzy. Because these concepts have no precise and widely accepted definitions, we presented some considered relevant by scientific literature. The paper also highlights the most important similarities and differences between these concepts.

IIoT can bring important benefits to companies in all types of industries. To seize the opportunities brought forth by IIoT, several obstacles would need to be overcome. Thus, for example, besides focusing only on the technical issues, the relationships between all stakeholders, from individuals to organizations, and from businesses to governments, must also be taken into consideration. Over the coming years, business and governments must intensify their efforts and escalate investments, but also must change their approach to education, skills and employment. Higher education must also intensify efforts aimed at tackling skills gaps and it must take additional measures in order to anticipate and address in a timely manner the issues related to the potential large-scale transformations of the employment landscape and the 'holistic' nature of skills requirements. In this respect, developing stronger employer-university ties for internships, training needs assessment, etc. is a valuable strategy needed to be encouraged and extended in the near future.

ACKNOWLEDGMENT

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

REFERENCES

[1] Daugherty, P., Berthon, B., Winning with the Industrial Internet of Things: How to Accelerate the Journey to Productivity and Growth, Dublín: Accenture, 2015.

[2] MRS Research Group. Global Industrial Internet of Things (IIoT) Market 2017 - Production, Sales, Supply, Demand, Analysis & Forecast to 2021, 2017.

[3] Manyika, J., Chui, M., Bisson, P., Woetzel, J., Dobbs, R., Bughin, J., Aharon, D., The Internet of Things: Mapping the value beyond the hype, McKinsey & Company, San Francisco, 2015, Available at: https://www.mckinsey.com/.

[4] *, The future of jobs: Employment, skills and workforce strategy for the fourth industrial revolution. In World Economic Forum, 2016.

[5] Giffi, C., McNelly, J., Dollar, B., Carrick, G., Drew, M., Gangula, B., The skills gap in US manufacturing: 2015 and beyond, Washington, DC: Deloitte and Manufacturing Institute, 2015.

[6] *, Internet of Things Global Standards Initiative. International Telecommunications Union, 2012.

[7] Sethi, P., Sarangi, S. R. Internet of things: architectures, protocols, and applications. Journal of Electrical and Computer Engineering, 2017.

[8] Hartmann, M., Halecker, B., Management of innovation in the industrial Internet of Things. In ISPIM Conference Proceedings (p. 1). The International Society for Professional Innovation Management (ISPIM), 2015.

[9] Banerjee Prith, Industrial Internet of Things Poised to Drive Unconventional Growth, 2014.

[10] Huberman, B. A., Ensuring Trust and Security in the Industrial IoT: The Internet of Things (Ubiquity symposium), 2016.

[11] "Industry 4.0: Building the digital enterprise", PwC, 2016.

[12] *, The Internet of Things (IoT) - essential IoT business guide, i-Scoop 2016, Available at: https://www.i-scoop.eu/internet-of-things-guide/

[13] Kagermann, H., Helbig, J., Hellinger, A., Wahlster, W, Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0: Deutschlands Zukunft als Produktionsstandort sichern; Abschlussbericht des Arbeitskreises Industrie 4.0. Forschungsunion, 2013.

[14] Albert, M. (2015). Seven things to know about the internet of things and industry 4.0. Modern Machine Shop, 88(4), 74–81.

[15] De Bernardini Luigi, Industry 4.0 and IIoT: different approaches to a smarter industry, 2016.

[16] Wan J., Lai C.F., Song H., Imran M., Jia D., Software-Defined Industrial Internet of Things, Wireless Communications and Mobile Computing, 2018.

[17] *, Industrial Internet of Things Market Analysis 2018 To 2022. Market Research Future, 2018.

[18] Sisinni, E., Saifullah, A., Han, S., Jennehag, U., & Gidlund, M. (2018). Industrial Internet of Things: Challenges, Opportunities, and Directions. IEEE Transactions on Industrial Informatics, 1–1. https://doi.org/10.1109/TII.2018.2852491

[19] PwC, Global Digital IQ Survey. PriceWaterHouse Coopers. September 2015.

[20] PwC, 2017 Global Digital IQ Survey: A decade of digital. Keeping pace with Transformation, 10th anniversary edition, PriceWaterHouse Coopers, 2017.

[21] Manyika, J., Chui, M., Bisson, P., Woetzel, J., Dobbs, R., Bughin, J., & Aharon, D., Unlocking the Potential of the Internet of Things. McKinsey Global Institute, 2015, Available at: https://www.mckinsey.com/.

[22] Edmondson M., Ward A., Tackling the disconnect between universities, Small businesses and graduates in cities and regions. Gradcore, 2017

[23] Broadband, Working Group on Education: Digital skills for life and work, Broadband Commission Working Group on Education, September 2017.

[24] O'Halloran, D., Kvochko, E., Industrial internet of things: unleashing the potential of connected products and services. In World Economic Forum (p. 40), 2015.

[25] McDevitt, V. L., Mendez-Hinds, J., Winwood, D., Nijhawan, V., Sherer, T., Ritter, J. F., Sanberg, P. R., More than money: The exponential impact of academic technology transfer. Technology & Innovation, 16(1), 75–84, 2014.

[26] EPO, Patents and the Fourth Industrial Revolution. European Patent Office, Handelsblatt Research Institute. Available at: https://www.epo.org/service-support/publications.html?pubid=163#tab3

Coverage-Aware Protocols in Wireless Sensor Networks: A Review

Jahangir Khan[1], Khalid Mahmood[2], Ansar Munir Shah[3], Babar Nawaz[4], Mahmood ul Hassan[5]
Department of Computer Science,
IIC University of Technology,
Phnom Penh, Cambodia

***Abstract*—Coverage plays a vital role in wireless sensor networks (WSNs), since it is used as one of the important measure to achieve the performance of the sensor network. The sensor nodes in WSN have limited power and energy resources. So, energy efficiency is an essential factor that should be considered along with coverage while designing the coverage protocols. During the past few years, many efforts have been made by the researchers on designing different coverage-aware protocols. Different coverage-aware protocols may impose different ways to solve the coverage and efficient utilization of energy among sensor nodes. In this paper, we present a review on coverage-aware protocols by highlighting their functionalities.**

Keywords—Wireless sensor network; coverage area; energy efficiency; coverage protocols

I. Introduction

In recent years, wireless sensor networks (WSNs) have been implemented in many application areas like wild fire detection, habitat monitoring, battlefield, industrial, agricultural and medical etc. [1]. In WSN, large numbers of sensors are deployed in the field of interest. The deployed sensors are low cost, small in size, and have limited power and communication capabilities. Moreover, sensor nodes communicate with each other and can cover the anticipated field of interest. In WSN, sensor nodes have the capability to sense the activities in an area of interest, process them, and can forward them to the base station [2]. The most essential factor to install an efficient sensor network is to search the best node installment schemes and efficient topology approaches [3] [4]. To manage the connectivity coverage is very essential for a sensor network.

Mostly sensors are supplied with limited energy resources. In particular applications, nodes are deployed in rigid surrounding, from where it is tough to recharge or change the batteries. Without any predefined planning, nodes are randomly placed in a desired region. Once nodes are deployed, they can automatically configure the network. In sensor network applications, numerous sensors are placed in the territory. Hence, there might be a chance that one or more sensor nodes are tracking a covering region. So, the data sensed by more than one sensor nodes have a small amount of redundancy. The optimal approach is to deploy lowest number of sensors in an area to get full coverage.

The life cycle of a sensor network can be determined by the time period in which a network is capable to achieve the sensing and communication task with the base station (BS) [5]. At the same time, a few sensor nodes may fail due to the hardware breakdown or power deficiency. Meanwhile there might be a chance of deploying more nodes to sustain the suitable coverage degree. In addition, when there is no more coverage possible and sensing voids appear then the coverage-aware protocols need to manage the sensing voids. The coverage redundancy appears in WSN, if more than one sensor nodes are sensing the same region which would result in more consumption of energy. Due to the limited energy resources in WSN, it is essential to detect the active redundant nodes and put them in OFF state unless they are needed to be in ON state. Due to the small power battery, we must use the sensor nodes in an efficient way so as to maximize the network life cycle [6].

The biggest key issue raised recently is how to deal with the coverage. Based on the object covering, the coverage issue can be categorized into path, region and target coverage. Each point in the region must be tracked by minimum number of nodes. To achieve the best coverage, the nodes should be placed with particular amount in order to cover the complete region. Therefore, coverage-aware protocols are needed to be designed in such a way that they can achieve full coverage with efficient utilization of power and communication resources. This paper provides a comprehensive review on coverage-aware protocols.

Rest of the paper is organized as follows. Section 2 highlights the coverage and connectivity problems based on the energy consumption and deployment strategies. Section 3 briefly provides an overview on coverage-aware protocols based on comparison study. Finally the section 4 concludes the review article.

II. Literature Review

Wireless sensor network is configured in an area of interest to provide the best coverage along with efficient utilization of energy. The proposed schemes for coverage in a wireless sensor network must consider several factors. The capacity of sensors should be reviewed. Mostly researchers emphasize on only one deployment model. Every sensor has a specific covering capability to cover an object in the deployed field. A node can cover a small range of field. Hence, in a WSN various sensing models can be composed according to the needs of the environment.

Some sensor nodes have uniform communication ranges; many sensors' radio transceiver has the capability of altering communication power continuously to acquire various transmission ranges. After the deployment of sensor nodes either distributed or centralized algorithm is run to find the sufficient coverage in the field. The coverage algorithm executes in a center node. Data from entire sensor nodes requires to be forwarded to central node. In distributed approach, the coverage algorithm is run in WSN which depends upon data from various nodes.

Mostly a huge number of sensors are placed in a territory to be supervised. In WSN, numbers of sensors are generally more than mobile-adhoc network [7]. In a WSN, placement of nodes can generally be classified as a dense deployment or a sparse deployment. In dense deployment, large numbers of sensor nodes are scattered in the field of interest. On the other hand in sparse deployment, few nodes are scattered in the field of interest. The deployment of sensor nodes in a WSN can be deterministic or random as shown in Fig. 1 and Fig. 2. In random approach it is compulsory to place extra nodes to get the full coverage.

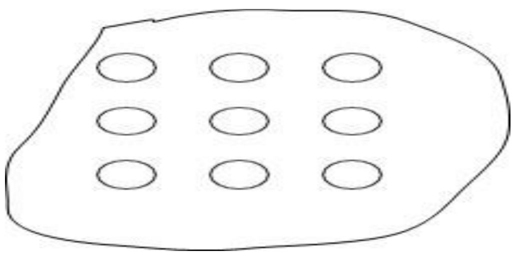


Fig. 1. Deterministic Deployment of Nodes.

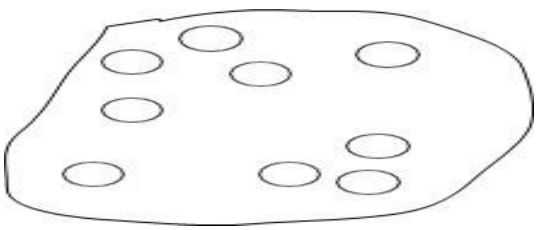


Fig. 2. Random Deployment of Nodes.

The following section provides a brief review on coverage-aware protocols for WSN.

A. *Low-energy Adaptive Clustering Hierarchy (LEACH)*

The LEACH protocol is a cluster-based scheme for small sensor networks which integrates the concepts of media access and energy- efficient cluster based routing that is combined with application specific data collection in order to get the application perceived quality, latency and good performance based on life cycle of the network [8]. The objective of LEACH is to switch off the nodes that are non-head nodes as much as possible. The function is distributed into periods and each period contains a steady phase and set-up phase. At the same time, set-up phase selects the cluster heads and every node gets a cluster by selecting the cluster head which needs small amount of transmission energy.

The limitation of LEACH is that the cluster head node consumes more energy as compared to non-cluster head nodes. The failure of cluster-head node breaks the communication of all of its connected nodes.

B. *Coverage Configuration Protocol (CCP)*

CCP is a decentralized protocol. It proposes that every sensor has a static circular covering area, having sensor at center [9]. CCP protocol also proposed that every sensor know the information of its location correctly. For the required coverage degree, CCP protocol configures a network. Various degrees of sensing coverage are needed by many of the applications, in which each location of area is tracked by many nodes. Each sensor node can be in one of the three modes in CCP protocol; active, listen and sleep. Sensor node actively covers the field for an object in an active mode. The node switch off its radio to preserve energy in sleep mode. The sleeping node repeatedly goes into listening mode to check whether to enter into active mode or not. The node remains in an active mode until or unless the CCP switch off condition is true. Each sensor node in CCP requires managing a neighborhood table. In this way node can find the coverage overlap to inquire switch off criteria.

The CCP is not energy efficient due to the fact that it consumes more energy with frequent switching of nodes from switch-on to switch-off and vice versa. Moreover, CCP cannot overcome the sensing holes due to low number of active nodes.

C. *Optimal Geographical Density Control (OGDC)*

OGDC is a density control protocol for full coverage. It finds the minimal number of engaging nodes [10]. OGDC makes use of localization methods in which nodes have the information of their exact position [11] [12]. It is also assumed in OGDC that the broadcasting range is at least double the covering range. In this protocol sensors can have one of three modes; OFF, ON and UNDECIDED. The procedure executes in cycles. In the starting of first cycle, all sensors join to choose an active node based on a switch on condition and wake up, adjust their mode to UNDECIDED. The first sensor of every cycle is selected randomly in decentralized way. All nodes change their modes to ON or OFF at the end of the execution of a cycle and still in that mode till starting of the second cycle.

D. *Lightweight Deployment-Aware Scheduling (LDAS)*

The objective of LDAS protocol is to sustain sufficient covering range and maximize the life time of WSN [13]. Among neighboring wireless sensors, the procedure of LDAS protocol evaluates the redundant covering area. As LDAS proposed that sensors have no attach device like GPS to get the information of location, on the probability of average partial and complete redundancy LDAS give tight upper and lower bounds.

E. *Obstacle-Resistant Robot Deployment (ORRD)*

The ORRD protocol utilizes one robot in placing static sensors for gaining full coverage with minimum energy consumption and deployment time [14]. The robot node in ORRD consumes less energy and provides better coverage with minimal deployment time. But, the algorithm gets stuck in dead edges with early dropped nodes or obstacles [15].

F. *Enhanced Configuration Control Protocol (ECCP)*

The coverage area of a WSN is decreased with the presence of sensing voids [15]. To avoid sensing void the CCP protocol is inadequate. In a sensor network to prevent from sensing voids ECCP algorithm gives a method however, additional number of active sensor nodes are required. The working of CCP is enhanced with the addition of a condition that the borders of the targeted area are sensed by neighboring sensors.

Although to maintain the more number of needed active nodes, ECCP protocol make sure that in the region there is no sensing void. It also ensures the full coverage of region. The limitation of ECCP is that additional active sensors involve as compared to CCP due to the ECCP's extra node switch off conditions.

G. Fuzzy Based Priority Coverage (FBPC)

FBPC coverage protocol is introduced in [16] to enhance the coverage in WSN. The movement of nodes depends upon the obstacles, distance from the borders and number of neighbors. The procedures in the proposed protocol depend on fuzzy inference engines and virtual forces. However, the average moving distance is relatively very long in the proposed protocol which consumes more energy resources.

H. Time Constrained Targets Patrolling (TCTP)

The TCTP algorithm is proposed, in which every target is designated a weight and moveable sensors which monitor every target according to its weight [17]. The algorithm establishes single path, and make sure that targets having high weight are visited continuously. The impairment in any section of patrol path badly affects the performance of the algorithm. Additionally, to sense all the targets the number of mobile sensors may not be enough.

I. Random Backoff Sleep Protocol (RBSP)

RBSP is a search based protocol which uses data regarding the remaining energy level of the on-duty nodes [18]. For the calculation of Backoff Sleep Timer, RBSP applies a novel backoff algorithm. Depending on residual energy (RE) of active node, RBSP protocol makes use of sleeping window in order to select a random value evenly. Through this method, the possibility of a neighboring nodes' switching on status is very down whereas on-duty node has high remaining energy.

In RBSP, when the residual energy of the on-duty node is less than the specified threshold, the neighboring off-duty node wake-up continuously. RBSP minimizes the network lifetime and energy wastage due to the unneeded frequent wake-up of a sleeping node. RBSP makes sure that sufficient numbers of active nodes are available for a long period of time.

J. Discharge Curve Backoff Sleep Protocol (DCBSP)

The DCBSP protocol provides better coverage established on normalized standard battery discharge curve. This curve is used to find the back-off sleep-time for neighboring sensor nodes which are in sleeping mode [19]. In DCBSP, each sensor has three functioning modes that are likely to ACTIVE, RBSP SLEEP and FLOAT. At start all sensors are in sleeping mode. The sleeping node awakens and switch in to a floating mode upon the expiration of Back-off sleep timer.

DCBSP is a location-unaware and statistical based protocol. To avoid irregular and unwanted continuously wake ups of off-duty sensors, the optimal back-off sleep-time obtained from normalized generic discharged curve. As a result, off-duty nodes activate only near to the deactivation of on-duty nodes.

K. Distributed Lifetime Coverage Optimization (DiLCO)

DiLCO protocol uses two techniques; network leader selection and scheduling to increase the life cycle of network and sustains coverage [20]. Firstly, the divide and conquer methodology is applied on an area of interest in order to divide it into subareas. DiLCO protocol is periodic protocol in which a period begins with discovery phase to send and receive data among sensors of likely subareas, to select a node (head) to achieve coverage area. Each period consists of four phases namely Leader Election, Information Exchange, Sensing, and Decision. For every period it uses the one cover set for sensing task. Periodic scheduling increases the robustness against the failures of node. In periodic scheduling, a node is excluded from the scheduling process if it is out of energy or fails before taking the decision.

III. PERFORMANCE COMPARISON

The performance comparison of above mentioned coverage protocols is based on their coverage degree, sensing area coverage, characteristics, objectives and node scheduling. The comparison is shown in Table 1.

TABLE I. COMPARATIVE ANALYSIS FOR COVERAGE PROTOCOLS

References	Protocols	Parameters				
		Coverage Degree	***Sensing Area Coverage***	***Characteristics***	***Constraint and Objective***	***Node scheduling***
[8]	LEACH	NE	2-D Circular	Distributed	As much as possible switching off non head nodes	Optimized base on cluster
[9]	CCP	NE	NE	Distributed	Configured for guaranteeing coverage	Random
[9-12]	OGDC	Cd = 1	2.12 %(+)	Distributed	For full coverage finds the least number of engaging node	Periodic based on coverage
[13]	LDAS	Cd = 1	2-D Circular	Distributed	To sustain covering area and maximize the network lifetime	Random

[14] [15]	ORRD	Cd = 1	DN	Centralized	-Minimized the number of sensors -Deployment time and consumed energy	NE
[9]	ECCP	Cd = 1 and 2	DN	Distributed	to avoid sensing voids in network get a structure	Based on eligibility test
[16]	FBPC	Cd = 1	NE	Distributed	-Binary Sensing -Enhance initial random coverage	Random
[17]	TCTP	Cd = 1	NE	Centralized	Sweep coverage for covering a set of line segment	NE
[18]	RBSP	Cd = 1	12.5 %(+)	Distributed energy efficient	Evenly select a random number from sleeping window depends on remaining energy	Depends on remaining energy
[18][19]	DCBSP	Cd = 1	31.23 %(+)	Distributed K-neighborhood	Maintain the sufficient count of ACTIVE node	Optimized base on battery level

*NE = Not examined, DN = Data not available, (+) = Higher , (-) Lower

IV. Conclusion

In WSN, coverage redundancy can be minimized by using various types of coverage protocols. Coverage is the important problem in wireless sensor networks (WSNs). Many algorithms have been proposed for the solution of coverage issues during past years. In this review article, we presented a brief overview of coverage optimization protocols. We take coverage optimization protocols along with energy efficiency factors into consideration and explained the various working techniques of both with comparison. It is observed that the basic purpose of coverage protocols is to retain the essential group of working nodes by switching off the redundant nodes in order to save the energy. By minimizing coverage overlap we can attain the energy conservation that maximizes the lifespan of the network.

References

[1] Khalid Mahmood, Muhammad Amir Khan, Mahmood ul Hassan, Ansar Munir Shah, Shahzad Ali, and Muhammad Kashif Saeed, "Intelligent On-Demand Connectivity Restoration for Wireless Sensor Networks," Wireless Communications and Mobile Computing, vol. 2018, Article ID 9702650, 10 pages, 2018. https://doi.org/10.1155/2018/9702650.

[2] Mittal, N., Singh, U., & Sohi, B. S. (2017). A stable energy efficient clustering protocol for wireless sensor networks. Wireless Networks, 23(6), 1809-1821.

[3] Kong, P., Fang, G., He, C., & Liu, Z. (2017, July). Topology optimization of port wireless sensor network based on small-world network. In Circuits, System and Simulation (ICCSS), 2017 International Conference on (pp. 157-161). IEEE.

[4] Zhao, J., Yağan, O., & Gligor, V. (2014, September). Connectivity in secure wireless sensor networks under transmission constraints. In Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on (pp. 1294-1301). IEEE.

[5] Mohamed, S. M., Hamza, H. S., & Saroit, I. A. (2017). Coverage in mobile wireless sensor networks (M-WSN): A survey. Computer Communications, 110, 133-150.

[6] Naranjo, P. G. V., Shojafar, M., Mostafaei, H., Pooranian, Z., & Baccarelli, E. (2017). P-SEP: a prolong stable election routing algorithm for energy-limited heterogeneous fog-supported wireless sensor networks. The Journal of Supercomputing, 73(2), 733-755.

[7] Bhowmik, S., & Giri, C. (2013). A novel fuzzy sensing model for sensor nodes in wireless sensor network. In Intelligent Informatics(pp. 365-371). Springer, Berlin, Heidelberg.

[8] Heinzelman, W. R., Chandrakasan, A., & Balakrishnan, H. (2000, January). Energy-efficient communication protocol for wireless microsensor networks. In System sciences, 2000. Proceedings of the 33rd annual Hawaii international conference on (pp. 10-pp). IEEE.

[9] Zhang, H., & Hou, J. C. (2005). Maintaining sensing coverage and connectivity in large sensor networks. Ad Hoc & Sensor Wireless Networks, 1, 14-28.

[10] Ye, F., Zhong, G., Cheng, J., Lu, S., & Zhang, L. (2003,May). PEAS: A robust energy conserving protocol for long-lived sensor networks. In Distributed computing systems, 2003. Proceedings. 23rd international conference on (pp. 28-37).

[11] Cerpa, A., & Estrin, D. (2004). ASCENT: Adaptive self- configuring sensor networks topologies. IEEE transactions on mobile computing, 3(3), 272-285.

[12] Li, J., Andrew, L. L., Foh, C. H., Zukerman, M., & Chen, H. H. (2009). Connectivity, coverage and placement in wireless sensor networks. Sensors, 9(10), 7664-7693.

[13] Lazos, L., & Poovendran, R. (2006). Stochastic coverage in heterogeneous sensor networks. ACM Transactions on Sensor Networks (TOSN), 2(3), 325-358.

[14] Chang, C. Y., Chang, C. T., Chen, Y. C., & Chang, H. R. (2009). Obstacle-resistant deployment algorithms for wireless sensor networks. IEEE Transactions on Vehicular Technology, 58(6), 2925-2941.

[15] Falcon, R., Li, X., & Nayak, A. (2010, June). Carrier- based coverage augmentation in wireless sensor and robot networks. In Distributed Computing Systems Workshops (ICDCSW), 2010 IEEE 30th International Conference on (pp. 234-239). IEEE.

[16] Farahani, B. J., Ghaffarian, H., & Fathy, M. (2009). A fuzzy based priority approach in mobile sensor network coverage. International Journal of Recent Trends in Engineering, 2(1).

[17] Chang, C. Y., Yu, G. J., Wang, T. L., & Lin, C. Y. (2014). Path construction and visit scheduling for targets by using data mules. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44(10), 1289-1300.

[18] More, A., & Raisinghani, V. (2014). Random backoff sleep protocol for energy efficient coverage in wireless sensor networks. In Advanced Computing, Networking and Informatics-Volume 2(pp. 123-131). Springer, Cham.

[19] More, A., & Raisinghani, V. (2015). Discharge curve backoff sleep protocol for energy efficient coverage in wireless sensor networks. Procedia Computer Science, 57, 1131-1139.

[20] Idrees, A. K., Deschinkel, K., Salomon, M., & Couturier, R. (2015). Distributed lifetime coverage optimization protocol in wireless sensor networks. The Journal of Supercomputing, 71(12), 4578-4593.

Wireless Internet of Things-based Air Quality Device for Smart Pollution Monitoring

Nurul Azma Zakaria[1], Zaheera Zainal Abidin[2], Norharyati Harum[3]
Center for Advanced Computing Technology
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka, Malaysia

Low Chen Hau[4]
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka, Malaysia

Nabeel Salih Ali[5]
Information Technology Research and Development Center
University of Kufa
Najaf, Iraq

Fairul Azni Jafar[6]
Advanced Manufacturing Center
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka, Malaysia

***Abstract*—Living in a healthy environment is a need for every human being whether indoor or outdoor. However, pollutions occur everywhere and most people are merely mindful of the importance of having clean outdoor air to breathe and are not concerned about the indoor air quality. Indoor air quality refers to the quality within the building, and relates to the health and comfort of the building occupants. Dangerous particles exist in the outside air, pollute the indoor environment and produce harmful conditions as the polluted air travels into the house or building through windows or doors. Therefore, a wireless Internet of Things-based air quality device is developed to monitor the air quality in the indoor environment. The proposed system integrates a low-cost air quality sensor, temperature and humidity sensors, a single-board computer (Raspberry Pi 2 microprocessor) and cloud storage. The system provides real-time air quality reading, transfers the data through a wireless network to the Internet and displays the data in dedicated webpage. Furthermore, it stores records in cloud storage and sends e-mail notification message to the user when unhealthy condition is met. The study has a significant impact on promoting affordable and portable smart pollution monitoring system as the development of the device utilizing low-cost and off-the-shelf components.**

Keywords—Internet of Things (IoT); single-board computer; cloud storage; smart pollution monitoring

I. Introduction

Most people are concerned about their health, whether the food is safe to eat or the water is clean to drink. The most important thing in healthy life is the air quality which most people pay little attention to the environment. The air pollution is frequently happening in Malaysia due to many sources such as open-fire, gas combustion from vehicles, and factories waste. These events release harmful gases and one of the examples is nitrogen oxide (NO_X) which is the primary component of acid rain formation. Harmful gases affect our wellness and can induce lung cancer or chronic heart disease even though in indoor environments. This is because the polluted outdoor air come into the indoor environment through windows and doors. Thus, a system which detects the air pollution is necessary so that people are able to monitor the air quality and perform the necessary action to keep the harmful gases away.

At present the sensor technology is improving as compared to previously. A number of researchers work with sensors to monitor the environment and resolve environmental issues. For example, a work by Kaur et al. [1] recognized environmental parameters such as the temperature, relative humidity, carbon monoxide (CO), carbon dioxide (CO_2) and luminosity by using SHT10, MQ7, T6615, LDR respectively. Likewise, [2] used MQ-135 and DHT-11 sensors to detect the temperature, humidity, CO, CO_2, smoke and alcohol. The DHT-11 sensor consists of Negative Temperature Coefficient (NTC) component to measure temperature and resistive type component to measure humidity. Ibrahim et al. [3] proposed an environmental monitoring device using a single board computer which allows remote access and control of a platform with Python programming language. The system measures concentrations of carbon monoxide and detects earthquakes through an assembled seismic sensor.

Furthermore, Devarakonda [4] implemented a vehicular-based mobile access for measuring fine-grained air quality in real-time. However, this work used expensive equipment at specified locations or dedicated mobile equipment and laboratories to perform the measurement of the pollution and it has portability issue. Another work by [5] depended on the base station to function. The value from every sensor node is sent after a request is received from the base station. It uses ultra-low power sensor based microcontroller and suitable for business sector. The drawback of this system is every sensor node needs to be ready and power up at all times to achieve total measurement, although the data are only required for a short period. This incurs high usage of power and it is costly. An option of a low power consumption device will benefit the research. Moreover, [6] proposed a scheme that comprises of a base station associated with the network and a few self-governing hubs furnished with various sensors to measure temperature, stickiness, light and air quality. A system with LabVIEW is used in order to administer the operation and estimates the system. Nevertheless, this work focus on a huge and comprehensive system which is suitable for large scale arrangement.

There are many commercial air quality sensor products and monitoring systems available in the market but the price is expensive and not affordable to employ. They consist of many features, functions and high-end equipment with complex technologies. Special skills and knowledge are required for system's handling and at times it is hard to set up for those with no expertise. Besides, the physical size of the commercial air quality system is usually huge and heavy and some of the existing systems do not provide the notification alert and data storage for further analysis. Thus a device with small scale size is necessary to increase the portability aspects of the arrangement so that the device can be used anywhere and at any time. Hence, this study has been carried out with three (3) research questions: (1) What is the suitable low-cost component that can be used to monitor the air quality? (2) How can the user be alerted in specific condition? (3) What is the appropriate mechanism for data display and storage? These research questions are mapped with the research objectives of this study which are to develop a wireless and affordable Internet of Things (IoT)-based device that can monitor the air quality which can affect human's health, to integrate the monitoring system with a cloud storage so that the information can be viewed online on a web, and generate an alert notification e-mail when the air quality is in unhealthy condition. The proposed solution intends to solve the highlighted issues by using IoT technology, wireless communication and cloud services.

A number of low-cost components, off-the-shelf and open-source products are involved in the development of the proposed system. For example Raspberry Pi 2 microprocessor which is an integrated circuit that has a processing unit inside the on-board chip, air quality sensor to sense harmful gases, temperature sensor and humidity sensor. A web page is created on open-source cloud platform to display and analyse the air quality data from the sensors. Furthermore, a notification message is transmitted to the user through the e-mail service when the air quality value is beyond the threshold limit which indicates the indoor atmosphere is in an unhealthy state. The transfer of air quality data and e-mail notification message to cloud storage and user are by using wireless network connections. So, with the proposed system the air quality in the indoor environment can be easily monitored in real-time.

The remainder of this report is organized into four (4) parts. Section 2 presents the methodology employed in this work and Section 3 provides a detailed explanation of results and discussion. The final section concludes the paper.

II. METHODOLOGY

In general the implementation of the proposed system was guided by Rapid Application Development model (RAD) [7]. RAD is a type of incremental model as depicted in Fig. 1. The RAD methodology involves four (4) main phases which are requirements planning, user design, construction and testing and cutover. The details of each phases are described as follows.

A. Requirements Planning

The objective of this phase is a process to determine the needs of the study such as establishment of the problem statement, scope, objectives, data and functional requirements, and types of hardware and software used in the implementation.

1) Data Requirement

The input of the system is the air quality sensor MQ-135, temperature and humidity sensor DHT-22. The MQ-135 sensor is used to sense the air quality which comprises of chemical substances such as ammonia (NH_3), nitrogen oxide (NO_X), alcohol, benzene and other harmful gases. The concentration scope of the detection for benzene vapour and ammonia (NH_3) are 10ppm to 1000ppm and 10ppm to 300ppm respectively, whereas the DHT-22 is used to sense the temperature and humidity in the air in degree Celsius and percentage. The output of the system is a notification email alert and online data display and storage.

2) Software Requirement

The main programming language used was Python. It was for code development of the system. It was integrated with several modules to realize the system's functionalities. Other related software requirements were Hypertext Markup Language (HTML) and Hypertext Preprocessor (PHP) for web development that incorporated with open source ThingSpeak [8] cloud storage. ThingSpeak requires an Application Programming Interface (API) in order to provide web services for connecting things or objects such as storing and retrieving data from the sensors using Hypertext Transfer Protocol (HTTP) at port 80 over the Internet. As for the alert notification part, the system communicated with Google e-mail server for e-mail service of personal account.

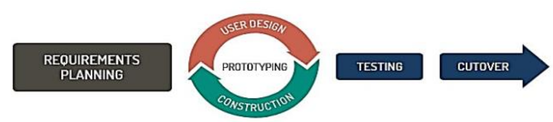


Fig. 1. RAD Methodology Phase [7].

TABLE I. SPECIFICATION OF RASPBERRY PI 2 MODEL1

Model	Raspberry Pi 2 Model B
Price	USD 30
Processor type	Broadcom BCM2836 ARMv7 SoC
Processor speed	Quad-core, 900MHz
Memory speed	1GB, 450MHz
Storage	microSD
Ethernet port	1 x 10/100
USB ports	4 x USB 2.0
GPIO	40 pin
Video	HDMI, Composite RCA (shared)
Audio	Multi-channel HD via HDMI, Stereo from 3.5mm jack (shared)
Power	5v microUSB, 800mA
Size	85 x 56 x 21mm

[1] Element14. Raspberry Pi models comparison, https://www.element14.com/community/servlet/JiveServlet/previewBody/82195-102-3-346675/PiPoster_14Jun16.pdf (accessed August 31, 2016).

3) Hardware Requirement

The core development of this IoT device comprised of a credit-card sized microprocessor computer i.e. Raspberry Pi 2 [9] that plugs into the monitor and a keyboard. It is the processing device used in this system. The specification of Raspberry Pi 2 is listed in Table 1.

Another key component is the sensor. A sensor is a device that can detect and sense some specific input from the physical environment such as light, heat, motion, moisture, and pressure. The output of the sensor is usually converted the signal to human-readable display at the sensor location or transmitted electronically using network environment for viewing or further processing. In this case, low cost MQ-135 air quality gas sensor and DHT-22 temperature and humidity sensor were used in the implementation with additional hardware such as Wi-Fi module for wireless communication, LEDs for indicator, MCP3008 component for analogue to digital converter, breadboard, and jumper wires.

B. User Design

In user design phase, model and prototype that represent all system processes, inputs, and outputs are developed. User design is a continuous interactive process that allows users to understand and modify a working model of the system that meets their needs. In this study, the air quality monitoring system model and prototype were developed based on system operations, inputs and outputs. The design for the proposed system is illustrated in Fig. 2. It demonstrates the integration of the sensors, LEDs, Internet and cloud storage with the processing platform i.e. Raspberry Pi 2 microprocessor.

C. Construction

The objective of the construction phase is to carry out the detailed plan of the proposed scheme. The design of the proposed system is initially described in the user design phase and completed in construction phase. In this research, the construction was divided into three (3) parts. The first part was to configure all hardware and program the Raspberry Pi 2 board based on the design to ensure the system works properly. Next, the cloud storage was configured so that the data acquisition can be stored in the cloud. The last part was the notification part; where a client and an e-mail server were configured in this system and alert notification message was sent to the user. The process cycles between the user design phase and construction until final prototype is created. Fig. 3 illustrates the flow of the proposed system, whereas Fig. 4 demonstrates the example of hardware connectivity, in this case is MQ-135 gas sensor to the Raspberry Pi board.

D. Testing and Cutover

In testing and cutover phase, the system combines the software and hardware components in order to acquire a complete system and test the system's functionalities. The hardware prototype and the web page developed in ThingSpeak cloud storage were tested together to ensure all sensors, e-mail and LEDs work properly based on the setup function. In this proposed system, the level of the air pollution was set in the construction step accordingly as indicated in Table 2. It is a simplified setting of Air Quality Index (AQI) level [2]. However, this setting can be easily modified by changing the configuration program to cater future requirement. When the device detected the level of the health concern was at good and moderate level, the green-coloured LED lighted up, whereas when the health concern was in unhealthy level, the red-coloured LED turned on and a notification message was sent to the user via an e-mail.

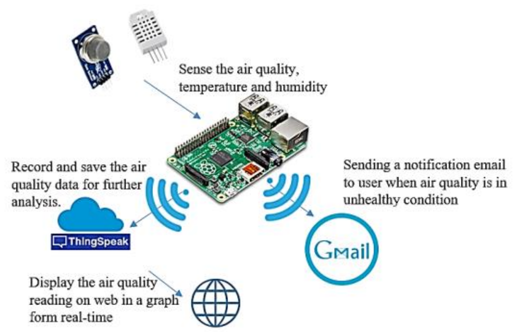


Fig. 2. System design of the proposed solution.

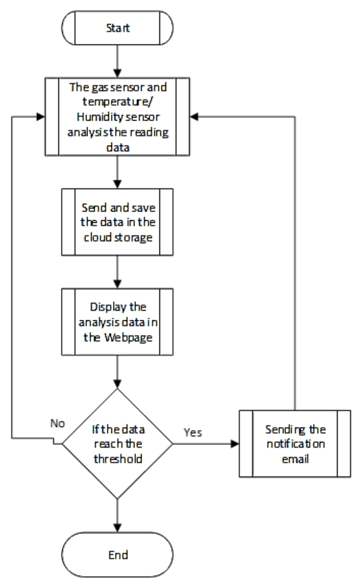


Fig. 3. Flow chart of the proposed system.

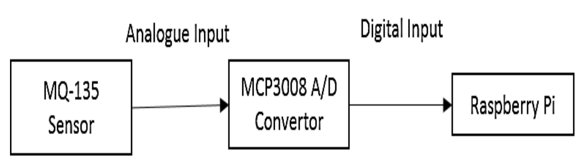


Fig. 4. Connectivity of MQ-135 gas sensor to Raspberry Pi.

TABLE II. AIR QUALITY INDEX (AQI) LEVEL

Air Quality Index	Level of Health Concern
0 to 100	Good
101 to 150	Moderate
151 to 600	Unhealthy/Hazardous

[2] U.S. Environmental Protection Agency. Air Quality Index (AQI) a guide to air quality and your health, https://www3.epa.gov/airnow/aqi_brochure_02_14.pdf (accessed 31 July, 2016)

III. Results and Discussion

In this study, three (3) experiments were conducted, namely connectivity test of sensors with Raspberry Pi, generation of the alert notification message by e-mail service test and integration test of cloud storage.

In the first experiment, the connectivity of MQ-135 and DHT-22 sensors with the Raspberry Pi board as depicted in Fig. 5 were tested. The connections of the sensors to the Raspberry Pi board were by using the general purpose input/output (GPIO) pins. Initially, the Raspberry Pi was switched on and Python script was run in the terminal by executing a command *sudo python analog.py*. This calibrated the air quality data. Then, the Raspberry Pi read the air quality data from the air sensor and displayed the data in the terminal of the Raspberry Pi. The DHT-22 sensor sensed the temperature and humidity reading whereas the MQ-135 gas sensor detected the air quality value. Fig. 6 illustrates the results of the air quality level which is in healthy condition with 83 ppm and temperature and humidity values are 32.7 degree Celsius and 67.4% respectively.

The second experiment was to send e-mail alert to user when the air quality value reaches the threshold limit based on the values listed in Table 2. The threshold value was configured in the main program (*analog.py*). The same command was executed to initiate the system and calibrate the readings from the sensors. E-mail notification was generated when the set threshold had been met. Fig. 7 presents the data retrieved from the sensors. When the air quality value was beyond the limit of the threshold, the air quality condition was stated as unhealthy/hazardous in the terminal and an e-mail alert was sent to the user. Fig. 8 represents the sample of the e-mail notification that is received by the user when the air quality level is at unhealthy or hazardous condition. With this warning notification, user aware of the current condition of the indoor environment and further action to amend the situation can be considered.

The third experiment was to examine the functionality of web-based cloud storage for record keeping and data presentation. All parts such as sensors, Raspberry Pi and ThingSpeak web-based cloud storage were set up accordingly. The command *sudo python analog.py* was executed to start the system. ThingSpeak cloud storage was used and proper login was required. Then, channel type Private was chosen to record the current reading of data from the system in graphical form. In addition, Data Import/Export option was selected for data saving. In this system the data were displayed in an online chart and saved on the cloud storage for further analysis.

Fig. 9 shows the web-based cloud storage platform "ThingSpeak.com". An account and channels were created in the implementation phase. The channel ID and API key were added into the main script (*analog.py*) to enable the reading process from the sensors and data saving in the cloud storage. Fig. 10 illustrates the data reading retrieved from the sensors which comprises of temperature, humidity and air quality values. The data from the sensors were sent to the cloud storage and displayed in graphical form. It shows the cloud storage receives the data successfully. Moreover, Fig. 11 illustrates the *.csv* file that was exported from the cloud storage. The information is downloadable by the system's user as reference report for further analysis of the air quality.

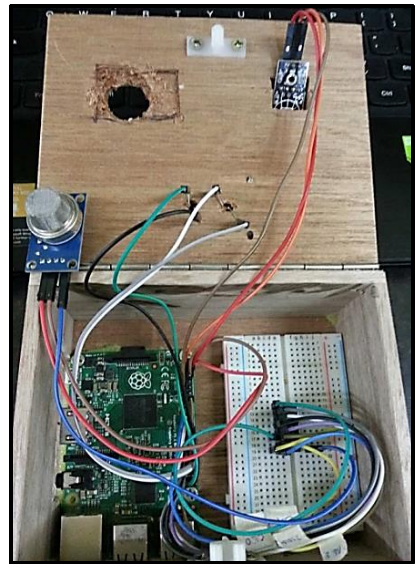


Fig. 5. Connectivity of sensors and the Raspberry Pi.

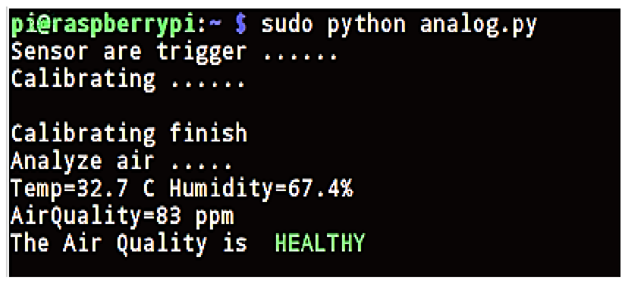


Fig. 6. Sensor shows the reading after calibrated.

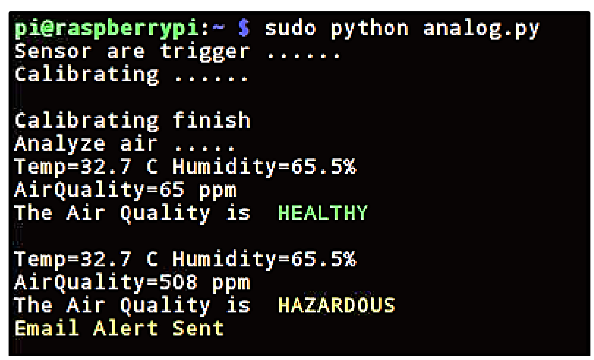


Fig. 7. Warning notification using e-mail alert.

Fig. 8. E-mail notification received by the user.

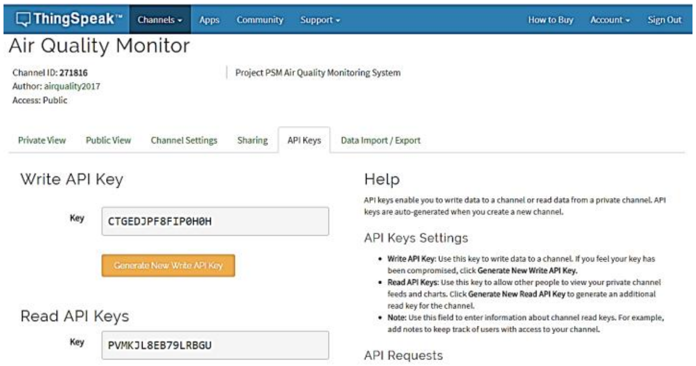


Fig. 9. Web-based cloud storage.

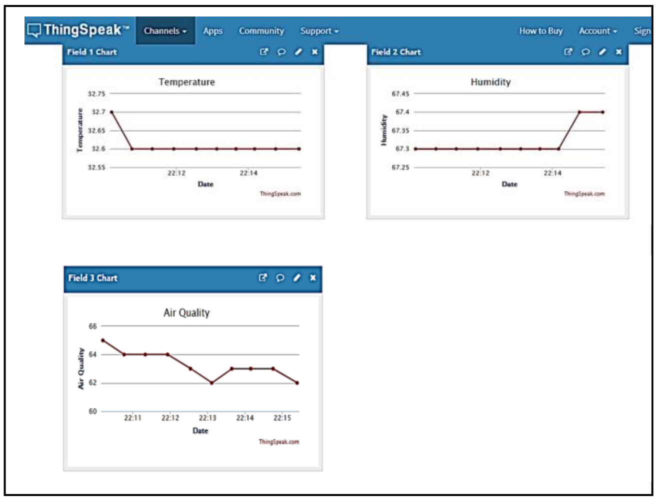


Fig. 10. Graphical data representation.

Date	Entry ID	Temp	Hum	AQ
2017-07-18 14:03:01 UTC	324	32.7	67.3	66
2017-07-18 14:03:48 UTC	325	32.7	67.3	66
2017-07-18 14:04:25 UTC	326	32.7	67.3	66
2017-07-18 14:04:57 UTC	327	32.7	67.3	65
2017-07-18 14:05:34 UTC	328	32.7	67.3	65
2017-07-18 14:06:05 UTC	329	32.7	67.3	65
2017-07-18 14:06:42 UTC	330	32.7	67.3	65
2017-07-18 14:07:14 UTC	331	32.7	67.3	65
2017-07-18 14:07:50 UTC	332	32.7	67.3	65
2017-07-18 14:08:32 UTC	333	32.7	67.3	65
2017-07-18 14:09:04 UTC	334	32.7	67.3	65
2017-07-18 14:09:35 UTC	335	32.7	67.1	65
2017-07-18 14:10:12 UTC	336	32.7	67.3	65
2017-07-18 14:10:46 UTC	337	32.6	67.3	64
2017-07-18 14:11:20 UTC	338	32.6	67.3	64
2017-07-18 14:11:56 UTC	339	32.6	67.3	64
2017-07-18 14:12:33 UTC	340	32.6	67.3	63
2017-07-18 14:13:07 UTC	341	32.6	67.3	62
2017-07-18 14:13:39 UTC	342	32.6	67.3	63
2017-07-18 14:14:10 UTC	343	32.6	67.3	63
2017-07-18 14:14:45 UTC	344	32.6	67.4	63

Fig. 11. Data export (.csv file) from cloud storage.

IV. Conclusion

In this paper, we have shown the development of the proposed solution which integrates Raspberry Pi microprocessor, air quality, temperature sensor and humidity sensor to form an Internet of Things (IoT) -based air quality monitoring system. The proposed system exhibits several strengths such as providing a real-time web-based cloud application that monitor, update, and display air quality data online, and sends an e–mail alert notification to the user when air quality is in unhealthy condition that is when the set threshold has been met. Further analysis of the air quality status can be carried out by the user, such as home owners, school operators, company managers, and manufacturing plant directors to ensure indoor conditions of living or working environment of the building is in healthy level for all occupants. This smart pollution monitoring system is easy to install and maintain, as it requires very low technical skills and knowledge for device handling and system operation which will benefit various users. However, most of the real-time functionality of this system is only limited to the area that have wireless network access. Thus, further research in the network communication area should be explored to enhance the usefulness of the system in a wider application area towards achieving conducive and healthy surroundings.

Acknowledgment

A high appreciation to Center of Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, and Universiti Teknikal Malaysia Melaka (UTeM) for supporting the work done in this paper.

References

[1] N. Kaur, R. Mahajan, D. Bagai and P.G. Student, “Air quality monitoring system based on Arduino microcontroller,” International Journal Innovative Research in Science, Engineering and Technology, vol. 3, pp. 9635-9646, 2016.

[2] G. Marques and R. Pitarma, “An indoor monitoring system for ambient assisted living based on internet of things architecture,” International Journal of Environmental Research and Public Health, vol. 13, no. 11, pp. 1152-1162, 2016

[3] M. Ibrahim, A. Elgamri, S. Babiker and A. Mohamed, “Internet of things based smart environmental monitoring using the Raspberry-Pi computer,” in 2015 Fifth International Conference on Digital Information Processing and Communications, Sierre, 2015, pp. 159-164.

[4] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode and B. Nath, “Real-time air quality monitoring through mobile sensing in metropolitan areas,” in 2nd ACM SIGKDD International Workshop on Urban Computing, Chicago, Illinois, 2013. doi: [10.1145/2505821.2505834]

[5] D.M.G. Preethichandra. “Design of a smart indoor air quality monitoring wireless sensor network for assisted living,” in International Instrumentation and Measurement Technology Conference, Minneapolis, 2013, pp.1306–1310.

[6] J. Lozano, J.I. Suarez, P. Arroyo, J.M. Ordiales and F. Alvarez, “Wireless sensor network for indoor air quality monitoring,” Chemical EngineeringTransactions, vol. 30, pp. 319–324, 2012.

[7] TatvaSoft. (2015). Top 12 software development methodologies and its advantages/disadvantages, [Online] Available: https://www.tatvasoft.com/blog/top-12-software-development-methodologies-and-its-advantages-disadvantages/

[8] S. Pasha, “Thingspeak based sensing and monitoring system for IoT with MATLAB analysis,” International Journal of New Technology and Research, vol. 2, no. 6, pp. 19-23, 2016.

[9] C. Severance, “Eben Upton: Raspberry Pi,” Computer, vol. 10, pp. 14-16, 2013.

Securing Locations of Mobile Nodes in Wireless Mesh Networks

Sultan Alkhliwi
Lecturer at Faculty of Science,
Northern Border University, KSA

***Abstract*—The current deployment of wireless mesh networks requires mobility management to track the current locations of mobile nodes around the network without service interruption. To do so, the Hierarchical Mobile IPv6 protocol has been chosen, which minimises the required signalling by introducing a new entity called the mobile anchor point to act as a local home agent for all visiting mobile nodes in a specific domain. It allows a mobile node to register its local/regional care-of addresses with a mobile anchor point by sending a local binding update message. However, the local binding update is quite sensitive; it modifies the routing to enable mobility in the wireless mesh networks. When a local binding update message is spoofed, an attacker can redirect traffic that is destined for legitimate mobile node either to itself or to another node. This situation leads to an increased risk of attacks. Therefore, this paper contributes to addressing this security issue based on wireless mesh networks by cryptography generation and verification of a mobile node's local and regional care-of addresses, as well as the application of a novel method to verify the reachability of mobile node at claimed local care-of address. This is called the enhanced mobile anchor point registration protocol. The Scyther tool has been used to ensure the proposed protocol accuracy. Furthermore, the performance, in terms of the mobile anchor point registration delay and signalling overhead, is evaluated by using the OPNET modeller simulator.**

Keywords*—*Wireless mesh networks; hierarchical mobile IPv6 protocol; authentication; secret key; Scyther tool; OPNET simulation

I. Introduction

The number of wireless devices, such as laptops, PDAs, Bluetooth devices and so on, is substantially increasing all over the world. As a result, the demand for broadband wireless access is increasing rapidly. Currently, it seems that wireless mesh networks (WMNs) [1] will play a major role in future anywhere–anytime communications. The WMNs have gained significant attention from the research community, as well as the industry and standard organisations, due to their wireless access flexibility, combined with their high coverage area, excellent reliability and proven cost efficiency. The WMNs have a wide range of applications, such as in-home broadband services, enterprises, communities, metropolitan areas, intelligent transportation, industrial automation, sensors and emergency/rescue networks. The WMNs have received considerable interest as a promising way for reliable wireless broadband services to gain access with minimal upfront investments. The WMN features (i.e., dynamic, self-organised and self-healing) can be deployed incrementally, one node at a time, as needed. As more nodes are installed, both reliability and connectivity will increase, which all users will enjoy[1],[2].

A WMN comprises dedicated backbone wireless access routers (ARs) and gateway routers to offer last-mile broadband connectivity to users. Since roaming is related to the desire to access the internet from a WMN, an efficient mobility management protocol is required [1]. To facilitate WMN abilities to locate a mobile node (MN) point of attachment for delivering data packets and to maintain an MN's connection as it continues to change its point of attachment in the domain, the Engineering Task Force (IETF) proposed the Mobile IPv6 (MIPv6) protocol [3] to both permit this roaming connectivity and reduce the required signalling with the Hierarchical Mobile IPv6 (HMIPv6) protocol [4]. Therefore, the HMIPv6 protocol has been chosen to support mobility location management for WMNs. The HMIPv6 is a lightweight protocol (compared with others) for the following reasons: (1) It supports hierarchical tunnelling approaches that are flexible, modular and scalable for supporting IP-based macro and micro mobility for WMNs. (2) It introduces a new node agent called the mobile anchor point (MAP) to act as a local home agent for the MNs. Thus, when an MN changes its attached point in the same domain, it will update its location with the MAP instead of the home link, as defined in the MIPv6 protocol [3]. As a result, the HMIPv6 protocol requires minimal bandwidth and computational resources, as well as reduces registration delays when tracking the current location of the MNs. Finally, the HMIPv6 protocol enables the deployment of a group of ARs into different subnets for easy employment of WMNs.

In the HMIPv6 protocol, an MN has three IPv6 addresses, including a permanent home-of-address, which identifies the MN in its home link and remains the same during the MN's movement, and two transients: the regional care-of address (RCoA), which is generated based on the MAP option that is included in the router advertisement (RA) message, and the local care-of address (LCoA), which is generated based on an AR advertisement. When an MN moves into a MAP domain and configures a new LCoA and RCoA, it initiates a process to register both its LCoA and RCoA with the MAP by sending a local binding update (LBU). The MAP then replies to the MN with a binding acknowledgement (BA) message. This registration allows the MAP to create a binding for the MN between its LCoA and RCoA. The MAP uses this binding to intercept all packets that are destined for the MN's RCoA from its home link and/or correspondent nodes and forwards these packets to the MN's current location in the MAP domain by using the MN's LCoA.

However, the LBU is obviously quite sensitive; it modifies the routing to enable mobility in the WMNs. When an LBU message is spoofed, an attacker can redirect traffic either to itself or to another node, preventing the original MN from receiving any traffic that is destined for it. This situation leads to an increased risk of attacks (e.g., denial-of-service attack, man-in-the-middle attack). Thus, the greatest security vulnerabilities of the HMIPv6 protocol are both the authentication and the authorisation of an LBU message. Therefore, the use of appropriate security provisions for the MAP registration process is fundamental to the HMIPv6 protocol. It is believed that the deployment of an HMIPv6 protocol without securing the MAP registration process could result in a breakdown of the entire internet [5][6].

This paper presents a novel scheme, called the Enhanced Mobile Anchor Point (E-MAP) registration protocol, to support the location authentication of MNs in the MAP in the WMN domain and to authorise the MN to use the services of the WMN domain. By executing the E-MAP protocol, the MAP is able to verify the ownership of the claimed LCoA and RCoA and confirm not only the authenticity of the LCoA but that it is indeed an MN's real location. The E-MAP registration protocol also allows the MAP to securely identify and establish a shared secret key with the MN. As a result, the E-MAP protocol can reduce the likelihood that a malicious MN can successfully steal a third party's node addresses (i.e., LCoA and RCoA), prevent a malicious MN from launching a flooding attack and protect any future binding update (BU) that could be sent from the MN to the MAP against the false BU attack.

The rest of this paper is organised as follows. Section II provides an overview of the cryptographically generated address (CGA) technique, discusses the proposed protocols to secure the MAP registration process and related works of the reachability test mechanism. Section III presents the preliminaries behind the design of the E-MAP registration protocol. Section IV covers an overview of this study's novel protocol, including the idea behind the public key certificate CGA-based technique and the idea of simultaneously conducting the LCoA reachability test and generating a shared secret key between the MAP and the MN. Section V describes the E-MAP protocol in detail. Section VI presents a formal security analysis of the proposed protocol using the Scyther tool. Section VII evaluates the performance of the E-MAP protocol compared with both the basic MAP (B-MAP) registration protocol and the most related work in terms of the MAP registration delay and signalling overhead, using the OPNET modeller simulation. Section VIII concludes this paper and suggests future work.

II. Related Works

This section provides an overview of cryptographically generated IPv6 addresses, surveying the existing protocols, which are used to secure the MAP registration process in the HMIPv6 protocol, and related works of the reachability test mechanism.

A. Cryptographically Generated Address (CGA) Protocol

The CGA technique is used to prevent the stealing and spoofing of existing IPv6 addresses [7]. A CGA is an IPv6 address for which the interface identifier part is generated, using a cryptographic one-way hash function that takes the address owner's public key and some auxiliary parameters as its input. The address owner can protect a message sent from the address by attaching its public key and auxiliary parameters to the message and signing it with the corresponding private key [8]. Thus, the owner asserts its ownership of the address by using the corresponding private key. Upon receipt of the signed message, the intended recipient verifies the binding between the public key and the address by recomputing and comparing the hash value with the interface identifier part of the address. Additionally, the recipient authenticates the address by verifying the signature. However, the CGA-based technique suffers from several limitations. First, it relies on the digital signature that is added to each message sent, but the IP header (the source address) is excluded from the signature. Therefore, an attacker could easily find and store its victim's messages while obtaining the victim's modifier and public key. Second, as a standalone solution, the CGA-based technique does not guarantee the owner's reachability at the claim address; an attacker can easily use its own public key to cryptographically generate a non-used address with a subnet prefix from the victim's network. Third, despite the CGA-based technique's ability to effectively prevent attackers from impersonating valid IPv6 addresses to launch attacks, it cannot prevent attacks on a network that involve redirecting data to a non-used address. Fourth, the CGA-based technique requires heavy computations to calculate and verify the digital signature, which could expose the network entities to denial-of-service attacks, particularly when the entity is an MN and has limited computation power or when it needs to verify digital signatures for many peers at the same time. Fifth, the MN can self-generate a public–private key pair that is not certified by a trusted third party. In this case, the malicious node can easily enter the network and use these keys to assign its care-of addresses from a specific domain, then access the network resources illegally. Finally, as the address owner's public key and digital signature and the auxiliary parameter values are carried in the message delivery procedure to generate the address cryptographically, a certain amount of overhead is incurred due to bandwidth consumption.

B. Existing Protocols to Secure Mobile Anchor Point (MAP) Registration in HMIPv6

The specification of the HMIPv6 protocol [4] suggests the use of the Internet Key Exchange version 2 (IKEv2) protocol [9] to establish a security association between the MN and the MAP, along with the IPSec protocol[10],[11], to protect the LBU and the BA messages. However, the IKEv2 protocol has limitations regarding communication and computational overhead, making it inefficient due to cryptographic operations and the need for four to six messages with two to three round trips to create a security association between the MN and the MAP. Additionally, the IKEv2-based key setup is difficult to achieve in a multi-hop communication environment with dynamic connections, such as the WMN [12]. Conversely, by using the IPSec protocol, the MAP registration protects against outside attacks (i.e., an attacker cannot send a spoofed LBU message instead of the MN). However, the IPSec protocol can authenticate neither the claimed LCoA nor the RCoA, and it cannot prevent the legitimate MN from sending a fake LCoA, which will cause the MAP to redirect traffic to the victim's location. Additionally, it does not provide a mechanism for the MAP to verify the ownership of the MN's RCoA, through which the MN can use the services in the MAP's domain. It also cannot prevent the attacker from replying to the LBU message that the MN sent earlier to the MAP. As a result, the MAP will redirect all subsequent traffic to the MN's old location. This situation can cause a denial-of-service attack to both the MN and the node that is currently located in the AR.

Kang-Park's security protocol [13] aims to secure the LBU and the fast handover in the HMIPv6. To protect the MAP registration process, the protocol leverages the authentication, authorisation and accounting (AAA) infrastructure [14], through which the MAP issues the authentication ticket to the MN. Two types of AAA servers are employed: one is operated by a home service provider (AAAH), and the other is operated by a service provider in a foreign network (AAAF). In this protocol, the MAP partially protects against resource exhaustion during denial-of-service attacks because the attacker cannot know the session key that is used to secure the LBU message. Moreover, it protects the entities against any replay attack, as the timestamps have been used. However, using the AAA infrastructure causes a long authentication and registration delay. Specifically, when the MAP receives the LBU message from the MN, it cannot authenticate the MN directly and asks the AAAH through the AAAF to generate and send the session key. If the distance between the AAAF and the AAAH is too long and frequent handovers occur, then large delays occur. Note that the MN cannot send and receive data sent from the correspondent node via the MAP until the MAP registration process is complete. This situation causes critical problems in the mobile network (e.g. registration delay). In contrast, Kang-Park's security protocol cannot verify the reachability of an MN at the claimed LCoA and the authenticity of the RCoA; thus, it is vulnerable to malicious flooding attacks from MNs and allows the visiting MNs to use the MAP domain resources illegally. Furthermore, it forces the MN to perform heavy computations, both to generate the session key to secure the LBU message and to verify the signature, which could expose the MN to a denial-of-service attack when the MN has limited computational power.

The ESS-FH protocol [15] is proposed to enhance Kang-Park's protocol [13] by combining the CGA technique [8] and the public-key cryptography operation. The protocol provides a strong key exchange and key independence based on both the public key encryption and the CGA technique. Additionally, it requests the MN to sign the LBU update message with its private key; thus, the MAP protects against redirect attacks. It also allows the MAP to verify the reachability of the MN at the claimed LCoA and the authenticity of the RCoA; thus, it prevents malicious MNs from launching flooding attacks against the MAP and prevents the MNs from using the MAP domain resources illegally. However, the ESS-FH protocol requires the MN to perform four public key operations to complete the registration with the MAP, which could expose the MN to denial-of-service attacks when the MN has limited computational power. Moreover, the protocol requires verification of the CGA-based technique [8] and the entity signature in each message; thus, the MAP registration delay, packet loss and signalling overhead could be increased. The protocol also requires the MN to self-generate its public–private key pair (unauthenticated key pair), and the RCoA is generated based on the CGA technique; thus, the malicious node can easily enter the network and use these keys to assign its RCoA from a specific domain, accessing the network resources illegally.

The HMIPv6sec protocol [16] aims to create an SA between the MN and the MAP based on the CGA technique [8]. It employs the following requirements: (1) The MN has a self-generated public–private key pair. (2) The LCoA of the MN is generated by using the CGA-based technique (3) The RCoA of the MN is generated by using the secret key shared between the MN and the AR, and the MAP's prefix is advertised by the MAP through the AR. (4) There are existing secure links between all the ARs located within the MAP tree. (5) The MN and the MAP use the Diffie-Hellman key exchange protocol to compute the secret shared key. The HMIPv6sec protocol requires the MN to sign the LBU message using the secret key shared with the AR, which partially protects the MAP against a denial-of-service attack. Furthermore, it increases security by ensuring the LCoA and the RCoA ownership, protecting the MAP against return-to-home spoofing attacks and preventing MNs from using the MAP domain resources illegally. However, this protocol cannot guarantee the reachability of the MN at the claimed LCoA; thus, it cannot protect third parties against denial-of-service attacks.

C. *Reachability Test*

A reachability test mechanism provides assurance that the MN is indeed located the claimed care-of address [17]. This section will examine various protocols have been adopted reachability test in their proposed.

Mobile IPv6 protocol used Return Routability procedure [3] to assist the correspondent node to assure that the MN can receive messages sent to the claimed home-of address and care-of address. The Return Routability procedure performs two reachability tests: a home-of address test and care-of address test. The MN performs the both tests simultaneously by sending a Home Test Init message its home agent and a Care-of Test Init message directly to the correspondent node. When the correspondent node receives the Home Test Init message, it sends a Home-of Test message to MN via home agent, including a secret home keygen token. Additionally, the correspondent node sends a Care-of Test message directly to MN in response to the Care-of Test Init message contains a secret care-of keygen token. However, the Return Routability procedure fails to provide sufficient protection for the correspondent registration. The attacker can sniff only the Home-of Test message from the correspondent node to forge and initiate a Care-of Test Init message by using its care-of address instead of the legitimate care-of address of the MN. Thus, the correspondent node replies to the Care-of Test message that the attacker uses in integration with the intercepted Home-of Test message to compute a binding management key. The attacker then impersonates the legitimate MN and sends a fake BU message to the correspondent node. As a result, all traffic redirects to the attacker instead of the MN.

The early binding update (EBU) protocol [18] is improved the Return Routability protocol [3] by shifting the home-of-address and care-of address reachability test to the handover phase where they cannot impact the registration delay. The home-of-address test is executed prior to the handover (i.e. the MN still in the old care-of address). After the handover, the care-of address test runs in parallel with data transfer to and from the new care-of address. However, the EBU protocol pays the cost for this reduction delay as the MN needs to run the home-of-address test periodically, that could increase the signalling overhead. Furthermore, it suffers from the same on-path attacks applicable to the Return Routability protocol.

Applying CGAs technique to optimise mobile IPv6 (CGA-OMIPv6) protocol [19] suggested to authenticate MN's home-of-address by using the CGA technique [8] together with an exchanging Home Test Init message and Home Test messages with the correspondent node to proof the reachability of MN at the home-of-address. Thus, the protocol partially protects against return-to-home spoofing attacks. On the other hand, to proof the reachability of MN at the care-of address, the MN exchanges Care-of Test Init message and Care-of Test message with the correspondent node, which partially protects correspondent node against denial-of-service attacks. However, the protocol required the MN and the correspondent node to perform two public key operations during the correspondent registration process, which is increased the complexity imposed on the both entities and led to increase the registration delay.

A novel scheme for supporting location authentication of mobile nodes [20] has been proposed to enhance the basic operation of the home registration defined in [3]. In this scheme, the authors suggested to add two extra mobility-related messages to allow the home agent to confirm the reachability of MN in the claimed care-of address. Both these messages are authenticated using the secret key shared between the MN and home agent. By using this method of reachability test, the scheme prevents the MN from lunching flooding attack. However, this method of reachability test suffers the following limitation. (1) Two additional messages are required to verify the reachability of the MN at claimed care-of address which could expose the MN to denial-of-service attack, as the MN in its nature is mobile entity and has limited computational power. (2) The MN wish to receive the packet from its correspondent node as soon as handoff to foreign network, the proposed reachability test introduced longer delay compared with basic operation defined in the MIPv6 to complete the home agent registration.

III. Preliminaries

Before presenting the E-MAP registration protocol, this section provides details of the assumptions and notations that the following sections use.

A. Assumptions

- Each AR with the MAP has a preconfigured security association for encrypting and authenticating communication exchanges. This assumption is justified by the fact that the HMIPv6 protocol and this study's contribution require ARs within the MAP tree to be involved in delivering the control mobility exchange messages and other packets sent by the MAP to MNs. Therefore, the IPsec Encapsulating Security Payload (ESP) protocol is used, and the AR and the MAP share the secret key ($K_{AR\text{-}MAP}$) [4], [16].
- A MAP has a public–private key pair (PK_{MAP}, SK_{MAP}). The private key SK_{MAP} is kept by the MAP and obtains a public key certificate ($Cert_{MAP}$) from a CA. The MAP possesses the key pair before the invocation of the protocol. The MAP then distributes its $Cert_{MAP}$ among the ARs that are located under the same coverage in the WMN domain. When the ARs receive the MAP's public key certificate, they check its validity with the Certificate Authority before the protocol is invoked.
- An MN and its attached AR have a preconfigured security association for encryption and authenticated communication after the MN completes a link layer handoff. They use the IPsec ESP protocol to protect mobility-related message exchanges. Thus, the MN and the AR share a secret key ($K_{MN\text{-}AR}$).

B. Notations

Table 1 lists the notations to be used in the E-MAP registration protocol description.

TABLE I. List of Notations

Notation	*Indication*		*Notation*	*Indication*
LCoA	Lcaol care-of address		Seq_X	A sequence number of node X
RCoA	Regional care-of address		Tx	A timestamp of node X
CoT	Care-of keygen token value		Ack	An acknowledgement value sent by MN to MAP
Modifier	A 128-bit value		MACx	A keyed hash value used to ensure the integrity and authenticity of the message
N_X	A nonce of node X		Kx-y	A shared secret key between two entities
ENG_{KBM}	An encyption value using K_{BM}		K_{BM}	A binding key management
IID	Interface identifire		||	Concatenation

IV. E-MAP Registration Protocol

The E-MAP registration protocol is designed based on two combined ideas. First, the E-MAP registration uses a novel, lightweight, improved version of the traditional CGA-based technique [8] to cryptographically generate and verify the MN's LCoA and RCoA. This is called the public key certificate CGA-based technique. Second, the E-MAP registration protocol uses the light-weight LCoA reachability test method to allow the MAP to confirm reachability of the MN at a claimed LCoA. In addition, the E-MAP registration protocol allows the MAP to securely identify and establish a shared secret key with the MN protect any future mobility messages that could be sent from the MN to the MAP against the possibility of a false mobility messages attack.

A. Public Key Certificate CGA-based Technique

The first idea of this study's proposed protocol aims to reduce the likelihood of a malicious MN stealing other nodes' addresses (i.e., LCoA and RCoA). It uses an improved version of the CGA-based technique [8], that is, a public key certificate CGA-based technique for cryptographic generation and verification of IPv6 addresses (i.e., LCoA and RCoA). The public key certificate CGA-based technique requires that a $Cert_{MAP}$ be distributed among the ARs in the WMN's domain by the MAP before the invocation of the E-MAP registration protocol. This step allows the ARs to cryptographically generate the MN's LCoA and RCoA on behalf of visiting MNs, and the MAP uses the same public key certificate to cryptographically verify the ownership of those addresses.

When the AR generates a CGA-based LCoA, it uses two input values: (1) a 64-bit subnet prefix of the AR and (2) the MAP's public key. The AR then uses the result of the MN's LCoA to compute the MN's RCoA by inputting two values: (1) the leftmost 64-bit of the output interface identifier (*IID*), which is computed based on the LCoA, and (2) the 64-bit prefix of the MAP, which is included in the RA message that is generated by the MAP. The AR runs the public key certificate CGA-based generation to compute the LCoA and the RCoA for the MN as soon as it receives the secure request message from the MN. The details are presented below and shown in Fig.1.

a) Generate a 128-bit random number called a modifier that is used to further randomise the LCoA generated from the same subnet prefix and the MAP's public key.

b) Concatenate from the left modifier and the AR subnet prefix. The AR then executes the HMAC_SHA1 function on the concatenation, using the MAP's public key (PKMAP), and obtains the leftmost 64 bits of the output. The result is HG: HG = First (64, HMAC SHA1 (PKMAP, (modifier || AR-subnet prefix))).

c) Form an IID from HG by setting both the U/L and the I/G bits to zero.

d) Concatenate the AR subnet prefix-(64 bits) and the IID-(64 bits) to form an IPv6 address-(128 bits) with the subnet prefix to the left and the IID to the right.

e) Perform a duplicate address detection [21] test at the LCoA. If an address collision is detected, increment the modifier by one and return to step *b*.

The outputs of the above steps are (1) a new CGA-based IPv6 address (an LCoA) and (2) the final value of the modifier-(128 bits). The AR then computes the RCoA as follows:

a) Concatenate the cryptographic generation of the LCoA with the MAP's public key (PK_{MAP}), execute the HMAC_SHA1 function to compute the *IID* that was used to compute the RCoA for the MN, and obtain the leftmost 64 bits of the output (*IID*). The result is (*IID*): *IID = First (64, HMAC SHA1 (*PK_{MAP}*|| LCoA)).*

b) Concatenate the MAP-subnet prefix-(64 bits) with 64 bits of the *IID* to form the 128-bit IPv6 address with the MAP prefix to the left and the *IID* to the right. The result is RCoA: *RCoA = (MAP-subnet prefix (64 bit) || IID" (64 bit)).*

The output of the above steps is a new IPv6 address (i.e., an RCoA). In this case, the LCoA and the RCoA address generation is completed. In parallel, the AR then securely sends a pre-binding update (PBU) message to the MAP and a reply request message to the MN. The PBU message includes the MN's LCoA and RCoA. When the MAP receives the PBU message, it creates a binding cache entry for the MN and stores the values of the LCoA and the RCoA. Once the cache entry is created, the MAP waits for a limited amount of time for the owner of those addresses to send the BU message. If no valid BU message is received during the binding cache entry's preconfigured lifetime, the MAP will delete this cache entry. When the MN receives the reply request message from the AR, it securely sends the LBU message to the MAP that includes the LCoA and the RCoA.

Upon receiving the LBU message, the MAP verifies the ownership of the claimed LCoA and the RCoA, using the public key certificate CGA-based verification. The process of verification is shown in Fig.2 and detailed below:

a) The MAP first verifies the claimed LCoA. The MAP divides the LCoA into a subnet prefix-(64-bit) and an *IID*-(64-bit).

b) Concatenate from the left of the modifier and the subnet prefix. Execute the HMAC_SHA1 function on the concatenation using the MAP's public key (PK_{MAP}) and obtain the leftmost 64 bits of the output. The result is H_V, that is, H_V *= First (64, HMAC_SHA1 (*P_{MAP}*, (modifier ||subnet prefix))).*

c) Compare the calculated hash value (H_V) obtained from step *'b'* with the *IID*-(64-bit) obtained from step *'a'*; the differences in the U/L and the I/G bits are ignored.

d) The MAP then verifies the claimed RCoA. The MAP divides the RCOA into the MAP prefix-(64-bits) and the *IID*-(64-bit).

e) Concatenate the LCoA with the MAP's public key (PK_{MAP}) to compute the *IID"*, that is, *IID" = First (64, HMAC_SHA1 (*PK_{MAP}*|| LCoA)).*

f) Compare the calculated *IID"*-(64-bit) value obtained from step *'e'* with the *IID*-(64-bit) obtained from step *'d'*. The MAP then forms the RCoA, that is, *RCoA = (MAP-subnet prefix (64 bit) || IID" (64 bit)).*

g) The MAP then performs a duplicate address detection test at the RCoA. If an address collision is detected, increment the modifier by one and return to step '*b'.*

If both address verifications are successful, the MAP will gain confidence that the CGA-based LCoA and RCoA were generated and sent first by the AR and then by the MN within the WMN domain, either belonging to the MN itself or are non-used addresses. The reason is that with the public key certificate CGA-based technique in place, a malicious MN will need about (2^{61}) attempts to produce either the LCoA or the RCoA that matches a third party's IPv6 address.

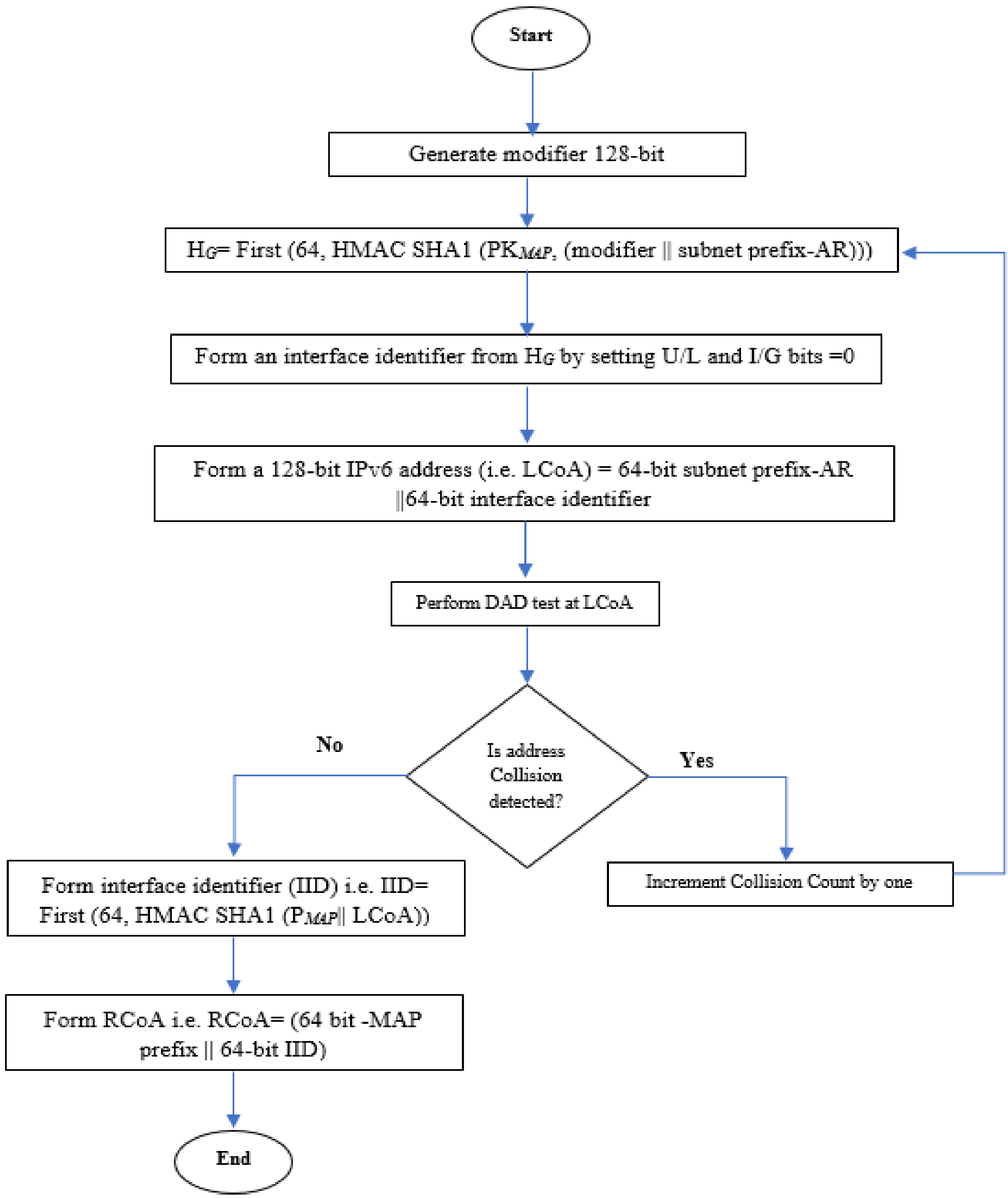


Fig. 1. LCoA and RCoA Generation Process.

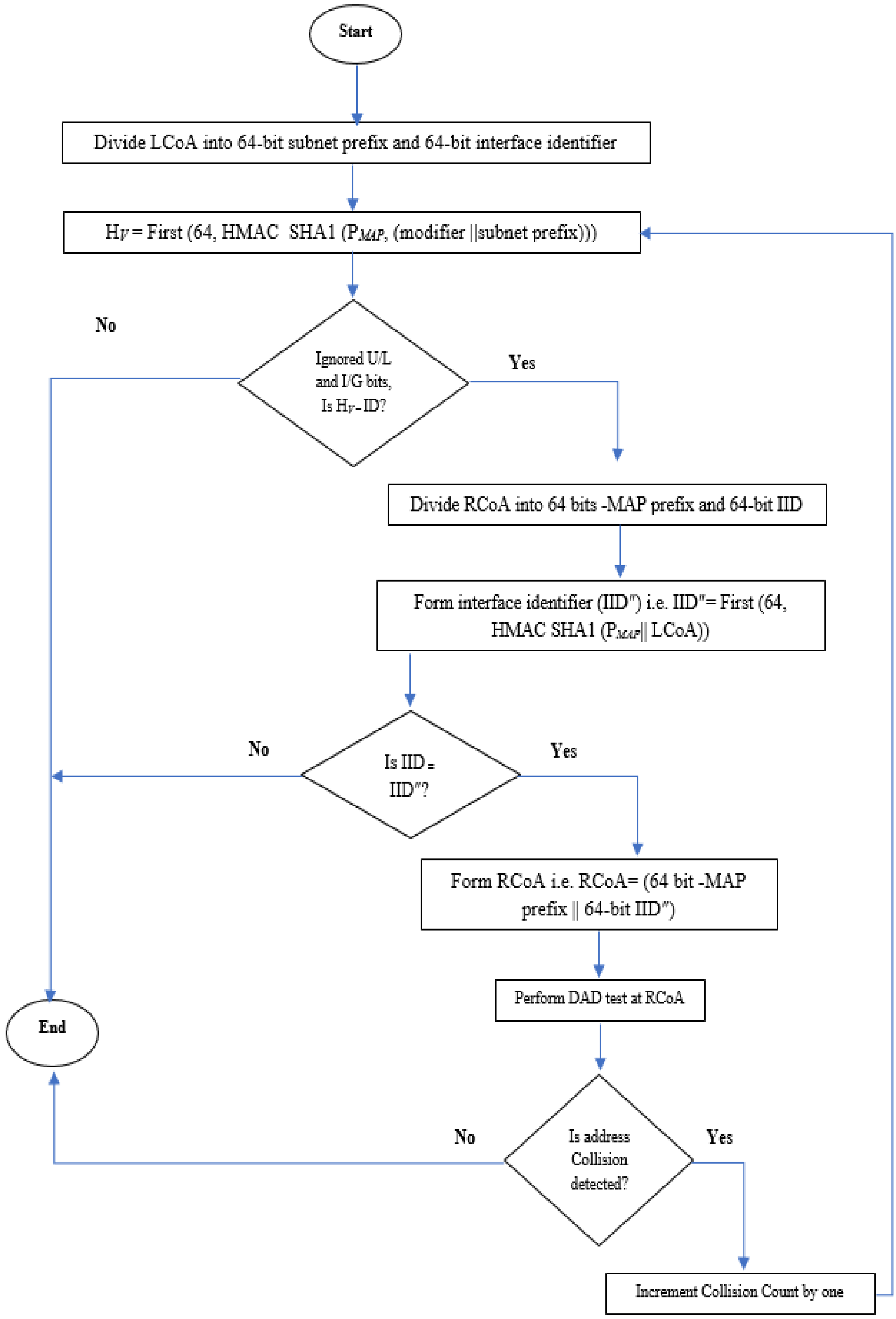


Fig. 2. LCoA and RCoA Verification Process.

B. *A Light-Weight LCoA Reachability Test Method*

The aim of the light-weight reachability test method to allow the MAP to confirm the reachability of the MN at the claimed LCoA. To doing so, the AR assists the MAP to ensure that the MN is indeed located at the LCoA. The AR generates a fresh care-of keygen token (CoT) value based on the secret key ($K_{AR\text{-}MAP}$) that is shared with the MAP, and the AR then securely sends the CoT value to both the MN and the MAP.

The reachability test is initiated as soon as the MAP receives a valid LBU message from the MN. The MN includes the received CoT value from the AR to show its presence at the claimed LCoA; in other words, the MN sends an LBU message containing the CoT value to the MAP. If the MAP successfully verifies the CoT value, it can then ensure that the MN is indeed located at the claimed LCoA. This method prevents a malicious MN from launching a flooding attack.

The LCoA reachability test is new in the HMIPv6 protocol context; this is the first test that incorporates the CoT's value to enable the MAP to verify the reachability of the MN at the claimed LCoA, and it does not affect the E-MAP registration protocol performance in terms of registration delay and signalling overhead because it does not require an extra message between the MN and the MAP to confirm the LCoA reachability.

V. E-MAP Registration Protocol in Detail

This section presents a detailed description of the E-MAP registration protocol. As shown in Fig. 3, the E-MAP registration protocol executes five messages to perform the MAP registration process. Each message has a specific name, as follows: Router Solicitation (RtSol); Pre-binding Update (PBU); Router Acknowledgement (RtAck); Local Binding Update (LBU); and Binding Acknowledgement (BA).

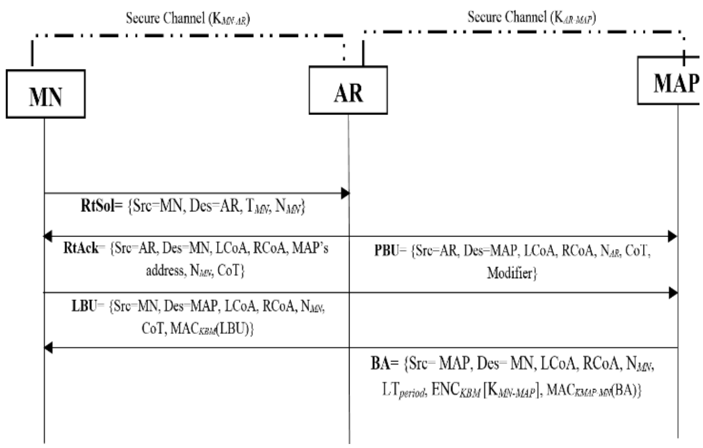


Fig. 3. E-MAP Registration Protocol Sequence Messages.

The E-MAP registration protocol is detailed in the following procedures:

- When the MN roam into the WMN domain, it registers with the MAP in the domain by using the E-MAP registration protocol. The MN initiates the E-MAP registration protocol by securely sending the RtSol message – as shown in (1) – to AR requesting to generate LCoA and RCoA. The MN includes a fresh timestamp (T_{MN}) and fresh nonce (N_{MN}). The T_{MN} is used to protect the AR against replay attack, and the N_{MN} is used to protect the MN against replay attack when it finds a response from the AR.

$$RtSol = \{T_{MN}, N_{MN}\} \quad (1)$$

- Upon the AR received message RtSol message, it checks the value of T_{MN} to confirm the freshness of the message. Upon successful verification, the AR runs the public key certificate CGA-based generation to configure LCoA and RCoA for MN as stated in Section (IV). The AR then generates a CoT – as shown in (2) – based on the secret key ($K_{AR\text{-}MAP}$) shared with the MAP and the fresh nonce N_{AR}.

$$CoT = First\ (64, HMAC\_SHA1\ (K_{AR\text{-}MAP}, (LCoA|| RCoA||N_{AR}))) \quad (2)$$

The AR then sends the PBU message – as shown in (3) – to the MAP. At the same time, the AR sends the RtAck message – as shown in (4) – to the MN via an IPSec ESP secure tunnel.

$$PBU = \{LCoA, RCoA, Modifier, N_{AR}, CoT, MAC_{KAR\text{-}MAP}(PBU)\} \quad (3)$$

$$RtAck = \{LCoA, RCoA, MAP_{address}, N_{MN}\} \quad (4)$$

- When the MAP received PBU message, it checks the value of CoT to verify the freshness of message. Upon successful verification, the MAP verifies the integrity and authenticity of the received message using the key $K_{AR\text{-}MAP}$ shared with AR. If any of these verifications fails, the MAP will discard the message without any further action. Otherwise, the MAP creates a binding cache entry for the MN, in which it stores the LCoA, RCoA, CoT and Modifier carried by the PBU message. Once the binding cache entry is created, the MAP waits for a limited amount of time for the owner of those addresses to send the LBU message. If no valid LBU message is received from the MN during the binding cache entry's preconfigured lifetime, then the MAP will delete it.

- When the MN received an RtAck message, it decrypts the message using the ($K_{MN\text{-}AR}$) key, and it checks the value of N_{MN} to verify the freshness of the message. If the verification fails, the MN will discard the message without any further action. Otherwise, the MN creates a binding list entry for the MAP and sets the status to Binding_Pending, indicating that it is waiting for acknowledgements from the MAP. The MN stores the LCoA and the RCoA values in its binding list entry and then generates fresh sequence numbers (Seq_{MN}) that it will use to detect any replay attack when it finds a response from the MAP. The MN then hashes a CoT value enclosed in the RtAck message to generate a binding management key (K_{BM}), as shown in (5). The MN then uses the MAP's address enclosed in the RtAck message to send an LBU message – as shown in (6) –to the MAP.

$$K_{BM} = SHA1\ (CoT) \quad (5)$$

$$LBU= \{LCoA, RCoA, Seq_{MN}, CoT, Ack, MAC_{KBM}(LBU)\} \quad (6)$$

- Upon the MAP received an LBU message, it checks the value of CoT to verify the freshness of the message and the reachability of MN the at claimed LCoA (as stated in Section IV). If the verification fails, the MAP will discard the message without any further action. A positive outcome from this verification check assures the MAP that the LBU message is fresh and comes from a node that is reachable in the LCoA, which provides some assurance of the MN's honesty before heavy computations are performed. This verification protects the MAP against replay attacks and resource exhaustion from denial-of-service attacks. Otherwise, The MAP verifies the ownership of the claimed LCoA and the RCoA, using a public key certificate CGA-based technique (as stated in Section IV). Aftre sucessful the ownership verfication, the MAP hashes the value of CoT to generate the binding key management (K_{BM}). The MAP then verifies that MAC_{KBM} (LBU) value. A positive result from this verification check assures the MAP that the LBU message is indeed from the MN and has not changed in transit. If any of above verifications fails, the MAP will discard the LBU message with no further action. Otherwise, the MAP updates the values of the MN's LCoA and RCoA and stores the Seq_{MN} value of the MN that was carried in the LBU message in a binding cache entry for the MN. The MAP then generates a fresh session key ($K_{MN\text{-}MAP}$), and sets a lifetime period between the LCoA and RCoA of MN in the binding cache entry to the maxium lifetime value to reduce the number of redundant binding refreshes and by extension, signalling overheads. Finally, the MAP sends the BA message – as shown in (7) – to the MN for an acknowledgement of the binding of the LCoA and the RCoA.

$$BA = \{LCoA, RCoA, Seq_{MN}, LT_{period}, ENC_{KBM}\,[K_{MN\text{-}MAP}], MAC_{KMN\text{-}MAP}(BA)\} \quad (7)$$

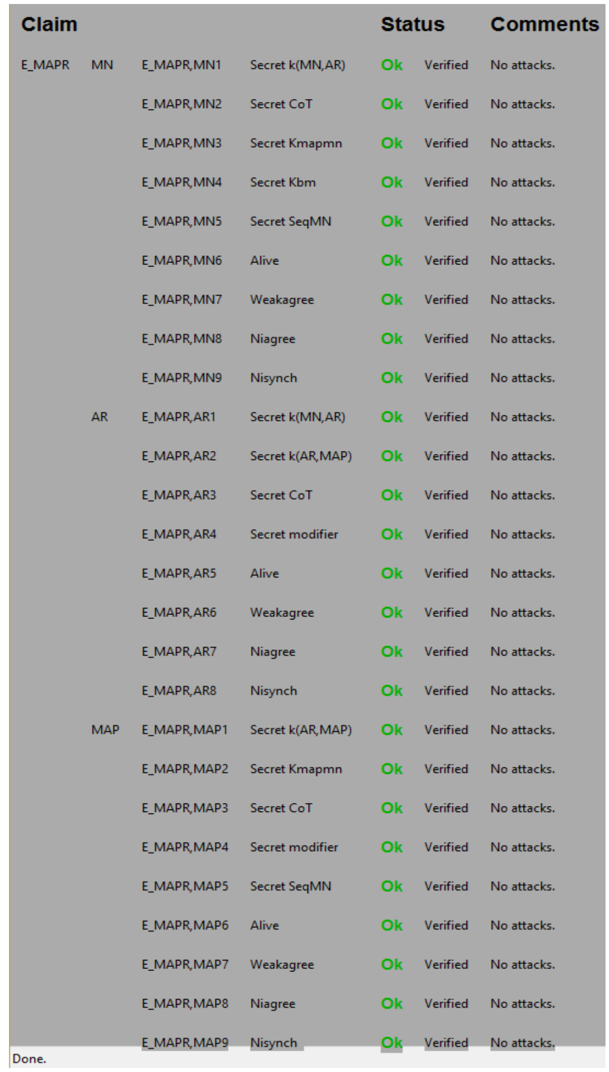
- When the MN receives a BA message, it will use the MAP's address as an index to search its Binding Update list. If a list entry is found with a Binding_Pending status, the MN will verify the freshness of the message. Upon successful verification, the MN decrypts the code session key $K_{MN\text{-}MAP}$ using the K_{BM}. The MN then verifies integrity and authenticity of the BA message. If any of these verifications fail, the MN will discard the message without any further action. Otherwise, the MN will story the values of LCoA, RCoA, lifetime period and $K_{MN\text{-}MAP}$, and change the status of binding Update list to Binding_Complete at match list entry. In this case, the E-MAP registration protocol is completed.

VI. Formal Security Analyses

This section formally verifies the accuracy of the E-MAP registration protocol presented in Section V. To doing so, the Scyther tool [22] has been used. According to [23], the Scyther tool is one of the fastest tools which stills finding attacks efficiently. To model the E-MAP registration protocol in the Scyther tool, the security protocol description language (SPDL) has been used.

Fig.4, summarises the experimental outcomes of the automatic formal verification of the E-MAP registration protocol. The figure shows that the results of the verification can confirm the security properties (i.e., the secrecy of the exchanged values) to guarantee that the information has not been stolen but exchanged safely. In addition, it shows that each role in the E-MAP registration protocol (i.e., MN, AR and MAP) meets the four major authentication forms those defined in the Scyther tool [24],[25] namely: Aliveness (*Alive*), Non-injective agreement (*Niagree*), Non-injective synchronisation (*Nisynch*) and Weak agreement (*Weakagree*).

To conclude, the outcome means that no attacks are found on Scyther's automatic security claim verifications for the E-MAP registration protocol.



Claim				Status		Comments
E_MAPR	MN	E_MAPR,MN1	Secret k(MN,AR)	Ok	Verified	No attacks.
		E_MAPR,MN2	Secret CoT	Ok	Verified	No attacks.
		E_MAPR,MN3	Secret Kmapmn	Ok	Verified	No attacks.
		E_MAPR,MN4	Secret Kbm	Ok	Verified	No attacks.
		E_MAPR,MN5	Secret SeqMN	Ok	Verified	No attacks.
		E_MAPR,MN6	Alive	Ok	Verified	No attacks.
		E_MAPR,MN7	Weakagree	Ok	Verified	No attacks.
		E_MAPR,MN8	Niagree	Ok	Verified	No attacks.
		E_MAPR,MN9	Nisynch	Ok	Verified	No attacks.
	AR	E_MAPR,AR1	Secret k(MN,AR)	Ok	Verified	No attacks.
		E_MAPR,AR2	Secret k(AR,MAP)	Ok	Verified	No attacks.
		E_MAPR,AR3	Secret CoT	Ok	Verified	No attacks.
		E_MAPR,AR4	Secret modifier	Ok	Verified	No attacks.
		E_MAPR,AR5	Alive	Ok	Verified	No attacks.
		E_MAPR,AR6	Weakagree	Ok	Verified	No attacks.
		E_MAPR,AR7	Niagree	Ok	Verified	No attacks.
		E_MAPR,AR8	Nisynch	Ok	Verified	No attacks.
	MAP	E_MAPR,MAP1	Secret k(AR,MAP)	Ok	Verified	No attacks.
		E_MAPR,MAP2	Secret Kmapmn	Ok	Verified	No attacks.
		E_MAPR,MAP3	Secret CoT	Ok	Verified	No attacks.
		E_MAPR,MAP4	Secret modifier	Ok	Verified	No attacks.
		E_MAPR,MAP5	Secret SeqMN	Ok	Verified	No attacks.
		E_MAPR,MAP6	Alive	Ok	Verified	No attacks.
		E_MAPR,MAP7	Weakagree	Ok	Verified	No attacks.
		E_MAPR,MAP8	Niagree	Ok	Verified	No attacks.
		E_MAPR,MAP9	Nisynch	Ok	Verified	No attacks.

Done.

Fig. 4. Automatic Security Claims from Scyther Tool.

VII. Simulation Setup and Performance Evaluation

This section evaluates the performance of the E-MAP registration protocol. The performance is measured in terms of the MAP registration delay and signalling overhead. The MAP registration delay is defined as the total time taken to complete the registration with the MAP in the WMN domain, measured in seconds. The signalling overhead is defined as the total amount of HMIPv6 protocol signalling traffic sent and received by all involved entities in each proposed protocol. The control signalling overhead is measured in bits/second. The OPNET™ modeller [26] (version 14.5) has been chosen to simulate the performance of the E-MAP registration protocol under varying network conditions.

Depicted in Fig.5, the chosen scenario involves one MAP that connects the IPv6 internet cloud via wired uplinks and five ARs that connect with each other through a multi-hop wireless backbone. A MAP has two separate interfaces for providing wired connectivity to the internet and wireless connectivity to form the mesh backbone. Every AR also has two interfaces, and both are for wireless connectivity. These interfaces support the operation of a router in two separate channels. One channel forms a wireless mesh backbone to route packets for the MNs. For the other channel for MN access, the AR uses an ordinary IEEE 802.11g MAC protocol with a data rate of 54 Mbps and a transmission power of 0.1 W with a power threshold of 20 dBm, and it operates within a 100-m range to simulate the AR so that Wi-Fi-compatible devices can easily join the mesh domain. This IEEE 802.11g MAC protocol is used in the WMN [1]. The figure also shows that the three movement trajectories by the white lines of the MN have been made and represent the three different scenarios of handoff of the MN with the AR. The scenarios dependent on the number of hops between the MN and MAP, e.g. one, three and five hops. When the MN enter the domain, the MN performs the link layer (L2) handoff with the AR and then initiates to perform the network layer (L3) handoff by operating the registration process with MAP.

The following subsection presents the analyses of the results obtained from the simulation study of the MAP registration delay and the control signalling overhead. It compares the results of the E-MAP registration protocol, the B-MAP registration protocol [4] and the HMIPv6sec protocol [16].

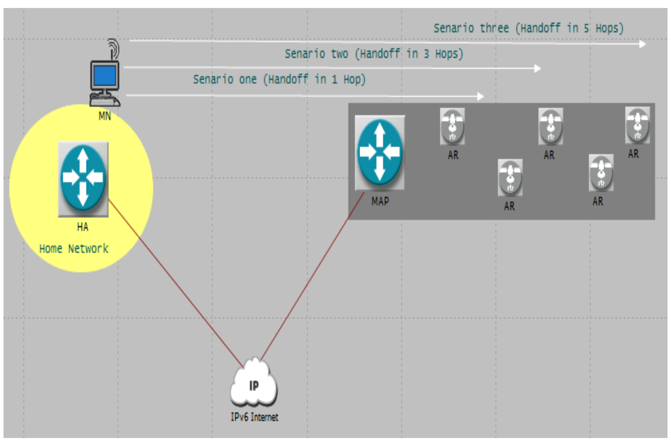


Fig. 5. MAP Registration Process Simulation Model.

A. *MAP Registration Delay*

This subsection presents an analysis of the MAP registration delay simulation results. A selection of the simulation results is shown in Figures 5-8. In these results, the time spent to examine the uniqueness of the LCoA and the RCoA using the duplicate address detection test is set to zero. The rationale for this setting is that during the implementation of the E-MAP registration protocol, the B-MAP registration protocol and the HMIPv6sec protocol, the duplicate address detection test obtained varying values ranging from 1 second to 1.4 seconds, which could mean that these values affect the accuracy of the collected results.

Fig.6 shows the MAP registration delay at varying numbers of hops (i.e., one, three and five). The figure shows that the MAP registration delay slightly increases as the number of hops increases. This is because the intermediate devices in the WMN domain require time to examine a packet header to determine where to direct it, causing the MAP registration delay to decrease.

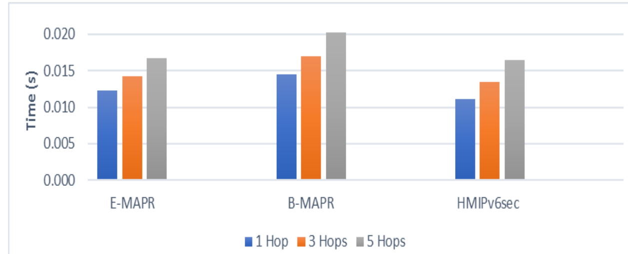


Fig. 6. MAP Registration Delay for E-MAP, B-MAP and HMIPv6sec Protocols Versus Varying Numbers of hops (one, three and five).

For further comparative analysis, Fig.7 shows the average MAP registration delay under different levels of background traffic among the ARs when there are three hops between the MN and the MAP. This figure illustrates that as the background traffic volume increases, ARs become more congested, and the MAP registration delay increases significantly under all protocols. Fig. 8 also shows that on average, the B-MAP registration protocol is around 17% and 19% higher than the E-MAP registration protocol and the HMIPv6sec registration protocol, respectively. This is because the B-MAP registration protocol requires eight messages (i.e., one message to receive the RA message includes the MAP option from the AR plus six messages in the IKEv2 protocol, and two additional messages to bind the LCoA and the RCoA at the MAP. It is also shown that the MN spends around 4% to register its LCoA and RCoA at the MAP in the E-MAP registration protocol, which is higher than in the HMIPv6sec protocol. This is because of (1) the additional two HMAC_SHA1 operations performed by the MAP to verify the reachability of the MN at the LCoA and (2) the increase in the size of the PBU and the LBU messages exchanged in the E-MAP registration protocol compared with the size of these messages in the HMIPv6sec protocol.

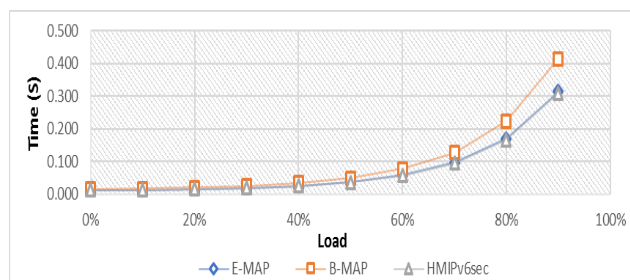


Fig. 7. Average MAP Registration Delay for E-MAP, B-MAP and HMIPv6sec Protocols Versus Load (one MN at three hops).

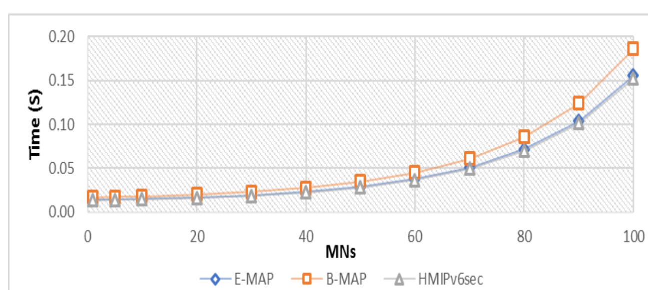


Fig. 8. Average MAP Registration Delay for the E-MAP, B-MAP and HMIPv6sec Protocols Versus the number of MNs.

Fig.8 compares the performance of the E-MAP, B-MAP and HMIPv6sec protocols based on the number of simultaneously visiting MNs served by MAP, which increases when those MNs are three hops away from the MAP, and the background traffic is zero. As shown in this figure, the channel contention at the MAP slightly increases when the number of simultaneously roaming MNs increases from one MN to 50 MNs, but when the number increases to more than 60 MNs, the network load and the channel congestion at the MAP increase exponentially. As also shown in the figures, the E-MAP registration protocol offers a lower rate of MAP registration delay than the B-MAP registration protocol. The reason is that the MAP in the E-MAP registration protocol is required to perform fewer operations during registration than the B-MAP registration protocol. Consequently, the queuing time on the MAP side of the E-MAP registration protocol increases faster than in the B-MAP registration protocol. On the other hand, the E-MAP registration protocol produces slightly more operations than the HMIPv6sec protocol. Therefore, as the number of simultaneously roaming MNs increases, the performance gap between the E-MAP registration protocol and the HMIPv6sec protocol grows slightly.

B. Signalling Overhead

This section analyses the control signalling overhead results for the E-MAP, B-MAP and HMIPv6sec protocols. Table 2 shows the numerical results for the control traffic received by and sent from the MN, AR and MAP sides.

The following observations can be made from Table 2: (1) Generally, the E-MAP registration protocol receives and sends a high amount of control traffic compared with the B-MAP and HMIPv6sec protocols because the E-MAP registration protocol requires an extra length of signalling messages received and/or sent at all that entities (2) The amount of control traffic received at the MN in the E-MAP registration protocol is around 19.4% and 16.3% higher than in the B-MAP and HMIPv6sec protocols, respectively; additionally, the amounts of control traffic sent at the MN in the E-MAP and B-MAP registration protocols are identical. (3) The amounts of control traffic received and sent at the AR in the E-MAP and HMIPv6sec protocols are somewhat identical, but zero bit/second is noted in the B-MAP registration protocol. (4) The amount of control traffic received and sent at the MAP in the E-MAP protocol is higher than in the B-MAP and HMIPv6sec protocols.

TABLE II. CONTROL TRAFFIC RECEIVED AND SENT ON THE MN, AR AND MAP SIDES IN E-MAP, B-MAP AND HMIPV6SEC PROTOCOLS

Protocol	**Control traffic (bit/second)**	**MN**	**AR**	**MAP**
E-MAP	Received	4.90	1.83	5.19
	Sent	3.66	2.38	2.83
B-MAP	Received	4.03	0	2.01
	Sent	3.66	0	1.83
HMIPv6sec	Received	4.16	1.83	4.43
	Sent	3.13	2.12	2.54

C. Discussion

The simulation study involving the E-MAP registration protocol reveals the following findings: (1) Increasing the number of hops between the MN and the MAP has an insignificant effect on the performance of the E-MAP, B-MAP and HMIPv6sec protocols in terms of the MAP registration delay. (2) Increasing the background traffic volume among the ARs in the WMN domain has a lower impact on the performance of the E-MAP registration protocol than that of the B-MAP registration protocol but has a slightly higher effect compared with the HMIPv6sec protocol. (3) The impact of increasing the number of simultaneously roaming MNs served by the same MAP on the performance of the E-MAP registration protocol is lower compared with the B-MAP registration protocol and higher compared with the HMIPv6sec protocol. (4) The E-MAP registration protocol receives and sends more control traffic at the MN, the AR and the MAP as a cost to pay for supporting the location authentication of the MN to the MAP, allow the MN to use the WMN domain services and compute the shared secret key between the MN and the MAP.

To conclude, if the E-MAP, B-MAP and HMIPv6sec protocols are compared based on efficiency and security, the E-MAP registration protocol emerges as the superior option.

VIII. CONCLUSION AND FUTURE WORK

This paper has presented the design of the E-MAP registration protocol that allows the MAP to verify the MN's ownership of the claimed LCoA and RCoA, as well as the reachability of the MN at the LCoA, and to securely identify and compute the shared secret key with the MN in the WMN domain. This study has combined two ideas. First, it has generated the MN's LCoA and RCoA using a public key certificate CGA-based technique. Second, it has applied a novel lightweight method to confirm the reachability of MN at claimed LCoA. In addition, the MAP computes the shared secret key with MN. Via these actions, the E-MAP registration protocol reduces the likelihood that a malicious MN can successfully steal a third party's node addresses (i.e., the LCoA and the RCoA), prevents a malicious MN from launching a flooding attack and protects any future BU against a false BU attack. The formal security analysis using the Scyther tool has demonstrated that no attacks haven found in the E-MAP registration protocol. Additionally, the performance evaluation has proven that increasing the number of hops in the WMN domain has little effect on the performance of the MAP registration process in the E-MAP, B-MAP and HMIPv6Sec protocols. The E-MAP registration protocol also offers a lower registration delay than the B-MAP protocol but higher than the HMIPv6sec protocol. Moreover, the E-MAP registration protocol introduces a higher signalling overhead compared with the B-MAP and HMIPv6sec protocols as a cost to support the MN location at the WMN domain, allow the MN to use the services of the WMN domain and configure the shared secret key between the MN and the MAP. Further research is recommended to extend the E-MAP registration protocol to securely handle MNs roaming within the WMN domain.

ACKNOWLEDGMENT

Sultan Alkhliwi gratefully acknowledges the Northern Border University, 'KSA' for financial support

REFERENCES

[1] I. F. A. and Xudong, Wireless Mesh Networks (Advanced Texts in Communications and Networking). John Wiley & Sons, 2009.

[2] C. K. Toh, Ad Hoc Mobile Wireless Networks: Protocols and Systems. Prentice Hall PTR, 2002.

[3] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," RFC 3775 (Proposed Standard), 2004.

[4] H. Soliman, C. Castelluccia, K. ElMalki, and L. Bellier, "Hierarchical Mobile IPv6 (HMIPv6) Mobility Management," RFC 5380 (Proposed Standard), 2008.

[5] T. Aura, "Mobile IPv6 Security," in Security Protocols, 2004, pp. 215–234.

[6] A. Moravejosharieh, H. Modares, and R. Salleh, "Overview of Mobile IPv6 security," Proc. - 3rd Int. Conf. Intell. Syst. Model. Simulation, ISMS 2012, pp. 584–587, 2012.

[7] G. O'Shea and M. Roe, "Child-proof authentication for MIPv6 (CAM)," Acm Sigcomm, vol. 31, no. 2, p. 4, 2001.

[8] T. Aura, "Cryptographically Generated Addresses (CGA)," RFC 3927 (Proposed Standard), 2005.

[9] E. Kaufman, C., "Internet Key Exchange (IKEv2) Protocol," RFC 4306 (Proposed Standard), 2005.

[10] S. K. and K. Seo, "Security Architecture for the Internet Protocol," RFC 4301 (Proposed Standard), 2005.

[11] S. Bridglal et al., The TCP / IP Guide. .

[12] R. Kandikattu and L. Jacob, "Comparative analysis of different cryptosystems for hierarchical mobile IPv6-based wireless mesh network," Int. J. Netw. Secur., vol. 10, no. 3, pp. 190–203, 2010.

[13] H.-S. Kang and C.-S. Park, "Authenticated Fast Handover Scheme in the Hierarchical Mobile IPv6," Wisa, vol. 4298, p. 211-224--, 2007.

[14] C. Perkins and P. Calhoun, "Authentication, Authorization, and Accounting (AAA) Registration Keys for Mobile IPv4," 2005.

[15] I. You, J. H. Lee, K. Sakurai, and Y. Hori, "ESS-FH: Enhanced security scheme for fast handover in hierarchical mobile IPv6," IEICE Trans. Inf. Syst., vol. E93–D, no. 5, pp. 1096–1105, 2010.

[16] W. Haddad, S. Krishnan, and H. Soliman, "Using Cryptographically Generated Addresses (CGA) to secure HMIPv6 Protocol (HMIPv6sec)," Expired Internet-Draft: draft-haddad-mipshop-hmipv6-security-06, 2006.

[17] C. Bauer, Secure and Efficient IP Mobility Support for Aeronautical Communications. KIT Scientific Publishing, 2013.

[18] C. Vogt, R. Bless, M. Doll, and T. Kuefner, "Early binding updates for Mobile IPv6," IEEE Wirel. Commun. Netw. Conf. WCNC, vol. 3, no. C, pp. 1440–1445, 2005.

[19] W. Haddad, L. Madour, J. Arkko, and F. Dupont, "Applying Cryptographically Generated Addresses to Optimize MIPv6 (CGA-OMIPv6)," Expired Internet-Draft: draft-haddad-mip6-cga-omipv6-04.txt, 2005.

[20] O. Elshakankiry, A. Carpenter, and N. Zhang, "A Novel Scheme for Supporting Location Authentication of Mobile Nodes," in In Proceeding of the Second International ICST Conference on Security and Privacy in Mobile Information and Communication Systems (MobiSec2010), 2010, pp. 91–102.

[21] T. J. S. Thomson, T. Narten, "IPv6 Stateless Address Autoconfiguration," RFC 4862 (Draft Standard), 2007.

[22] C. J. F. Cremers, "Scyther : Unbounded Verification of Security Protocols," no. 572, pp. 1–18, 2006.

[23] C. J. F. Cremers, P. Lafourcade, and P. Nadeau, "Comparing State Spaces in Automatic Security Protocol Analysis," in Formal to Practical Security: Papers Issued from the 2005-2008 French-Japanese Collaboration, V. Cortier, C. Kirchner, M. Okada, and H. Sakurada, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 70–94.

[24] G. Lowe, "A hierarchy of authentication specifications," 10th Comput. Secur. Found. Work., pp. 31–43, 1997.

[25] C. J. F. Cremers, S. Mauw, and E. P. de Vink, "Injective synchronisation: An extension of the authentication hierarchy," Theor. Comput. Sci., vol. 367, no. 1–2, pp. 139–161, 2006.

[26] "Riverbed Modeler." [Online]. Available: https://www.riverbed.com/gb/products/steelcentral/opnet.html. [Accessed: 24-Jan-2016].

Voice Pathology Recognition and Classification using Noise Related Features

HAMDI Rabeh[1], HAJJI Salah[2], CHERIF Adnane[3]
Faculty of Mathematical, Physical and Natural Sciences of Tunis, University of Tunis El Manar, Tunisia
Laboratory of Analysis and Processing of Signals And Electric and Energy System, 1060 Tunis, Tunisia

***Abstract*—Nowadays, the diseases of the voice increase because of bad social habits and the misuse of voice. These pathologies should be treated from the beginning. Indeed, it is no longer necessary that the diseases of the voice lead to affect the quality of the voice as heard by a listener. The most useful tool for diagnosing such diseases is the Acoustic analysis. We present in this work, new expression parameters in order to clarify the description of the vocal signal. These parameters help to classify the unhealthy voices. They describe essentially the fundamental frequency F0, the Harmonics-to-Noise report (HNR), the report Noise to Harmonics Ratio (NHR) and Detrended Fluctuation Analysis (DFA). The classification is performed on two Saarbruecken Voice and MEEI pathological databases using HTK classifiers. We can classify them into two different types: the first classification is binary which is used for the normal and pathological voices; the second one is called a four-category classification used in spasmodic, polyp, nodule and normal female voices and male speakers. And we studied the effects of these new parameters when combined with the MFCC, Delta, Delta second and Energy coefficients.**

Keywords— HTK; MFCC; MEEI; SVD; pathological voices

I. Introduction

Many pathologies, may affect the voice as nodules, spasmodic folds, polypoid causing irregular vibrations due to the malfunction of many factors that contribute to vocal vibrations. Beside, pathologies of the voice may affect differently the vibration of the vocal field, first of all it depends on the type of disorder, but also the location of the disease in the folds of the voice, so it allows them to produce different shades of base [15].

To tackle those problems of the voice, digital processing on voice signals is a found tool that helps with nonvasive analysis for doctors. It allows identification of vocal disorders especially from the beginning [16].

The disease affecting more people is the dysphonia because of the disruption of the speech. There are various types of dysphonia. First, the dysfunctional dysphonia is characterized in some obstacles of pronunciation but without changing the organic composition of the vocal cords. The dysfunctional dysphonia can lead to organic dysphonia because of the application of compensation by the patient. Second the organic dysphonia is a pathological change in the vocal cords. Third, we note neurological dysphonia. To evaluate and determine the therapy, the evaluation of the voices is very relevant. The quality of voice could be assessed by diagnosis or by the laryngostroboscopy testing as. Two different approaches are involved: the perception and the objective approach.

On the other hand, to establish the subjective measurement of voice quality it should be based on the individual experience. The subjective measurement may vary. The method detection of automatic voice-pathology can be accomplished by various types of signal analysis which can be long term or short-term. These parameters can be determined using cepstral coefficients with Mel frequency [13] [14], linear predictive cepstral coefficients (LPCC) [12], and so on. The old research presented different tools to establish an evaluation. Obviously, in the related work many methods of acoustic diagnosis of pathological voices have been proposed. Between them, a big attention was given to the automatic classification of the troubled voice. For classification of pathological voices there are very important classifiers such as: hidden Markov model (HMM) [19] as well as the neural networks [17], the support vector machines (SVM) [20] and finally the Gaussian mixing model (GMM) [18].

A normal binary / pathological classification of vocal samples [1, 2] has been proposed in the literature, the best performances are obtained by using specific parameters of the HMM classification. However, few studies that have classified the pathologies [3] and the obtained results were not effective.

In the present work, the classification of pathological voices is studied using the method of extraction of the parameters MFCC with energy, derivative and acceleration combined with the prosodic parameters, noise-to-harmonic ratio (NHR), harmonic-to-noise ratio (HNR), Detrended Fluctuation Analysis (DFA) and fundamental frequency (F0) which are calculated for each frame. To validate this work we used two bases to give MEEI Database and Saarbruecken Voice Database. The aim of this work is to show the ability of these parameters to detect and classify pathologies of the voice, using a scenario where these parameters are used alone with MFCC and hybrid.

II. Methods and Materials

The classical characteristics are derived from the benchmarks used in the domain of acoustic recognition. These parameters are essentially the analysis of trend fluctuations (DFA), the harmonic/noise ratio (HNR), the fundamental frequency F0 and the Cepstral coefficient of the frequency Mel (MFCC) combined with the energy and the first and second derivatives.

The characteristics involved in the pathological voice which are the most common are described in the section below.

A. Fundamental Frequency

The fundamental frequency (F0) is one of the essential parameters in acoustic measurement. This frequency expresses the vibration rate of the voice fold. This setting describes the voice state. It is sometimes used with the Mel-Frequency (MFCC) Cepstral coefficients in the form of conjunction.

B. Mel-Frequency Cepstral Coefficients

The parameter MFCCs is used to decrease the voice signal redundancy. It is also used in other areas such as voice recognition [4]. The calculation of these coefficients is done by using the method of weighting the signal Fourier transform through a bank of filters distributed on a "Mel" scale, then from this weighted spectrum by calculating the cepstrum and at the end calculate the discrete cosine transformation for this cepstrum.

MFCC belongs to a family of parameters that are used in speech processing. Based on the human knowledge of the sounds, MFCC does a frequency analysis of the signal. By listening to the signal an experienced therapist can detect the presence of a speech disorder [2]. For each frame, the extraction procedure is done after a 16 kHz interpolation, with a bank of 29 Mel filters and a 25 ms with a 10 ms step, to get 12 MFCC plus log-energy, Delta and Delta seconds.

C. Noise to Harmonics Ratio and Harmonics to Noise Ratio

The harmonic-noise ratio HNR measures objectively the feeling of perception in a hoarse voice [5]. The calculation of the harmonic-noise ratio, the signal must be dropped sampled at 16 KHz, and divided into 25 ms length of the frames, with a step of 10 Ms. The comb filter is applied to the signal in each frame, in order to calculate the energy in the components Harmonics. For the logarithm of this quantity, the logarithmic energy of the noise is inferred to obtain the HNR.

D. Detrended Fluctuation Analysis (DFA)

Detrended Fluctuation Analysis characterizes the extent of turbulent noise in the speech signal, quantifying the stochastic self-similarity of the noise caused by turbulent air flow in the vocal tract, e.g. incomplete vocal fold closure can increase the DFA value. It is applied to parole signals, shows the ability to detect voice disorders in general. [6]

III. The Results

A. Databases

In all this work we use two different databases MEEI data base and Saarbruecken database. In the first data base the voices samples are based mainly on the phonation of vowels [a] whose duration is about 3 or 4 s by men and women. And in the second data base we used the recording of the phrase "Guten Morgen, wie geht es Ihnen?" (" Good morning, how are you?").

Table 2 gives the number of samples of the pathological voices of each base.

TABLE I. An Overview of the Measures of Dysphonia used in this Study

Measure	Motivation	Number of features
Mel Frequency Cepstral Coefficients (MFCC)	The vocal pathologies cause a drop of the control of the articulators (vocal music channels). and MFCCs are trying to analyze the voice conduit regardless of the vocal cords	13-26-39, depends on extracted components and the use of delta and delta-delta coefficients
Harmonics-to-Noise Ratio (HNR) and Noise to Harmonics Ratio (NHR)	In the pathologies of the voice, the incomplete closure of the vocal fold cause an increase of the noise due to turbulent air flow. MST and NHR measure information report real signal to noise.	2
Fundamental Frequency (F0)	Average vocal fundamental frequency	1
Detrended Fluctuation Analysis (DFA).	Quantify the stochastic self-similarity of the noise caused by turbulent airflow	1

1) MEEI Database

MEEI-KayPENTAX is the database that was invented by the Massachusetts Eye and Ear Infirmary Voice and Speech Labs Corp and was published in 1994. The recordings are manifested in a sustained phonation of the vowel/Ah/(53 normal and 657 pathological) and the statement of the first sentence of the Rainbow passage (53 normal and 662 pathological). They are pronounced by patients who have these types of diseases like neurological, organic, traumatic and psychogenic speech disorders from the beginning of the disease to the complete elaboration. The recording environment of speech samples has the following characteristics 16 bit of resolution and the sampling frequency is about 25 khz or 50 khz. We chose a subset of voices which comprise 53 normal and 60 pathological [7].

2) Saarbruecken Voice Database

Recently the Saarbruecken Voice Database was published online [8]. This database is a collection of voice recordings of more than 2000 people; a recording session contains the following recordings:

- recordings of vowels /a/, /i/, /u/ produced at normal, high, low and low-high-low pitch.
- recording of sentence" Guten Morgen, wie geht es Ihnen?" (" Good morning, how are you?").

Each session contains 13 registration files. Moreover, for each case of the electroglottogram signal (EGG) is saved in a separate folder. These files contains vowels whose length is 1-3 seconds. All recordings are made in a controlled environment at 50 kHz and their resolution is 16-bit. These recordings contain 71 different pathologies, including organic and functional. 1320 pathological voices are divided into 609 males and 711 females. In other hand there are 650 normal voices (400 males and 250 females). We worked with a subset of voices which comprise 133 normal and 134 pathological.

3) Hidden Markov Model Toolkit

The Hidden Markov Model Toolkit (HTK 3.4.1) is a portable toolkit for building and manipulating hidden Markov models. HTK is primarily used for speech recognition research although it has been used for numerous other applications including research into speech synthesis, pathological voice recognition [9]. A hidden Markov model with a Gaussian mixing density (HMM-GM), five observation states (a simple model from left to right) and four diagonal state mixtures were formed for each pathological voice. [10]

TABLE II. THE DIFFERENT PATHOLOGICAL VOICES FOR THE TWO BASES

	MEEI Database		Saarbruecken Voice Database	
Diseases	**Test**	**Train**	**Test**	**Train**
Nodules	6	14	5	20
Spasmodic	6	12	21	43
Polyploid	7	15	15	30
Normal	17	36	45	88

We have developed a parametrization method that extracts the MFCC coefficients with energy, derivative and concatenated acceleration with the parameters that measure the disturbance of the vocal signal (prosodic parameters such as: F0, HNR, NHR, DFA). These parameters are calculated for each frame.

B. Global Rate Recognition for MEEI Database

Table 3 below gives the results of the rate of recognition of pathological voices for the MEEI database. For the first database, acoustic modeling is refined, estimating four-Gaussian probability densities. The recognition having the best rates are respectively obtained MFCC with all the parameters (94.44%), MFCC_NHR_DFA (91.67%) and MFCC_DFA (88.89%).

TABLE III. GLOBAL RECOGNITION RATE OF PATHOLOGIES FOR FOUR GAUSSIANS OF MEEI DATABASE

Parameters	*Recognition rate*
MFCC	72.22
MFCC_HNR	77.78
MFCC_F0	77.78
MFCC_NHR	80.56
MFCC_DFA	**88.89**
MFCC-HNR-F0	86.11
MFCC-DFA-F0	77.78
MFCC_NHR_F0	72.22
MFCC_NHR_DFA	**91.67**
MFCC_HNR-NHR-DFA-F0	**94.44**

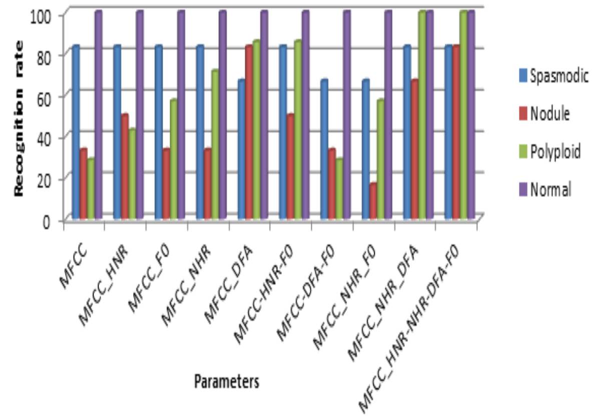


Fig. 1. Recognition rate by pathology of different techniques for 4 Gaussian of the MEEI Database.

The combination of Harmonics to Noise Ratio, Noise to Harmonics Ratio, Detrended Fluctuation Analysis and the fundamental frequency parameters has the ability to recognize the voice disease while, the Noise to Harmonics Ratio combined with the fundamental frequency shows the disability to recognize the diseases. Moreover, it appears that with the MFCC coefficients the recognition of normal voice is with high rate.

C. Global Rate Recognition for Saarbruecken Voice Database

Table 4 below gives the results of the rate of recognition of the pathological voices for the Saarbruecken Voice Database. For the second database, acoustic modeling is refined, estimating two-Gaussian probability densities. The best results are obtained when using the parameters MFCC_NHR_DFA (94.19%) and MFCC_NHR (91.86%).

The Noise to Harmonics Ratio combined with Detrended Fluctuation Analysis appear that this combination was the most able of knowing and distinguishing the types of pathologies. While, the MFCC is not able to distinguish pathological voices.

TABLE IV. GLOBAL RECOGNITION RATE OF PATHOLOGIES FOR TWO GAUSSIANS OF SAARBRUECKEN VOICE DATABASE

Parameters	*Recognition rate*
MFCC	77.91
MFCC_HNR	84.88
MFCC_NHR	**91.86**
MFCC_DFA	88.37
MFCC_F0	89.53
MFCC-HNR-F0	90.70
MFCC_NHR_F0	90.70
MFCC_DFA_F0	86.05
MFCC_NHR_DFA	**94.19**
MFCC_HNR-NHR-DFA-F0	88.37

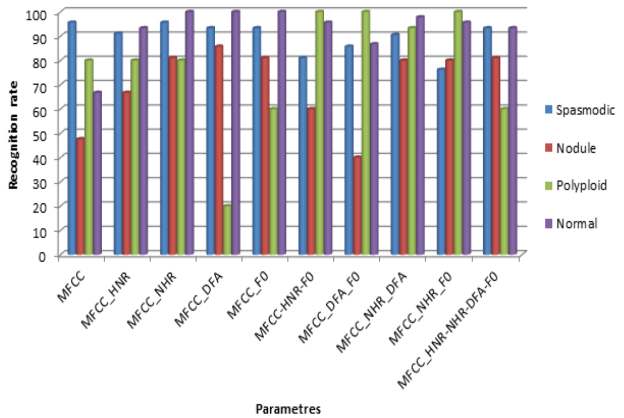


Fig. 2. Recognition rate by pathology of different techniques for 4 Gaussian of Saarbruecken Voice Database.

D. Pathologic/Normal Classification.

The results of the different experiments used for the classification and the detection of pathology are expressed by

These terms:

- Accuracy(ACC): It is the ratio of the correctly detected samples by the total number of samples used.
- Sensitivity(SN): It represents the proportion of pathological samples correctly identified.
- Specificity (SP): It is the proportion of normal samples that are negatively identified.

The following distinct equations show how to calculate these terms:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$SN = \frac{TP}{TP + FN} \quad (2)$$

$$SP = \frac{TN}{TN + FP} \quad (3)$$

The expression where true negative (TN) can be explained as follows: the system detects a normal subject as a normal subject, while the true positive (TP) means that the system detects a pathological subject as a pathological subject, besides the false negative (FN) means that the system detects a pathological issue as a normal subject and ultimately false positives (FP) means that the system detects a normal subject matter as a pathological subject. [11]

The extracted parameters from the two different databases must be checked in the detection and classification processes.

The experimental analysis shows that the data obtained varied between the databases and varied according to the types of parameters (HNR, NHR, DFA, F0 and their combination) in the same database.

The two tables 5 and 7 represent the Confusing matrix for the normal / pathological classification respectively of the MEEI and SVD databases.

Tables 6 and 8 give the results of the sensitivity, specificity and accuracy calculations of the different combinations of parameters MFCC, HNR, NHR, F0 and DFA; for normal / pathological classification of MEEI database and Saarbruecken Voice Database, respectively. These results are deduced using tables 5 and 7 and equations (1), (2) and (3).

These tables indicate the best precisions obtained for each database with the different types of parameters. Its show that the accuracy varied from one database to another for the same used characteristic, in other hand the accuracies obtained also varied for the same database as a function of parameters used to carry out the experiment.

Generally, the highest achieved accuracies are 100% for MEEI Voice Database, in the case of using the MFCC_HNR, MFCC_F0, MFCC_DFA and MFCC_F0. While using the MFCC_HNR we get the highest acquired accuracies which is 100% for Saarbruecken Voice Database.

TABLE V. CONFUSING MATRIX OF THE NORMAL / PATHOLOGIC CLASSIFICATION USING ALL VOICES OF THE MEEI DATABASE

Parameters		**Normal**	**Pathologic**
MFCC	Normal	**100**	0
	Pathologic	5.3	**94.7**
MFCC_HNR	Normal	**100**	0
	Pathologic	0	**100**
MFCC_F0	Normal	**100**	0
	Pathologic	0	**100**
MFCC_NHR	Normal	94.1	5.9
	Pathologic	5.3	94.7
MFCC_DFA	Normal	**100**	0
	Pathologic	0	**100**
MFCC-HNR-F0	Normal	94.1	5.9
	Pathologic	5.3	94.7
MFCC-DFA-F0	Normal	**100**	0
	Pathologic	0	**100**
MFCC_NHR_F0	Normal	**100**	0
	Pathologic	5.3	94.7
MFCC_NHR_DFA	Normal	94.1	5.9
	Pathologic	5.3	94.7
MFCC_HNR-NHR-DFA-F0	Normal	94.1	5.9
	Pathologic	5.3	94.7

TABLE VI. BEST DETECTION ACCURACIES IN THE MEEI DATABASE USING VARIOUS PARAMETERS

Parameters	Sensitivity	Specificity	Accuracy
MFCC	**100**	94.97	95.20
MFCC_HNR	**100**	**100**	**100**
MFCC_F0	**100**	**100**	**100**
MFCC_NHR	94.14	94.67	95.17
MFCC_DFA	**100**	**100**	**100**
MFCC-HNR-F0	94.14	94.67	95.17
MFCC-DFA-F0	**100**	**100**	**100**
MFCC_NHR_F0	**100**	94.97	95.20
MFCC_NHR_DFA	94.14	94.67	95.17
MFCC_HNR-NHR-DFA-F0	94.14	94.67	95.17

For the MEEI database, we obtained a total recognition of normal and pathological samples in each case (TN = 100% and TP = 100%) for the combinations MFCC_HNR, MFCC_F0, MFCC_DFA and MFCC_DFA_F0.

For the other combinations the recognition is not perfect in each case for the different types of voice samples.

While for Saarbruecken Voice Database we did not get total recognition of normal and pathological samples in each case, but all normal type samples are recognized as normal for MFCC_DFA and MFCC_NHR_F0. While pathological voices are recognized as pathological voices for MFCC_HNR only.

TABLE VII. CONFUSING MATRIX OF THE NORMAL / PATHOLOGIC CLASSIFICATION USING ALL VOICES OF THE SAARBRUECKEN VOICE DATABASE

Parameters		Normal	Pathologic
MFCC	Normal	**97.8**	2.2
	Pathologic	1.1	**98.9**
MFCC_HNR	Normal	97.8	2.2
	Pathologic	0	**100**
MFCC_F0	Normal	97.8	2.2
	Pathologic	1.1	98.9
MFCC_NHR	Normal	97.8	2.2
	Pathologic	11.5	88.5
MFCC_DFA	Normal	**100**	0
	Pathologic	1.1	98.9
MFCC-HNR-F0	Normal	97.8	2.2
	Pathologic	1.1	98.9
MFCC-DFA-F0	Normal	97.8	2.2
	Pathologic	1.1	98.9
MFCC_NHR_F0	Normal	**100**	0
	Pathologic	11.5	88.5
MFCC_NHR_DFA	Normal	97.8	2.2
	Pathologic	13.8	86.2
MFCC_HNR-NHR-DFA-F0	Normal	95.6	4.4
	Pathologic	14.9	85.1

TABLE VIII. BEST DETECTION ACCURACIES IN THE SAARBRUECKEN VOICE DATABASE USING VARIOUS PARAMETERS

Parameters	Sensitivity	Specificity	Accuracy
MFCC	97.82	98.89	99.39
MFCC_HNR	97.85	**100**	**100**
MFCC_F0	97.82	98.89	99.39
MFCC_NHR	97.57	89.48	88.89
MFCC_DFA	**100**	98.91	99.40
MFCC-HNR-F0	97.82	98.89	99.39
MFCC-DFA-F0	97.82	98.89	99.39
MFCC_NHR_F0	**100**	89.69	89
MFCC_NHR_DFA	97.51	87.63	86.69
MFCC_HNR-NHR-DFA-F0	95.08	86.52	85.58

IV. DISCUSSIONS AND VALIDATION

In this study we used the parameters that measure the disturbance of the vocal signal, in two databases for the detection and classification of vocal pathologies.

Indeed, the obtained results are better or comparable than the other results reported using the MEEI and SVD databases.

TABLE IX. COMPARISON OF ACCURACIES BETWEEN METHODS (PATHOLOGY DETECTION)

Method	*MEEI*	*SVD*
In this paper	100%	100%
Method [11]	99.96%	92.79%
Method [15]	88.21%	99.68%
Method [21]	94.07%	-
Method [22]	94.80%	81%

For example, in al-Nasheri et al. [15] used the MEEI and SVD databases with the SVM classifier and used the MDVP parameters to detect pathologies, obtaining accuracies of 99.68% and 88.21% respectively for SVD and MEEI.

In addition, al-Nasheri et al. [11] used the SVD and MEEI databases and used the autocorrelation and entropy parameters for the detection and classification of pathologies, obtaining respectively 99.96% and 92.79% accuracy for MEEI and SVD.

Thus Godino_Liorente et al. [21] used the MEEI database and reported an accuracy of 94.07%. While Marinez et al. [22] used the SVD database and the SVM classifier and achieved an accuracy of 81%. The authors also used the MEEI database and the accuracy gained was 94.80%.

In our study, the accuracy obtained in the case of SVD and MEEI is better at the accuracy obtained in other cases.

Table 9 illustrates the comparison between our contribution and other contributions mentioned in the related work using both bases MEEI and SVD. In our work, we obtained a high 100% accuracy for detection.

V. Conclusion

In this study, we presented our approach which is manifested in the addition of new and classical parameters to each other .Also we presented the study of the effect of the classical parameters formed by the MFCC coefficients, the energy, their first and second derivatives in the classification performances. In addition, we classified all speakers who have pathological and normal voices in binary classification.

Our contribution is tested on two pathological voice databases: SVD and MEEI only

The acoustic modeling is refined, estimating the probability densities respectively at four Gaussian for the first database MEEI and at two Gaussian for the second database SVD. The best recognition rates of the MEEI database are respectively obtained MFCC with all the parameters (94.44%), MFCC_NHR_DFA (91.67%) and MFCC_DFA (88.89%) while for the SVD base using the parameters MFCC_NHR_DFA (94.19%) and MFCC_NHR (91.86%) we have obtained the best result of recognition rates.

For the normal / pathological classification of MEEI database and Saarbruecken Voice Database, respectively.

Generally, the highest achieved accuracies are 100% for MEEI Voice Database, in the case of using the MFCC_HNR, MFCC_F0, MFCC_DFA and MFCC_F0. While the values found are the highest in the case of using MFCC_HNR is 100% for Saarbruecken Voice Database.

References

[1] J.I. Godino-Llorente, P. Gomez Vilda, N. Saenz-Lechon1, M. Blanco-Velasco, F. Cruz-Roldan, and M. Angel Ferrer-Ballester, “Support Vector Machines Applied to the Detection of Voice Disorders”, Springer-Verlag, Berlin Heidelberg, pp.219-230, 2005.

[2] A.A. Dibazar, T.W. Berger, and S.S. Narayanan, “Pathological Voice Assesment”, In: IEEE 2006 EMBS 2006; New York.

[3] J. Iagnacio Godino-Llorente , Member, IEEE, P. Gomez Vilda , Member, IEEE, M. Blanco-Velasco, Member, IEEE, “Dimensionality Reduction of a Pathological Voice Quality Assesment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters”, In: IEEE 2006 Transactions on Biomedical Engeneering, October 5, 2006.

[4] J. Benesty, M.M. Sondhi, and Y. Huang, “Springer Handbook of Speech Processing”, Springer, Berlin, Germany 2008.

[5] Teixeira, João Paulo, Carla Oliveira, and Carla Lopes. "Vocal acoustic analysis–jitter, shimmer and HNR parameters." Procedia Technology 9: 1112-1122. 2013

[6] Little MA, McSharry PE, Roberts SJ, Costello DA, Moroz IM. “Exploiting Nonlinear recurrence and Fractal scaling properties for voice disorder detection”. Biomed Eng Online. 2007 Jun 26

[7] MEEI Database, Disorder database model 4337. Massachusetts Eye and Ear Infirmary Voice and Speech Lab, Boston, MA. 2002.

[8] W. J. Barry and M. Putzer, “Saarbruecken Voice Database”, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/.

[9] http://htk.eng.cam.ac.uk/

[10] Young, S., G. Evermann, M. Gales, T. Hain and D. Kershaw et al., 2009. The HTK book version 3.4.1. Cambridge University Engineering Department, Cambridge, U.K.

[11] A. Al-Nasheri et al. “Voice Pathology Detection and Classification Using Auto-Correlation and Entropy Features in Different Frequency Regions”. IEEE volume 6, March 9, 2018.

[12] Anusuya MA, Katti SK. “Front end analysis of speech recognition”: a review. Int J Speech Technol,14:99–145. 2010

[13] Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 2010.

[14] Ali Z, Aslam M, Ana María ME. “A speaker identification system using MFCC features with VQ technique”. Proceedings of 3rd IEEE International Symposium on Intelligent Information Technology Application, pp. 115–119, 2009.

[15] Al-nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T. A., Farahat, M., ... & Bencherif, M. A. “An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification”. Journal of Voice, 31(1), 113-e9. 2017.

[16] Al-Nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Malki, K. H., Mesallam, T. A., & Ibrahim, M. F. Voice “pathology detection and classification using auto-correlation and entropy features in different frequency regions”. IEEE Access, 6, 6961-6974. 2018.

[17] Ritchings, R. T., McGillion, M., & Moore, C. J. “Pathological voice quality assessment using artificial neural networks”. Medical engineering & physics, 24(7-8), 561-564. 2002.

[18] Godino-Llorente, J. I., Gomez-Vilda, P., & Blanco-Velasco, M. “Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters”. IEEE transactions on biomedical engineering, 53(10), 1943-1953. 2006.

[19] Daniel Jurafsky & James H. Martin. “Hidden Markov Models” Speech and Language Processing chapter 9. Draft of August 7, 2017.

[20] Abe S. “Support Vector Machines for Pattern Classification”. Berlin, Heidelberg, New York, NY: Springer-Verlag; 2005.

[21] Godino-Llorente JI, Gomes-Vilda P, Blanco-Velasco M. “Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters”. IEEE Trans Biomed Eng. 53:1943–1953. 2006.

[22] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, ‘‘Voice pathology detection on the Saarbrücken voice database with calibration and fusion of scores using multifocal toolkit, in Advances in Speech and Language Technologies for Iberian Languages”. Berlin, Germany: Springer, pp. 99–109 , 2012

Developing a Model for Predicting the Speech Intelligibility of South Korean Children with Cochlear Implantation using a Random Forest Algorithm

Haewon Byeon
Department of Speech Language Pathology
Honam University
Gwangju, Republic of Korea

***Abstract*—The random forest technique, a tree-based study model, predicts the results by using random decision trees based on the bootstrap technique. Therefore, it has a high prediction power and fewer errors, which are advantages of this method. This study aimed to provide baseline data regarding the language therapy after cochlear implantation by identifying the factors associated with the speech intelligibility of children with cochlear implantation. This study evaluates the factors associated with the articulation accuracy of children with cochlear implantation. This study targeted 82 hearing-impaired children, who lived in Seoul, Incheon, and Suwon areas, were between 4 and 8 years old, and had been worn cochlear implant at least one year and less than five years. Explanatory variables used in this study included gender, age, household income, the wear time of a cochlear implant, vocabulary index, and corrected hearing. Speech intelligibility was analyzed using the 'speech intelligibility test tool' composed of nine sentences. The predictive model for speech intelligibility of children with cochlear implants was developed using random forest. The major predictors of the articulation accuracy of children with cochlear implantation were the wear time of a cochlear implant, the time since cochlear implantation, vocabulary, household income, age, and gender, in the order of the magnitude. The final error rate of the random forest model developed by generating 500 bootstrap samples was 0.22, and the prediction rate was 78.8%. The results of this study on a prediction model suggested that it would be necessary to implement cochlear implantation and to develop a customized aural rehabilitation program considering the linguistic ability of a subject for enhancing the speech intelligibility of a child with cochlear implantation.**

Keywords*—*Random forest; hearing impairment; vocabulary index; speech intelligibility; risk factor; data mining

I. INTRODUCTION

The number of patients with hearing loss in South Korea is continuously increasing. Health Insurance Review & Assessment Service (2017)[1] reported that the number of patients who visited a doctor's office due to hearing loss increased by 18.7% in five years (from 746,499 in 2012 to 886,091 in 2016). Considering that a large number of people with hearing problems do not visit a doctor's office, the number of potential patients with hearing loss is expected to be higher than the estimate. National Survey of Ministry of Health and Welfare (2016)[2] reported that one out of two people with hearing loss did not visit a doctor until it prohibits them from living normal lives and only 12.6% of people who were diagnosed with hearing loss wore a hearing aid. It is expected that the number of patients with hearing loss will increase steadily because the noise environment is expanding to the advancing industrialization. The more population with hearing loss is a problem affecting various domains of a community and a nation, such as the loss of workforce and the support of special education.

As the effectiveness of cochlear implantation has been proven, a wider range of patients has become a target of this surgery. Since the Food and Drug Administration (FDA) approved cochlear implantation in 1990, the number of cochlear implant recipients has been increasing worldwide [3]. The cochlear implantation has been covered by the national insurance system in Korea since 2005, so patients are only responsible for 20% of testing and surgical cost. Consequently, the size of the market is growing rapidly.

Many studies have proven the effectiveness of cochlear implant. The cochlear implant is expected to increase the speech perceptivity through biaural gain, noise signal discrimination under a noise environment, sound localization identification, head shadow effect, biaural squelch, and biaural summation [4-7]. The cochlear implantation has been introduced worldwide more and more as more advantages (e.g., the increase of speech discrimination in noise, improvement of sound quality, improvement of signal-to-noise ration, enhancement of equilibrium sensation, suppression of tinnitus, and cost-effectiveness of cochlear implants) of it are known [8,9]. Generally, the target patients of cochlear implantation are when patients are from 12 months to 17 years old, sensorineural hearing loss for both ears above 90dBHL and, when patients are equal to or older than 18 years, hearings loss for both years above 70dB, who do not show advancement in hearing ability and linguistic ability even after wearing a hearing aid long than 3 months [9].

Numerous studies have already reported that hearing-impaired children with cochlear implantation had better voice perceptivity than hearing-impaired children with hearing aids [10,11]. These studies suggested that children with cochlear implantation had better speech perceptivity than those with hearing aids because the cochlear implant is connected directly to the cochlear hair cell in the labyrinth to compensate the severe hearing loss effectively, unlike the hearing aid that only amplifies sound [12].

Numerous studies have already reported that hearing-impaired children with cochlear implantation had better voice perceptivity than hearing-impaired children with hearing aids [10,11]. These studies suggested that children with cochlear implantation had better speech perceptivity than those with hearing aids because the cochlear implant is connected directly to the cochlear hair cell in the labyrinth to compensate the severe hearing loss effectively, unlike the hearing aid that only amplifies sound [12].

The speech intelligibility of hearing-impairment children can be effectively improved if auditory compensation is made in the early stage [13]. However, many children with hearing impairment receive distorted auditory feedback due to the acoustic limitations of hearing aids, which cannot amplify sounds above 4000 Hz, resulting in the production of misarticulation [14]. On the other hand, the cochlear implant can deliver a wide frequency band from 100 Hz to 8000 Hz without distortion, unlike the hearing aid, so it is more beneficial in recognizing consonants belonging to the high-frequency band.

It has been reported that the onset of hearing loss, the level of language comprehension, the wear time of a cochlear implant, the time of cochlear implant operation, and the number of cochlear implant electrodes affect the voice perceptivity of children with cochlear implantation [7,9,13,15]. It has been also reported that the time of cochlear implantation, speech perceptivity, speech production ability, hearing ability, and language ability are related to the speech intelligibility [16]. Although it is expected that the factors influencing voice perceptivity are also associated with speech intelligibility, there are not enough studies evaluating the factors related to the speech intelligibility of children with cochlear implantation in South Korea. Furthermore, the previous studies evaluating the predictive variables of the speech intelligibility of children with cochlear implantation were conducted using linear regression models [17-19]. The linear regression model is a statistical technique that estimates the effects of independent variables on a dependent variable. Since this technique assumes the linear relationship between independent variables and a dependent variable, it does not have a good predictive power when the data have non-linear attributes.

Recently, the data mining technology has been advanced greatly and it is very useful to extract information from big data. Owing to the advancement, this technique has been used in various fields such as marketing, education, and healthcare [20, 21, 22, 23]. Among them, the random forest technique, a tree-based study model, predicts the results by using random decision trees based on the bootstrap technique. Therefore, it has a high prediction power and fewer errors, which are advantages of this method. This study aimed to provide baseline data regarding the language therapy after cochlear implantation by identifying the factors associated with the speech intelligibility of children with cochlear implantation.

II. Materials and Methods

A. Subjects

It is a descriptive study evaluating the factors associated with the articulation accuracy of children with cochlear implantation. This study targeted 82 hearing-impaired children, who lived in Seoul, Incheon, and Suwon areas, were between 4 and 8 years old, and had been worn cochlear implant at least one year and less than five years. The subject selection criteria were as follows. The subjects of this study were hearing-impaired children (1) who had been wearing cochlear implant at least one year, (2) who had received aural rehabilitation regularly after receiving the surgery, and (3) who used oral speech in conversation. Children accompanying other disorders such as visual impairment, cognitive impairment, emotional impairment, and intellectual impairment were excluded from the study. The appropriate sample size was estimated as 80 people by using G-Power 3.1 program when there are seven predictive variables at alpha=0.05, power (1-B) = 0.8, and effect size (f2) of 0.2. Therefore, the sample size of this study met the appropriate sample size for conducting statistical tests (Figure 1) (Figure 2).

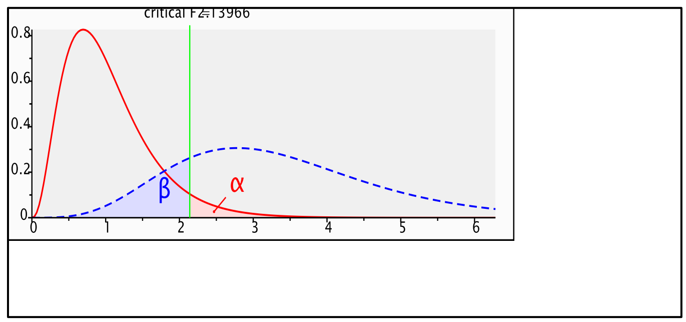


Fig. 1. Power and sample size calculations

B. Explanatory Variable

Explanatory variables used in this study included gender, age, household income, the wear time of a cochlear implant, vocabulary index, and corrected hearing. The corrected hearing was defined as the mean threshold decibel (dB) of hearing tests measured at 250 Hz, 500 Hz, 1 kHz, 2 kHz, and 4 kHz ranges after wearing a cochlear implant. When both ears had cochlear implants, a mean threshold value was used. On the other hand, when only one ear wore a cochlear implant and the other ear used a hearing aid, only the hearing of the cochlear implant side was used.

C. Speech Intelligibility

Speech intelligibility was analyzed using the 'speech intelligibility test tool' composed of nine sentences and developed by Yoon et al. (2005) [24]. This tool includes eight consonant oriented sentences (i.e., a nasal sound, a liquid consonant, a bilabial plosive, an alveolar plosive, a velar plosive, an alveolar fricative, a glottal fricative, a palatal affricate) and one vowel oriented sentence and all sentences are composed of two phrases. The test items are shown in Table 1.

The number of collected words and the number of correctly pronounced words were scored as a percentage (%). All test words produced by subjects were recorded with ESONIC MQ-

94 (4GB) voice recorder. In order to evaluate the consistency between examiners, two graduate students majoring in speech pathology evaluated the speech intelligibility after listening to all recordings of the speech intelligibility tests. The consistency between the two examiners was 91%. The estimated speech intelligibility was transformed into quartiles: the 1st, 2nd, and 3rd quartiles were defined as low-rank groups of speech intelligibility and the 4th quartile was defined as a high-rank group.

F tests – Linear multiple regression: Fixed model, R^2 deviation from zero

Analysis: A priori: Compute required sample size

Input: Effect size f^2 = 0.2

α err prob = 0.05

Power (1–β err prob) = 0.8

Number of predictors = 7

Output: Noncentrality parameter ë = 16.0000000

Critical F = 2.1396555

Numerator df = 7

Denominator df = 72

Total sample size = 80

Actual power = 0.8061255

Fig. 2. Estimation of Appropriate Sample Size

D. Vocabulary Index

Receptive vocabulary and expressive vocabulary were measured using the receptive and expressive vocabulary test (REVT)[25]. The age equivalent-month of receptive vocabulary capability and expressive vocabulary capability was calculated from the test results and then it was converted to the vocabulary index. Equation (1) is the vocabulary index formula.

Vocabulary Index = Receptive Vocabulary Age Equivalent-Month + Expressive Vocabulary Age Equivalent-Month / 2 (1)

TABLE I. THE TEST ITEMS OF SPEECH INTELLIGIBILITY TEST TOOL

No	Target phoneme	Test sentence
1	Nasal	noran ɲyaŋmarine.
2	Liquids	norɛrɯl pullʌyo.
3	Bilabial plosive	pʰirirɯl punda.
4	Alveolar plosive	tʰadzoga taʎʎinda.
5	Velar plosive	kʰok̀ʼirinɯn kʰɯda.
6	Palatal affricate	tsadoŋtsʰa undzʌnhɛyo.
7	Alveolar fricative	sʰasʰɯmi sʰumʌyo.
8	Glottal fricative	horaŋiga ʌhɯŋhɛyo.
9	Vowel	agiga urʌyo.

III. ANALYSIS METHODS

A. Bagging Algorithm

When the variability of the classifier is large while the data is changed even a little, a classifier can be acquired by using a bootstrap method to reduce the variability of a predictor. This method is called a bagging (Bootstrap Aggregating), which is an ensemble algorithm using the bootstrap method [26].

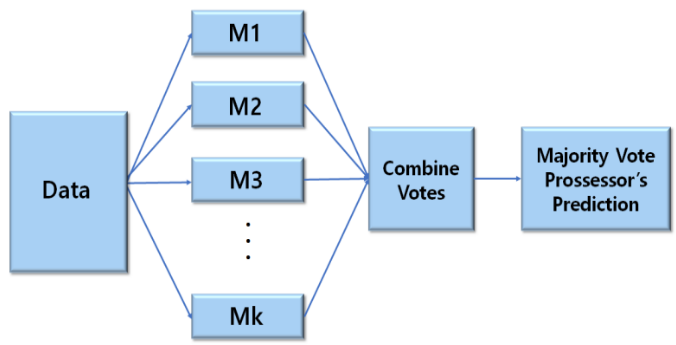


Fig. 3. Bootstrap Aggregating algorithm

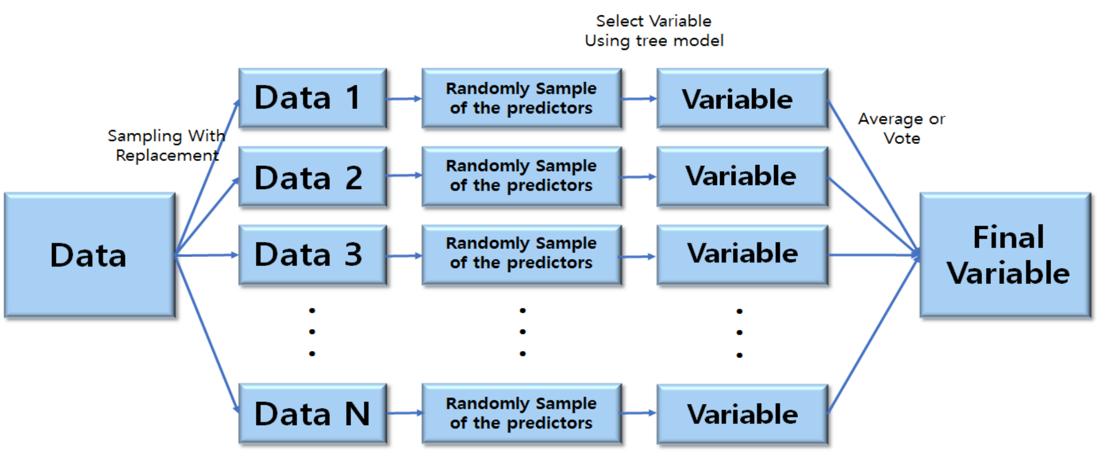


Fig. 4. Random Forest algorithm

The bagging algorithm process generates the bootstrap dataset by sampling with replacement from the training dataset, which is extracted from the population (Figure 3). N bootstrap datasets are generated by repeating this step N times and N classifier groups were obtained after forming each single decision tree by applying an appropriate classification algorithm (e.g., decision tree) to each bootstrap dataset [26]. There are several ways to combine these single classifiers: the mean is used when the response variable is continuous, while a vote is used when it is categorical [27-29].

If a training dataset is unstable, the classification performance is improved by combining bagging classifiers [30, 31]. However, if a training dataset is stable, the bagging-based decision tree obtained from the bagging process is similar to the single decision tree obtained from the training dataset [32, 33]. The bagging is effective in increasing the classification accuracy and improving the stability. Moreover, it has the advantages of reducing the variance and preventing the over fitting [34].

B. Random Forest Model

The core of the bagging method is to predict the classification algorithm results of the bootstrap dataset by averaging or voting the bootstrap dataset. The random forest is an algorithm that adds a random additional layer to this bagging algorithm [32].

Unlike the decision tree expressing each node as a partition providing the most optimal results by using all variables, the random forest uses a method providing the most optimal results among the selected explanatory variable groups by randomly selecting explanatory variables for expressing each node (Fig. 4).

The random forest algorithm process generates bootstrap datasets by sampling with replacement from the training dataset extracted from the population. N bootstrap datasets were generated by repeating this process N times. Moreover, when applying the decision tree algorithm, m explanatory variables are randomly selected from each node. By using these selected variables, this study obtained a group composed of N single classifiers after identifying the most optimal split combination. If a response variable is continuous, a single classifier is combined using a mean. If a response variable is categorical, a single classifier is combined using a vote, just the same as the bagging algorithm.

One of the advantages of the random forest method is that it has less variance than the bagging algorithm because it reduces the correlation between trees. Moreover, it provides more accurate results than other algorithms and is useful to identify an important variable from a large data because it can utilize thousands of independent variables without removing a variable. Especially, it is known that it provides a similar or better predictive power than bagging or boosting when there are many input variables [35]. The analysis is performed by using R version 3.4.3. Figure 5 presents the Random Forest algorithm source of R program.

IV. Results

A. General Characteristics of Subjects

The general characteristics of the whole subjects showed that the mean age was 6.3 years (standard deviation (SD)=3.1), and 32% of them were females (n=26) and 68% of them were males (n=56). The mean vocabulary index was 41.2 (SD=23.6). The mean wear time of cochlear implant was 38 months (SD=8.7) and the mean corrected hearing was 33.8 dB (SD=4.9).

```
Usage
## S3 method for class 'formula'
randomForest(formula, data=NULL, ..., subset, na.action=na.fail)


## Default S3 method:
randomForest(x, y=NULL, xtest=NULL, ytest=NULL, ntree=500,
mtry=if (!is.null(y) && !is.factor(y))
max(floor(ncol(x)/3), 1) else floor(sqrt(ncol(x))),
replace=TRUE, classwt=NULL, cutoff, strata,
sampsize = if (replace) nrow(x) else ceiling(.632*nrow(x)),
nodesize = if (!is.null(y) && !is.factor(y)) 5 else 1,
maxnodes = NULL,
importance=FALSE, localImp=FALSE, nPerm=1,
proximity, oob.prox=proximity,
norm.votes=TRUE, do.trace=FALSE,
keep.forest=!is.null(y) && is.null(xtest), corr.bias=FALSE,
keep.inbag=FALSE, ...)


## S3 method for class 'randomForest'
print(x, ...)
```

Fig. 5. Random Forest algorithm source of R program

B. The Potential Factors related to the Articulation Accuracy of Children with Cochlear Implantation

The general characteristics and potential factors of the subjects according to the articulation accuracy are shown in Table 1. The articulation accuracy of the top group was 19 people out of 82 people (23.1%). The results of the chi-square test showed that the high-rank group and the low-rank group were significantly different in gender, age, the wear time of a cochlear implant, and vocabulary index ($p<0.05$).

C. Major Predictors in the Random Forest Model

Table 3 shows the 'significance of variables' for the articulation accuracy of the children with cochlear implantation based on the random forest. The major predictors of the articulation accuracy of children with cochlear implantation were the wear time of a cochlear implant, the time since cochlear implantation, vocabulary index, household income, age, and gender, in the order of the magnitude. In other words, the wear time of a cochlear implant was the most important predictor of the articulation accuracy. The final error rate of the random forest model developed by generating 500 bootstrap samples was 0.22, and the prediction rate was 78.8%.

TABLE II. POTENTIAL FACTORS RELATED TO THE ARTICULATION ACCURACY OF CHILDREN WITH COCHLEAR IMPLANTATION, %

Characteristics	**Articulation accuracy**		**p**
	High-rank group (n=19)	**Low-rank groups (n=63)**	
Sex, n (%)			
Male	11 (19.6)	45 (80.4)	<0.001
Female	8 (30.7)	18 (69.3)	
Household income (1000 KRW), mean±SD	3,850±1,230	3,370±1,750	<0.001
Age, mean±SD	5.7±2.8	4.6±1.3	0.002
Wear time of a cochlear implant (month), mean±SD	42.5±9.3	33.8±8.6	<0.001
Time since cochlear implantation	45.1±10.1	35.3±9.2	<0.001
Vocabulary index, mean±SD	47.8±28.3	36.1±20.7	<0.001
Corrected hearing (dB), mean±SD	33.2±5.2	34.5±4.8	0.850

TABLE III. THE SIGNIFICANCE OF VARIABLES' FOR THE ARTICULATION ACCURACY OF THE CHILDREN WITH COCHLEAR IMPLANTATION BASED ON THE RANDOM FOREST

Wear time of a cochlear implant	44.580
Time since cochlear implantation	39.905
Vocabulary index	35.734
Household income	23.108
Age	15.605
Sex	15.334

V. DISCUSSION

Recently, the healthcare field tries to provide the foundation necessary for preventing the spread of a disease by identifying the source of it, providing a personalized treatment, predicting the future, and identifying a risk factor by applying the big data technology. This study developed a prediction model for predicting the articulation of school children with cochlear implantation by using a random forest, which is an ensemble based machine learning algorithm. This study established a model for predicting articulation production with considering multiple explanatory variables. It was found that the most important factor was the wear time of a cochlear implant, followed by the time since a cochlear implant operation and receptive vocabulary.

The age at the time of cochlear implantation is known to be an important factor affecting language development of hearing-impaired children. Numerous studies have shown that hearing-impaired children improved their speech production ability after they received cochlear implantation [36-38]. Moreover, the age at the time of implantation, the wear time of a cochlear implant, the remaining hearing before cochlear implantation were reported as factors affecting the ability to produce speech after cochlear implantation [12]. These studies revealed that children who received cochlear implantation before age 5 had better speech intelligibility than those after 5 years old [36]. Moreover, the age at the time of cochlear implantation had the highest effects on speech intelligibility than other factors and the longer wear time of a cochlear implant showed better speech intelligibility [37]. Additionally, even for a study targeting children under 5 years old, subjects had faster language acquisition when they received the implantation earlier [38]. On the other hand, subjects who received the implantation after adolescence had poorer auditory performance than those who received it in childhood [12].

Vocabulary index is another major factor for predicting the improvement in communication ability after cochlear implantation. The linguistic ability has been reported as a significant variable for predicting the speech intelligibility along with the degree of hearing loss and age [38]. Particularly, the naming test and the vocabulary test were significantly correlated with the speech intelligibility of hearing-impaired children [39].

In summary, the results of this study suggested that the time of cochlear implantation and the wear time of a cochlear implant were more important factors than gender, household income, and age for predicting the speech intelligibility of children with cochlear implantation. Furthermore, the speech intelligibility after cochlear implant implies that linguistic ability such as receptive vocabulary is important in addition to the wear time of a cochlear implant.

The model, based on the random forest technique and developed for predicting the speech intelligibility of children with cochlear implantation, had better prediction power than the existing prediction models based on a regression model [24]. Moreover, it was reported that it provided more stable results because it is based on an ensemble algorithm, in addition to a good prediction performance [27, 40]. Therefore, it is believed that to use a random forest model would be more effective for estimating the importance of a variable among many variables than to use a regression model.

VI. CONCLUSION

The results of this study on a prediction model suggested that it would be necessary to implement cochlear implantation and to develop a customized aural rehabilitation program considering the linguistic ability of a subject for enhancing the speech intelligibility of a child with cochlear implantation.

REFERENCES

[1] Health Insurance Review & Assessment Service, Statistical index of Health Insurance Review & Assessment Service, Health Insurance Review & Assessment Service, Seoul, 2017.

[2] Korea Centers for Disease Control and Prevention, Korean National Health and Nutrition Examination Survey, Korea Centers for Disease Control and Prevention, Sejong, 2017.

[3] F. R. Lin and J. K. Niparko, Measuring health-related quality of life after pediatric cochlear implantation: a systematic review. International Journal of Pediatric Otorhinolaryngology, vol. 70, no. 10, pp. 1695-1706, 2006.

[4] P. Schleich, P. Nopp and P. D'haese, Head shadow, squelch, and summation effects in bilateral users of the MED-EL COMBI 40/40+ cochlear implant. Ear and hearing, vol. 25, no. 3, pp. 197-204, 2004.

[5] T. Y. Ching, P. Incerti and M. Hill, Binaural benefits for adults who use hearing aids and cochlear implants in opposite ears. Ear and hearing, vol. 25, no. 1, pp. 9-21, 2004.

[6] C. Morera, M. Manrique, A. Ramos, L. Garcia-Ibanez, L. Cavalle, A. Huarte, C. Castillo and E. Estrada, Advantages of binaural hearing provided through bimodal stimulation via a cochlear implant and a conventional hearing aid: a 6-month comparative study. Acta Oto-Laryngologica, vol. 125, no. 6, pp. 596-606, 2005.

[7] K. Vermeire and P. Van de Heyning, Binaural hearing after cochlear implantation in subjects with unilateral sensorineural deafness and tinnitus. Audiology and Neurotology, vol. 14, no. 3, pp. 163-171, 2009.

[8] R. Laszig, A. Aschendorff, M. Stecker, J. Müller-Deile, S. Maune, N. Dillier, B. Weber, M. Hey, K. Begall, T. Lenarz, R. D. Battmer, M, Bohm, T. Steffens, J. Strutz, T. Linder, R. Probst, J. Allum, M. Westhoten and W. Doering, Benefits of bilateral electrical stimulation with the nucleus cochlear implant in adults: 6-month tt results. Otology & Neurotology, vol. 25, no. 6, pp. 958-968, 2004.

[9] M. Bond, S. Mealing, R. Anderson, J. Elston, G. Weiner, R. S. Taylor, M. Hoyle, Z, Liu, A. Price and K. Stein, The effectiveness and cost-effectiveness of cochlear implants for severe to profound deafness in children and adults: a systematic review and economic model. Health Technology Assessment, vol. 13, no. 44, pp. 1-330. 2009.

[10] L. Spencer, N, Tye-Murray and B. Tomblin, The production of english inflectional morphology, speech production and listening performance in children with cochlear implants. Ear and Hearing, vol. 19, no. 4, pp. 310-318, 1998.

[11] N. Tye-Murray, L. Spencer, E. Bedia and G. Woodworth, Differences in children's sound production when speaking with a cochlear implant turned on and turned off. Journal of Speech and Hearing Research, vol. 39, no. 3, pp. 604-610, 1996.

[12] E. D. Thoutenhoofd, S. M. Archbold, S. Gregory, M. E, Lutman, T. P. Nikolopoulos and T. H. Sach, Paediatric cochlear implantation: Evaluating outcomes, Wiley Blackwell, Virginia, 2005.

[13] R. Y. Litovsky, A. Parkinson and J. Arcaroli, Spatial hearing and speech intelligibility in bilateral cochlear implant users. Ear and hearing, vol. 30, no. 4, pp. 419-431, 2009.

[14] T. A. Serry and P. J. Blamey, A 4 year investigation into phonetic inventory development in young cochlear implant users. Journal of Speech, Language, and Hearing Research, vol. 42, no. 1, pp. 141-154, 1999.

[15] M. A. Svirsky, A. M. Robbins, K, I. Kirk, D. B. Pisoni and R. T. Miyamoto, R, Language development in profoundly deaf children with cochlear implants. Psychological science, vol. 11, no. 2, pp. 153-158, 2000.

[16] C. Yoshinaga-Itano and A. Sedey, Early speech development of children who are deaf or hard of hearing: interrelationship with language and hearing. The Volta Review, vol. 100, no. 5, pp. 181-211, 2000.

[17] E. A. Tobey, A. E. Geers, C. Brenner, D. Altuna, and G. Gabbert, G, Factors associated with development of speech production skills in children implanted by age five. Ear and hearing, vol. 24, no. 1, pp. 36-45, 2003.

[18] S. H. Ferguson and D. Kewley-Port, Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. The Journal of the Acoustical Society of America, vol. 112, no. 1, pp. 259-271, 2002.

[19] P. J. Blamey, J. Z. Sarant, L. E. Paatsch, J. G. Barry, C. P. Bow, R. J. Wales, M. Wright, C. Psarros, K. Rattigan and R. Tooher, Relationships among speech perception, production, language, hearing loss, and age in children with impaired hearing. Journal of Speech, Language, and Hearing Research, vol. 44, no. 2, pp. 264-285, 2001.

[20] H. Byeon, D. Lee, and S. Cho, Assessment for the model predicting of the cognitive and language ability in the mild dementia by the method of data-mining technique. International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, 2016.

[21] H. Byeon, The factors that affects the experience of discrimination in children in multi-cultural families using QUEST algorithm: focusing on Korean language education. Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, vol. 4, no. 2, pp. 303-312, 2014.

[22] H. Byeon, The prediction model for self-reported voice problem using a decision tree model. Journal of the Korea Academia-Industrial cooperation Society, vol. 14, no. 7, pp. 3368-3373, 2013.

[23] H. Byeon, Chi-Square automatic interaction detection modeling for predicting depression in multicultural female students. International Journal of Advanced Computer Science and Applications, vol. 8, no. 12, pp. 179-183, 2017.

[24] M. S. Yoon, H. S. Sim, S. O. Chang and C. S. Kim, Predictor variables of speech intelligibility after cochlear implant in Korean prelingually deafened children. Communication Sciences & Disorders, vol. 10, no. 3, pp. 57-70, 2005.

[25] Y. T. Kim, G. H. Hong and K. H. Kim, Content and reliability analyses of the Receptive and Expressive Vocabulary Test (REVT). Communication Sciences & Disorders, vol. 14, no. 1, pp. 34-45, 2009.

[26] H. Byeon, A prediction model for mild cognitive impairment using random forests. International Journal of Advanced Computer Science and Applications, vol. 6, no. 12, pp. 8-12, 2015.

[27] H. Byeon, Developing a model to predict the occurrence of the cardio-cerebrovascular disease for the Korean elderly using the random forests algorithm. International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, pp. 494-499, 2018.

[28] A. Singh, M. N. Halgamuge and R. Lakshmiganthan, Impact of different data types on classifier pe-rformance of random forest, naive bayes, and k-nearest neighbors algorithms. International Journal of Advanced Computer Science and Applications, vol. 8, no. 12, pp. 1-10, 2017.

[29] H. Byeon, S. Yu, H. Jin and S. Cho, The development of mild cognitive impairment prediction model based on random forest with focusing on visuospatial function and language ability, Journal of Engineering and Applied Sciences, vol. 13, no. 2, pp. 2885-2888, 2018.

[30] A. Statnikov, L. Wang and C. F. Aliferis, A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC bioinformatics, vol. 9, no. 1, pp. 319, 2008.

[31] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis, Conditional variable importance for random forests. BMC bioinformatics, vol. 9, no. 1, pp. 307, 2008.

[32] L. Breiman, Random forests. Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[33] D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson and J. J. Lawler, Random forests for classification in ecology. Ecology, vol. 88, no. 11, pp. 2783-2792, 2007.

[34] G. Biau, Analysis of a random forests model. Journal of Machine Learning Research, vol. 13, pp. 1063-1095, 2012.

[35] A. Cutler, D. R. Cutler and J. R. Stevens, Random forests, In Ensemble machine learning, Springer, Boston, MA, 2012.

[36] N. Tye-Murray, L. Spencer, and G. Woodworth, G, Acquisition of speech by children who have prolonged cochlear implant experience. Journal of Speech and Hearing Research, vol. 38, no. 2, pp. 327-337, 1995.

[37] M. Mondain, M. Sillon, A. Vieu, M. Lanvin, F. Reuillard-artieres, E. Tobey and A. Uzeil, Speech perception skills and speech production intelligibility in French children with prelingual deafness and cochlear implants. Archives of Otolaryngology-Head & Neck Surgery, vol. 123, no. 2, pp. 181-184, 1997.

[38] D. M. Houston and R. T. Miyamoto, Effects of early auditory experience on word learning and speech perception in deaf children with cochlear implants: implications for sensitive periods of language development. Otology & Neurotology, vol. 31, no. 8, pp. 1248-1253, 2010.

[39] M. Yoon, Y. Lee, and H. S. Sim, The relationship between speech intelligibility and related factors of speakers in prelingually hearing impaired children using hearing aids. Korean Journal of Communication Disorders, vol. 5, no. 2, pp. 144-158, 2000.

[40] H. Byeon, H. Jin and S. Cho, S. Development of Parkinson's disease dementia prediction model based on verbal memory, visuospatial memory, and executive function. Journal of Medical Imaging and Health Informatics, vol. 7, no. 7, pp. 1517-1521, 2017.

Ranking Method in Group Decision Support to Determine the Regional Prioritized Areas and Leading Sectors using Garrett Score

Heru Ismanto[1]
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science.
Department of Informatics Engineering, Faculty of Engineering, Musamus University, Merauke, Indonesia

Suharto[2], Azhari[3]
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, Yogyakarta, Indonesia

Lincolin Arsyad[4]
Department of Economics, Faculty of Economics and Business, Universitas Gadjah Mada, Yogyakarta, Indonesia

***Abstract*—The main objective of regional development is to achieve equal development in different regions. However, the long duration and complexity of the process may result in the unequal development of some regions. In order to achieve a fair development process for each region, a standard approach must be developed to select a suitable priority area that can support other underdeveloped regions that require attention. One of the approaches taken is to determine the prioritized areas and the leading sectors in the region where the region is expected to be a support for other regions that still need attention and handling on development priorities. This research was conducted to provide a new alternative in the process of determining the prioritized areas, not only by observing the development data, but also involving decision-making components consisting of government and community (including non-governmental organizations and academicians). This study used group decision support approach with the Garrett ranking technique. The results of the research on the determination of the prioritized areas using Garret Score showed that there are 5 of 29 Regencies/Municipalities in Papua Province that can be used as prioritized areas, namely Jayapura Regency, Jayapura Municipality, Mimika, Merauke and Nabire. Then, there are three leading sectors for development, namely agricultural, mining and Industrial and Processing sectors. The test of ranking results was conducted by calculating the Spearman's correlation coefficient of the Garrett ranking results and obtained a coefficient of 0.807 which means that the ranking results are very strong.**

***Keywords*—Priority area; leading sector; garrett; decision group support; spearman correlation**

I. Introduction

The main objective of regional development is to achieve an equal development that improves the standard of living in different regions. However, the long and complex process involved in regional development makes it difficult to achieve equal development. One of the factors responsible for the difficulty in achieving regional development goals is unequal development in different regions (particularly regions with different economic, cultural and geographic conditions). Unequal development can be reduced by selecting a priority area and leading sector that will serve as a support for the development of surrounding areas that are still largely underdeveloped [1]. The priority area is an area that is considered to be developed and well-established with human resources, nature and an adequate regional growth center. The area must also have better infrastructure compared to surrounding areas [2].

There are different approaches that can be used to identify a suitable priority area. However, the standard method of regional analysis used to identify priority areas is the shift-share approach [3], [4]. In order to determine the priority area using this approach, the achievement data from each region is analyzed and the shift-share coefficient value is calculated. However, the use of this approach in the selection of a priority area and leading sector often produces results that do not match the actual facts on the field. There are many instances in which the regions report good development achievements, even though the facts on the field are different.

Therefore, there is a need to develop an approach that does not rely solely on development and achievement data for the selection of a priority area. As a result, the most prominent area that will be chosen. This study developed a group decision support model for the selection of priority areas. This model combines the results of the selection of priority areas based on the development achievement data (shift-share) and a subjective assessment of decision makers that are familiar with the actual conditions on the field. These decision makers worked together to give a valid and reliable assessment of the actual conditions on the field. The most suitable priority area was then chosen in the selection stage, using the Garrett method of ranking. The decision makers that participated in this study include experts (in the field of government and academics) and the general public, including non-governmental organizations. The decision makers were people who really understand the conditions and facts in their respective regions. Thus, the group decision support model is more complex because it involves diverse assessors.

The Garrett technique is often applied in alternative ranking using simple steps. Some studies have used this technique to carry out research such as identification of the damage caused by natural factors on betel leaf products [5], identification of the effects of work stress on factory workers [6], [7], employee talent management [8], identification of constraints in cattle crossing processes [9], analysis of constraints to the cultivation of agricultural products [10], [11], analysis of the efficiency of resources used in the field of banana cultivation [12], factor analysis of passengers in determining the choice of travel using railroad facilities [13], analysis of clinical control constraints [14], document ranking for information retrieval [15], ranking of candidate synonyms in the biomedical field on the concept of text mining [16].

This research is divided into seven sections: the first section is the introduction, the second to the fourth section is the literature review, the fifth section is the stage of research, the sixth section presents the results and discussion while the last section present the conclusions generated from this research.

II. Priority Area

The priority area is an area that has many advantages over other regions. The priority area supports development activities in areas that require special treatment [2]. The determination of the priority area begins with the collection of data related to policies implemented in the region. This data is then used to determine the priority area. The next stage is the analysis of the priority areas in a particular area, region or province. The priority area may be located in more than one region in a province.The final stage is the review of legislation. At this stage, the results generated from the analysis of the priority area are integrated with existing regulatory reviews about the area. In addition, the role of local institutions, the conditions of regional institutions were also examined for the purpose of this study.

III. Leading Sector

Regional development policies are basically government decisions and interventions, both nationally and regionally to support the overall regional development process. This analysis is very important in order to accelerate regional economic growth, increase the supply of employment and reduce poverty in underdeveloped regions and Border areas. All of this is needed to be able to improve the regional and regional development processes to improve the welfare of local communities.

Efforts to achieve the objectives of regional economic development, the main policy that needs to be done is to make every effort possible so that regional development priorities are in accordance with the potential of the regions. This is related to the development potential of each region that varies greatly, so each region must determine the dominant economic sector [17].

IV. Group Decision Support

Group decision support is fast becoming an important in the determination of strategic plans related to the development of specific alternatives. Group decision support can be used to avoid personal conflicts of interest that arise from the decision making process of individuals [18]. The basic concept of group decision support is concerned with fine-tuning the decision making process in the organizational environment by prioritizing the contribution of appointed experts [19].

The concept of group decision support has been used for various purposes such as the needs of human resource management planning [20], management and endurance of emergency conditions in coastal areas [21], also including an applications in the economic field [22].

V. Garrett Method

Garrett method is often used to complete the ranking of an alternative based on the ratings of respondents that are converted into certain ranks [23]. This ranking is done by determining the most significant factor from the respondent's answer. The ranking of alternatives using Garrett method is done by calculating the respondent's data as a factor of the percentage position value using the following equation:

$$Percent\ Position\ (PP) = \frac{100(R_{ij}-0.5)}{N_j} \quad (1)$$

Where, R_{ij} is the value of the i variable given by the respondents to j, while N_j is the number of variables assessed by as many as j respondents. The results of the percentage position are then converted into Garrett Values using the Garrett ranking conversion table. The value of R_{ij} is then multiplied by the Garrett Value to determine the Total Garrett Score. The average Garrett Score is then calculated by dividing the Total Garrett Score by the amount of alternatives. The alternative ranking is done based on the highest average value.

VI. Research Methodology

This research began with the collection of the results of the assessment of three elements of decision makers on 29 Regencies / Cities in the Papua Province. The three elements came from the government, the general public including non-governmental organizations and experts in academics. The three elements were asked to provide subjective assessments. These assessments were based on actual experiences in the field related to the district / city that is a suitable priority area and leading sector in the Papua Province. The assessment was conducted by distributing online questionnaires. The questionnaire had an answer scale of 1 to 4, which showed the level of achievement and development of an area according to the Klassen typology.

The four scales used in this study include the following, Scale 1: The region is advanced and growing rapidly, HIGH PRIORITY. Scale 2: Developed but depressed region, MEDIUM PRIORITY. Scale 3: Region has advanced potential, MEDIUM PRIORITY. Scale 4: The area is relatively lagging, LOW PRIORITY.

Then, the assessment of the regional leading sector was conducted by choosing an answer on a scale of 1 to 9 consists of nine development sectors. The nine sectors are S1=Agriculture, S2=Mining, S3=Processing Industry, S4=Electricity, Gas and Water utility, S5=Construction, S6=Trade, Hotels and Restaurants, S7=Transportation and Communication, S8=Financial, Rental and Business Services, S9=Services. The results of the questionnaires were then analyzed in a decision support model scheme in which Garrett technique was used as analysis tool.

The analysis carried out using the Garrett technique was intended to rank the areas that are eligible to be used as a prioritized area according to the assessment of the three elements that have been determined and to determine the rank of the 9 regional leading sectors of the region. The results of the ranking of prioritized areas and regional leading sectors were then tested for the correlation using Spearman Correlation.

VII. RESULTS AND DISCUSSION

Based on the results of questionnaires, there were fifty-three respondents consisting of three elements of decision makers (government, community (including non-governmental organizations) and academicians). Each assessor was asked to assess the 29 regencies/municipalities in Papua Province which are feasible to be the prioritized areas and regional leading sectors. Subjective assessment at this stage was very possible. The results obtained from the questionnaires are shown in Table 1 and Table 2.

The next stage was to calculate the percentage of the position of each assessment scale which is the level of feasibility of an area to be a prioritized area and the leading sector of the region.

TABLE I. THE RESULTS OF THE ASSESSMENT OF DECISION MAKERS

No.	**Territory**	**Priority Level**			
		1	**2**	**3**	**4**
1	Asmat	3	10	29	11
2	Biak	9	17	27	0
3	Boven Digul	14	6	30	3
4	Deiyai	1	9	24	19
5	Dogiyai	1	14	23	15
6	Intan Jaya	3	19	18	13
7	Kab Jayapura	37	15	1	0
8	Jayawijaya	3	26	19	5
9	Kerom	5	14	25	9
10	Lani Jaya	1	16	16	20
11	Memberamo Raya	2	15	23	13
12	Memberamo Tengah	2	11	23	17
13	Mappi	8	19	17	9
14	Merauke	22	21	10	0
15	Mimika	32	14	7	0
16	Nabire	13	17	23	0
17	Ndunga	1	13	24	15
18	Paniai	2	18	26	7
19	Pegunungan Bintang	3	12	22	16
20	Puncak Jaya	4	12	22	15
21	Puncak	2	14	17	20
22	Sarmi	4	14	29	6
23	Supriori	2	11	32	8
24	Tolikara	3	14	23	13
25	Waropen	3	16	27	7
26	Yahokimo	3	11	18	21
27	Yalimo	1	7	19	26
28	Yapen	2	16	21	14
29	Kota Jayapura	37	14	2	0

TABLE II. THE RESULTS OF ASSESSMENT OF LEADING SECTORS THROUGH QUESTIONNAIRE

No	**Territory**	**Leading Sector**								
		S1	S2	S3	S4	S5	S6	S7	S8	S9
1	Asmat	10	1	6	5	5	7	11	1	7
2	Biak	9	3	13	2	0	14	2	3	7
3	Boven Digul	13	9	13	2	2	0	12	1	1
4	Deiyai	16	5	3	3	5	1	9	2	9
5	Dogiyai	16	8	3	4	7	1	8	1	5
6	Intan Jaya	17	16	5	1	3	3	6	0	2
7	Kab Jayapura	3	4	11	2	2	20	1	8	2
8	Jayawijaya	12	16	3	3	0	3	9	1	6
9	Kerom	30	2	2	4	4	1	4	1	5
10	Lani Jaya	15	9	2	2	5	3	9	3	5
11	Memberamo Raya	17	5	6	6	3	2	8	1	5
12	Memberamo Tengah	16	6	7	6	2	0	10	1	5
13	Mappi	19	3	8	1	3	1	11	3	4
14	Merauke	43	0	2	1	2	2	0	3	0
15	Mimika	3	42	1	3	0	0	2	0	2
16	Nabire	17	4	12	5	3	6	1	1	4
17	Ndunga	9	10	6	4	5	2	8	3	6
18	Paniai	18	8	3	5	5	2	9	1	2
19	Pegunungan Bintang	11	12	4	6	5	3	6	1	5
20	Puncak Jaya	10	16	4	2	5	4	7	1	4
21	Puncak	15	10	3	2	5	3	10	1	4
22	Sarmi	21	2	9	3	4	4	4	2	4
23	Supriori	14	5	7	3	2	4	10	1	7
24	Tolikara	16	5	9	3	3	1	10	1	5
25	Waropen	12	5	7	5	3	3	13	1	4
26	Yahokimo	15	12	6	2	3	2	7	0	6
27	Yalimo	16	9	2	4	5	1	8	1	7
28	Yapen	13	6	2	7	5	4	9	1	6
29	Kota Jayapura	2	4	14	0	3	18	0	10	2

The results of the calculation of the percentage of positions as well as conversions into Garrett Value are shown in Table 2.

Furthermore, the results of the calculation of the percentage of positions are converted into Garrett Value. This Garrett Value then becomes a multiplier component for each result of the assessment carried out by the three decision support elements. The results of the calculation of the percentage of positions as well as conversion into Garret score are shown in Table 3 and Table 4.

In Table 3 and Table 4, it can be seen that the Garret score automatically shows the best ranking sequence based on the percentage position value of each scale. In other cases, the ranking position may differ depending on the value of the questionnaire generated.

TABLE III. THE VALUE OF CONVERSION OF THE POSITION AND GARRET SCORE OF PRIORITIZED AREAS

Ranking Scale of Prioritized Areas	**Percentage Position**	**Garret score**
1	12,5	73
2	37,5	57
3	62,5	44
4	87,5	28

TABLE IV. THE VALUE OF CONVERSION OF THE POSITION AND GARRET SCORE OF LEADING SECTORS

Ranking Scale of Leading Sectors	**Percentage Position**	**Garret score**
S1	5.56	81
S2	16.67	69
S3	27.78	62
S4	38.89	56
S5	50.00	50
S6	61.11	44
S7	72.22	38
S8	83.33	31
S9	94.44	19

TABLE V. THE AVERAGE GARRET SCORE AND FINAL RANK OF PRIORITIZED AREAS

Territory	**Total Garret Score**	**Average Score**	**Ranking**
Kab Jayapura	3.6	67,92	1
Kota Jayapura	3.587	67,68	2
Mimika	3.442	64,94	3
Merauke	3.243	61,19	4
Nabire	2.93	55,28	5
Biak	2.814	53,09	6
Boven Digul	2.768	52,23	7
Jayawijaya	2.677	50,51	8
Mappi	2.667	50,32	9
Sarmi	2.534	47,81	10
Kerom	2.515	47,45	11
Waropen	2.515	47,45	11
Paniai	2.512	47,40	13
Intan Jaya	2.458	46,38	14
Supriori	2.405	45,38	15
Tolikara	2.393	45,15	16
Memberamo Raya	2.377	44,85	17
Yapen	2.374	44,79	18
Asmat	2.373	44,77	19
Puncak Jaya	2.364	44,60	20
Pegunungan Bintang	2.319	43,75	21
Dogiyai	2.303	43,45	22
Ndunga	2.29	43,21	23
Memberamo Tengah	2.261	42,66	24
Puncak	2.252	42,49	25
Lani Jaya	2.249	42,43	26
Yahokimo	2.226	42,00	27
Deiyai	2.174	41,02	28
Yalimo	2.036	38,42	29

TABLE VI. THE AVERAGE GARRET SCORE AND FINAL RANK OF LEADING SECTORS

Leading Sectors	**Total Garret Score**	**Average Score**	**Ranking**
S1=Agriculture	34,668	1,195	1
S2=Mining	16,353	564	2
S3=Industry and Processing	9,989	344	3
S7=Transportation and Communication	7,752	267	4
S4=Electricity, Gas, and Water Utility	5,376	185	5
S6=Trade, Hotels, and Restaurants	5,060	174	6
S5=Construction	4,950	171	7
S9=Services	2,489	86	8
S8=Financial, Rental, and Business Services	1,674	58	9

The next step was to calculate the total Garrett Score by multiplying each value given by the decision maker on the results of the questionnaire (Table 1 and Table 2) by Garrett Value (Table 3 and Table 4), and were then summed up. This treatment applies to the calculation for each region. Then, the average Garret score was calculated by dividing the total Garrett Score in each region by the number of regions being ranked. The final result of Garrett score calculation is region ranking based on the highest average Garrett value. Table 5 shows the average results of Garret score as well as the ranking of the prioritized area, while Table 6 shows the average results of Garret score as well as the ranking for the regional leading sectors.

The results of the Garret ranking showed that Jayapura Regency is in the strongest position to become a prioritized area in the Papua Province. Followed by four other regions; Kab Jayapura, Kota Jayapura, Mimika, Merauke and Nabire. Whereas, getting sorted from ten regions, the other five regions are Biak, Boven Digul, Jayawijaya, Mappi and Sarmi.

The results of leading sectors ranking showed that the Agriculture is the strongest sector to be the leading sector of Regional development in Papua Province. Followed by Mining, Transportation and Communication, and Industry and Processing.

The results of the Garret ranking of prioritized areas and regional leading sectors were then analyzed for correlation of the results of the both ranking. Based on the spearman correlation calculation, it was found that the correlation coefficient is 0.807 with a standard error value of 0.11. This shows that the results of the ranking carried out on both showed a very strong correlation.

VIII. Conclusion

Based on the results of the analysis, it can be seen that there is a very strong correlation between the ranking of the prioritized areas and the leading sectors using the Garrett method. This is indicated by the spearman correlation coefficient which is close to +1 of 0.807. The results of Garret ranking on 29 regencies/municipalities in Papua Province show that there are five main areas that can be used as prioritized areas, namely Jayapura Regency, Jayapura Municipality, Mimika, Merauke and Nabire. While the regional leading sectors are Agriculture, Mining, and Industry and Processing. Thus, Garret ranking can be used for alternative ranking technique especially for group decision support model that combines the results of ranking with other approaches.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments and suggestions that helped to improve the quality and presentation of this paper. This research is supported by Ministry of Research, Technology and Higher Education in the research scheme Doctoral dissertation in 2018.

References

[1] Direktorat Pengembangan Kawasan Khusus dan Tertinggal, Priority Area-Based Regional Economic Development: Building a Model for Management and Development of Related Program Linkages, Paper of Deputy for Regional Autonomy and Regional Development of BAPPENAS, Government of the Republic of Indonesia, access at July, 03,2018, available at https://www.bappenas.go.id/ files/4513/5080/2311/10pengembangan-ekonomi-daerah-berbasis kawasan-abilitas-membangun-model-pengelol-dan-pengembangan-keterkaitanprogram__20081123002641__9.pdf, *in* Bahasa

[2] Pemerintah Propinsi Kalimantan Utara, Priority Area, access at July, 11, 2018, available at http://simtaru.kaltaraprov.go.id/web/ktr/ kawasan-andalan, *in* Bahasa

[3] Yuda, D.K., and Navitas, P., "Economic Development Direction for Lamongan Regency Based on Leading Sectors (Case Study: Agriculture Sector)," *Jurnal Teknik Pomits,* Vol.3, No. 2, 2014, pp. C.136 - C.141, *in* Bahasa

[4] Nazipawati, "Application of Static Models and Dynamic Location Quotients and Shift-Share in Regional Economic Planning: A Case Study of Ogan Komering Ulu District, South Sumatra Province," Jurnal Eko-Regional, Vol. 2 No. 2, 2007, pp. 81 – 86, *in* Bahasa

[5] Mohanasundaram, P., " Marketing Probles Faced by Betel Leaf Cultivators," International Journal of Advanced Research, Volume 3, Issue 5, 2015, pp. 1447-1451

[6] Oyelaran, O.A., Tudunwada, Y.Y., Abidoye, J.K., Sanusi, O.M., "A Study on Impact of Work Stress among Mat Factory Workers in Kano – Nigeria in," *Iranian Journal of Health, Safety & Environment*, Vol.4, No.2, 2015, pp.746-751

[7] Jins, J.P., Radhakrishnan, R., "A Study of Work Stress and Coping Strategies among Tile Factory Workers in Calicut District in Kerala," International Journal of Scientific and Research Publications, Volume 4, Issue 4, 2014, pp. 1 - 9

[8] Suryawanshi, A., and Maharshi, S., "Talent management: Success Mantra for Small and Medium sized Enterprises," Industrial Engineering Letters, Vol.3, No.11, 2013.

[9] Manoharan, R., Selvakumar, K.N., and Pandian, S.S., "Constraints in Milk Production Faced by The Farmers in Pondicherry Union Territority," Indian J. Anim. Res, Vol. 37, No. 1, 2003, pp. 68 – 70

[10] Nirmala, B. and Suhasini, K., " Armer's experience with hybrid rice technology: A case study of Khunti district of Jharkhand State of India," African Journal of Agricultural Research, Vol. 8, No. 29, 2013, pp. 3973 – 3975

[11] Rohit, J., Dubey, S.K., Singh, P., Singh, B.K., and Kumbhare, N.V., "An Assessment of Constraints Faced by the Farmers in Peri-Urban Vegetable Cultivation," International Journal of Current Microbiology and Applied Sciences, Vol. 6, No. 10, 2017, pp. 2245 - 2251

[12] Balaganesh, G., Gautam, Y., Anoop, M., and Singh, H.P., "An Empirical Analysis On Resource Use Efficiency And Constraints In Adoption Of Precision Farming In Banana In Theni District - Tamil Nadu, " Journal Economic Affairs, Vol. 61, No. 3, 2016, pp. 375-380

[13] Vishnuvarthani, S., and Selvaraj, A., "Factors Influencing the Passengers to Select Train Travel: A Study in Salem Division of Southern Railway,*"* International Journal of Scientific and Research Publications, Volume 2, Issue 10, 2012, pp. 1 – 7

[14] Christy, R.J., "Garrett's Ranking Analysis of Various Clinical Bovine Mastitis Control Constraints in Villupuram District of Tamil Nadu," IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), Vol. 7, No. 4, 2014, pp. 62 – 64

[15] Mbarek, R., Tmar, M., Hattab, H., and Boughanem, M., "A Re-Ranking Method Based on Irrelevant Documents in Ad-Hoc Retrieval," 5th Symposium on Languages, Applications and Technologies (SLATE'16), 2016, pp. 2:1 - 2:10

[16] Jagannatha, A.N., Chen, J., and Yu, H., "Mining and Ranking Biomedical Synonym Candidates from Wikipedia, " Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), 2015, pp 142–151

[17] Sjafrizal, Ekonomi Regional, Teori dan Aplikasi. Cetakan Pertama. Padang, Baduose Media 2014, In bahasa

[18] Rigopoulus, G., "A Group Decision Support System for Collaborative Decisions Within Business Intelligence Context," American Journal of Information Science and Computer Engineering, Vol. 1, No. 2, 2015, pp. 84-93

[19] Rayed, C.A., "Developing an Expertise Interaction Meta- Model for Group Decision Support System (GDSS)," Computer Science and Information Technology, Vol. 1, No. 2, 2013, pp. 105-110

[20] Group Support Systems: Tools for HR Yao, J., Wang, J., Xing, R., and Lu, J., "Decision Making," Proceedings of the Third European Academic Workshop on electronic Human Resource Management, Bamberg, Germany, CEUR-WS.org, ISSN 1613-0073, Vol. 570, online: CEUR-WS.org/Vol-570/ , 2010, pp. 400-409.

[21] Zhao, X., Chen, Y., Ku, M., Rich, E., Deegan, M., and Luna-Reyes, L.F., "Group Decision Support Systems for Emergency Management and Resilience: CoastalProtectSIM, " Proceedings of the 50th Hawaii International Conference on System Sciences | 2017, pp. 2489 – 2497

[22] Istudor, I., and Duta, L., "Web-Based Group Decision Support System: an Economic Application, " Informatica Economică , Vol. 14, No. 1/2010, 2010, pp. 191 – 200

[23] Dhanavandan, S., "Application Of Garrett Ranking Technique: Practical Approach," International Journal of Library and Information Studies, Vol. 6, No. 3, 2016, pp. 135 – 140.

A Hybrid Technique for Tunneling Mechanism of IPv6 using Teredo and 6RD to Enhance the Network Performance

Zulfiqar Ali Zardari[1]
Faculty of Information Technology
Beijing University of Technology
Beijing, China

Munwar Ali[2]
Department of Software Engineering. MUET SZAB
Khairpur Mirs', Pakistan

Reehan Ali Shah[3]
Computer System Engineering
University College of Engineering & Technology (UCET)
The Islamia University Bahawalpur

Ladha Hussain Zardari[4]
Department of Information Technology, QUEST,
Nawabshah, Pakistan

***Abstract*—Currently, Internet Protocol version 4 (IPv4) addresses have been depleted. Many Internet Service Providers (ISPs), researchers and end users are migrating from IPv4 to IPv6 due to strong features of IPv6 and limitation in IPv4. Different tunneling techniques have been deployed to migrate on IPv6 for ordinary users. However, these techniques create many issues such as compatibility, complexity, connectivity and traffic. Due to the dissimilar header structure and incompatibility of IPv4 and IPv6, the devices are unable to communicate with each other because devices do not support IPv6 addresses directly. The performance of network is also compromised due to huge increment in data transmission traffic. In this paper, we proposed a technique to provide full IPv6 connectivity and enhancing the network performances by combining two tunneling techniques such as IPv6 Rapid Deployment (6RD) and Teredo. To increase the throughput of the network, jumbo frames are used to carry huge amount of data. The main objective of using both techniques is to provide a hybrid network rendering for full IPv6 connectivity. The proposed technique provides not only the IPv6 but also provides better performance in the network. Simulation results show that throughput and packet delivery ratio achieve maximum gain by 9000 bytes and 98%, respectively.**

Keywords—Hybrid network; IPv4; IPv6; 6RD; teredo tunneling mechanism; network performance jumbo frames

I. Introduction

In February 2011, internet assigned number authority (INNA) declared that there are no IP addresses available on IPv4 pool. After this announcement the shortfalls of IP addresses create problems to provide connectivity for end users. In such condition the only solution is to migrate on IPv6 which contains trillion of trillion IP addresses (128 bits). It is not an easy task to shift from IPv4 to IPv6 quickly but a big challenge of next generation of internet. Many translation, transitions and tunneling techniques are available such as IPv4-In-IPv6 tunneling, Teredo, 6to4, dual stack and other. These methods are planned to extend IPv6 connectivity using IPv4 network [1-4]. But out of these two techniques Teredo and 6RD are widely used by research community. This research focuses on both Teredo and 6RD.Currently, illustrated issue is the connectivity of IPV4 and IPV6 under the same environment. Devices are not upgraded to support IPv6 connectivity and Compatibility. There is a need of hybrid network so that IPv4/IPv6 can communicate across both the network infrastructure. It is difficult for ISPs to control the flow of packets in 6to4 tunnel prefix from customers for this reason 6RD is proposed. The rapid use of social networks and multimedia application create high data transmission load and cause hindering the traffic to raise the problem that how to move this traffic on IPv4 network. The jumbo frames has high capacity and is specially designed to carry heavy load of traffic in the network [5-7].

This paper focuses on deployment of IPv6 with IPv4 network. The proposed work merge Teredo and 6RD together to create a hybrid network. The features of both networks are combined to facilitate the current IPv4 addresses, the issues of 6RD can be covered in Teredo and vice versa. IPv6 have a unique feature of IPsec and huge amount of IPs over IPv4 [8]. Proposed hybrid network provides full connectivity to customers/ISPs using these two techniques. After deployment of hybrid network the performance of the overall network is highly reliable.

The contributions of this proposed work are: (1) an adaptation of an improved hybrid-based IPv6 mechanism is deployed, which enables IPv6 (2) Examination and illustration of IPv6 development system is performed incorporating traffic characteristics tunneling. (3) A solution is provided to handle the increased payload. (4) Achieved results for packet delivery ratio in hybrid network jumbo frame are included. This approach facilitates the Internet service providers (ISPs), companies and vendors in migration of IPv4 and especially for those ISPs having large network across the any province or country.

The rest of the paper is divided into following parts, section 2 describes the related work, section 3 is about problem statement, section 4 is methodology, section 5 covers the simulations and results, and section 6 is about conclusion.

II. Related Work

Ala Hamarsheh et al [9] proposed mechanism for known as D4across6. It consists of deployment of IPv4 locally for IPv6 connectivity to access network. Mechanism allows IPv4 devices to communicate with IPv4-only content providers over an IPv6-only infrastructure. It carries IPv4 packets and routing information of the IPv4 protocol over an IPv6-only network infrastructure. The features of the proposed protocol are: offers IPv4 connectivity across IPv6-only access networks, stateless operation, cost-effective solution, and requires simple and automatic configuration at customers' hosts and no change in network infrastructure is needed. Results are calculated in one parameter throughput which varies from 64 bytes to 1024 bytes with number of nodes 100 to 1500, while other parameter is Mbps which is dynamically changes as number of bytes are increasing. Compared with 4rd, (A+P), DS-Lite, D4across6 and it shows that performance of D4across6 is comparatively is better all other techniques. The main drawback of this is mechanism it can't handle huge amount of traffic load, whenever traffic increases it may takes more time transfer the data packets.

Fatema Siddika et al [10] proposed a solution to migration IPv6 in Bangladesh by applying transition mechanism to provide connectivity to IPv6 hosts. Specific routing protocols are implemented in networks i.e. Border Gateway Protocol (BGP), Open Shortest Path First (OSPF), and OSPFv3 for addressing and connectivity of end nodes. The proposed solution is divided into three phases first, apply a transition mechanism to the IPv4 dominant network to provide IPv6 support and make it an IPv6 dominant network. Second, apply the transition mechanism to the IPv6 dominant network to provide IPv4 support until the remaining IPv4 hosts became IPv6 hosts. Third, the network is completely in IPv6. Prevent different nodes in the network from talking with each other with different IP protocols, if necessary, upgrade the applications or use the proxy at application layer. The results carried out with that dual stack tunneling techniques deployed in this network gains better performance than others.

Shaneel Narayan et al [11] proposed a solution for transferring high traffic load one to end to another. This scheme utilizes jumbo frames to transfer the data packets during communication. This can be achieved by connecting two powerful systems and two network card (windows server) with each other. Compare the performance of jumbo frames to that of normal frames on a test-bed. To do so, two different operating systems were implemented with both IPv4 and IPv6 and it has been shown that jumbo frames undoubtedly give higher throughput than normal frames. When used on networks with IPv6, favorable delay values are attained and packet drop rates are generally better than that on IPv4 implementation. Jumbo frames overall give better network performance metrics than when normal frames are implemented on a network.

Zhao Qin et al [12] proposed seamless converging system for IPv4/IPv6 transition which is based on two parts zone-1 state full and zone-2 stateless protocol modules. It can share to addresses of IPv4 or IPv6 to access IPv4 connectivity across IPv6 network. In zone-1 is designed as Metropolitan Area Network (MAN) in IPV6 addresses and (MAN) is also designed for IPv4 addresses. In zone-2 core routers (CR) are placed of both addresses are connected with state full translator. In zone-2 CR router of IPv4 connected with stateless of zone-1, in this both zone are communicating each other. Service Router (SR) and Broadband Access Server (BAS) of both networks IPv4 and IPv6 are placed to transfer data packets. It supports connectivity of IPv6 quickly, additionally it supports both addresses simultaneously without affecting each other.

S. Aravind et al [13] proposed dual stack and tunneling techniques for migration of IPv6 to transmit data packets of IPV addresses to IPv4 in both scenarios i.e. static and dynamic of the network. In the network five routers and four switches are placed in simulation for transferring data packets between both the networks. Using Graphical Network Simulator-3 (GNS3) network simulator the results are carried in dynamic, static routing network and Open Shortest Path First (OSPF) protocols to transmit IPv6 packets on IPv4 network. Performance comparison shows that low Latency and high throughput in IPv6 with respect to IPv4.

M. F. Suleiman et al [14] has point out the factor, challenges and analyzed the network by deployment of IPv6 address in Nigeria. In first phase the author survey the IPv6 awareness training, planning and deployment of IPv6. After successfully deployment of IPv6 addresses they have expressed the results and finally phase simulations experiments to explains the whole mechanism of dual stack and 6to4 techniques in detail. A real time business case was the discussed to motivate and promote the benefits, cost estimation, risk of IPv6 addresses.

S. Javid et al [15] implemented OSPFv3 in IPv6 addresses to survival of IPs because of various features of OSPFv3 protocol. OSPFv3 is used for large and complex network to provide IPv6 connectivity to end users. The migration of IPv6 is in slow process it may takes long time to fully migrate on IPv6. The reason behind implementation is big organizations are fully depended on internet and they don't want to bear any loss if any time their services can be downtime. So avoid this issue the author has implemented to provide full connectivity to large organizations.

III. Problem Statement

Nowadays IPv6 infrastructure is a necessity of internet service providers. However some challenges, difficulties and delusion are the major concerns during deployment. These concerns are enlisted below:

- ISPs below Tier-3 are not IPv6 supported
- Deployment of IPv6 takes long time
- Replacement of equipment is a big challenge
- Unavailability of IPV6 supported devices on vender end
- Payload increases in the network
- People are irrevocable and reluctant to IPv4 and IPv6 respectively

- At this time migration can be done slowly and gradually, although IPv6 is in practical form but still it takes long time to put into operations. [16, 17, 18]

Keeping above stated reasons in consideration migration can be done slowly and, although IPv6 is in practical form but still it takes long time to put into operations. The transition techniques used for IPv6 connectivity are summarized in Fig. 1.

The following Fig. 1 shows the hierarchy of transition which shows the different layer there are various transition techniques are used for IPv6 connectivity.

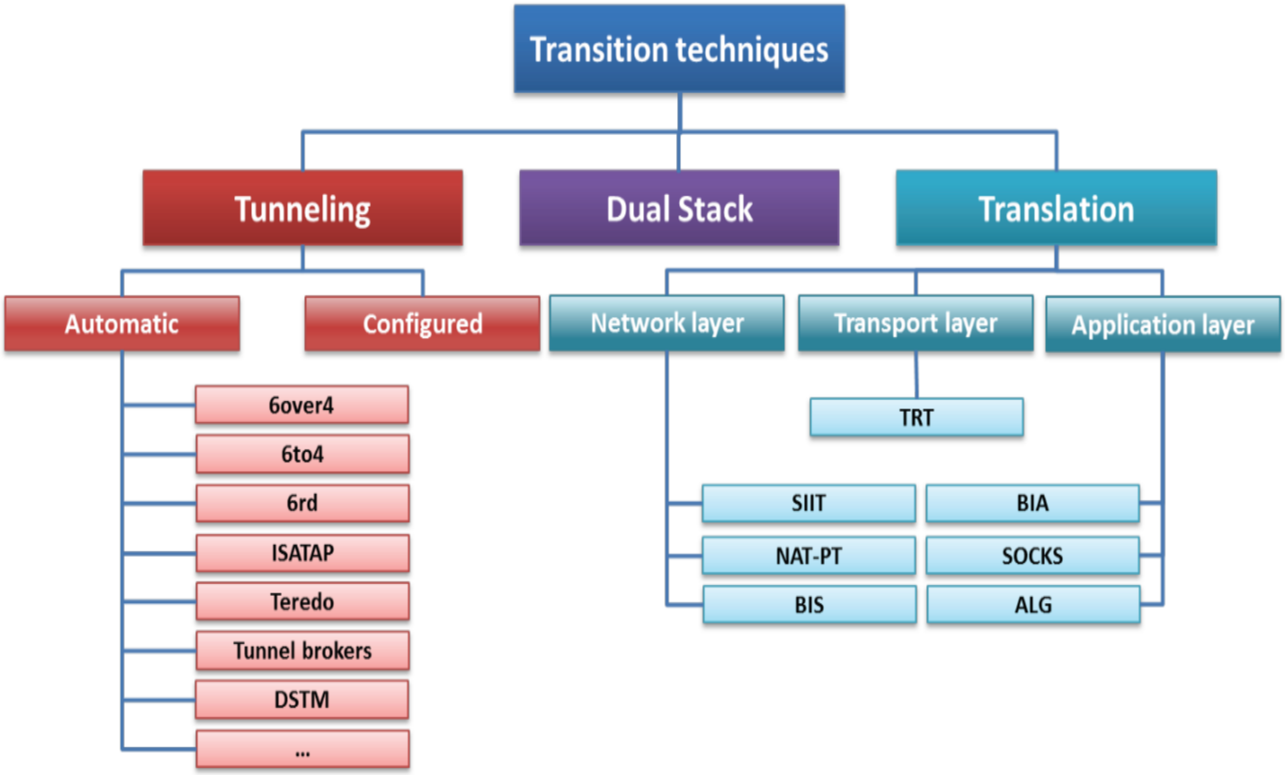


Fig. 1. Summarization of different transition techniques

In Fig. 1 transitions techniques are divided into three main techniques, tunneling, dual stack and translation. In tunneling automatic techniques are widely used for communication between IPv4 and IPv6.Internet service provider (ISP) configures tunneling and translation techniques to providing IPv6 connectivity in real time environment [19].

The research finding and contributions of this work presented as follows:

- Propose a hybrid technique based on 6RD and Teredo, which provides relieve IPv4 addresses shortage.
- The proposed hybrid technique delivers an introduction of IPv6 addresses.
- The proposed technique gives an easy way to migrate from IPv4 to IPv6 addresses
- The hybrid technique is capable of running both addresses on the same network without affecting each other.

IV. Proposed Technique

After study of all mechanisms their advantages and disadvantages in mind we have proposed a technique which is based on hybrid network that is the combination of two techniques which are Teredo & 6RD, used to transfer data from IPv4 to IPv6. The proposed hybrid technique supports both networks IPv4 and IPv6. It can provides IPv4 addresses as well as IPv6 addresses. The hybrid network contains 4 Cisco routers connected with IPv6 hosts, IPv4 backbone network of 6RD & Teredo relay devices. The traffic moves from IPv6 to IPv4 network via these connected devices. Addresses of IPv4 & IPv6 are configured at every router and tunnel is configured on the boarder routers to provide the access of both networks. It is unable to provide the IPv6 connectivity to users without boarder routers because one can't send the addresses of IPv6 directly. When huge amount of data come over the network it might be some interruption in the network. By the use of jumbo frames, throughput can be increased, because it can carry data up to 9000Bytes in the network. The proposed hybrid network a whole traffic of IPv6 is divided into two parts that are local traffic and tunneled traffic. The UPD and TCP protocols are used to encapsulate the tunneled traffic. To provide the connectivity of IPv6 behind the NAT devices is responsible to is responsibility of Teredo. It depends on two main components that are Teredo relay and Teredo Server. Teredo Server is used to initialize the Teredo, while relays are used for routing connectivity of IPv4 and IPv6. The upper level contain IPv6 network and lower level has IPv4 network pool. The BR router and Teredo Server functions like bridge between IPv4 and IPv6 networks. Native traffic routed by IPv6 host is encapsulated in IPV4 by CE router and forwarded towards the BR router or directed to 6rd. The received encapsulated 6rd and BR traffic is de-capsulated by CE router and forwarded to the end user. Server assist client to get IP from IPV6 for the internet access. IPv6 packets are encapsulated in to IPv4 UDP Teredo client address by using Teredo relay and vice versa. The 3FFE:831F::/32 IPv6 suffix is advertised by Teredo relay for internet. Different IPv6 hosts send the packets to Teredo client routed by Teredo relays to share dynamic load over every Teredo client. Fig.2 shows hybrid network technique, the communication link between two terminals 6RD and Teredo techniques. The IPv6 traffic flows from upper section of the network, after that de-capsulated on the edge router to the next endpoint crossing IPv4 network. To ensure the security of data packets thorough tunneled traffic is needs to be de-capsulated.

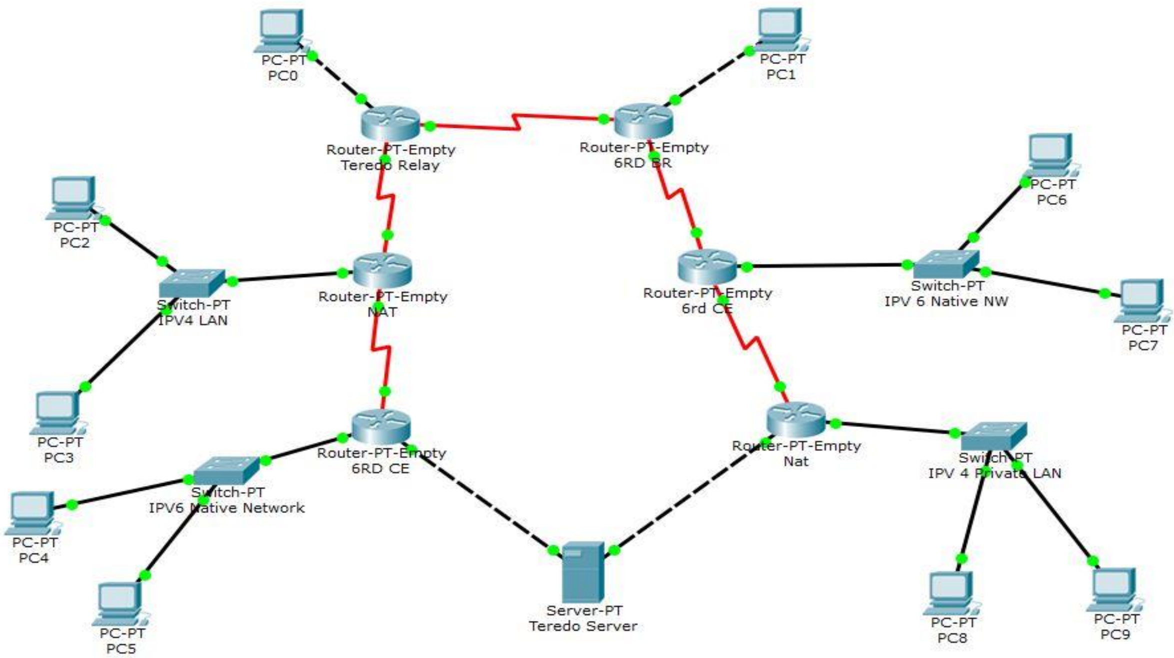


Fig. 2. Hybrid Network topology

V. SIMULATIONS AND RESULT ANALYSIS

In our proposed hybrid technique we have deployment IPv4 and IPv6 addresses successfully. In order to check the performance of the hybrid network tests are conducted amongst IPv4 & IPv6 to show the connectivity of devices results. Simulation work scenario is created using graphical network simulator 3 (GNS3). Results analysis is carried out for evaluation among the performance metrics of these strategies in order to estimate any statistically-important variation between them. The main purpose of this research is to rank the IPv6 transition strategies and categorize the better strategy that offers high throughput, high packet delivery ratio and lowest delay. This section presents the simulation results by comparing different parameters like throughput, packet delivery ratio, average delay, tunneling overhead and jitter. Proposed network demonstrate the bundle conveyance proportion in cross breed network. Hybrid network connected with core of Teredo and 6RD routers along with IPv6 do not require header checksum, flags and identifier therefore, the IPv6 performs faster than other counterparts. Initially the throughput is very low because of less number of packets transferred as the number of packets increases throughput also increased. When the jumbo frames are used in the network it is favorable for increased throughput and packet delivery ratio as compare to 6RD and Teredo technique. The following subsections discusses these analyzed parameters and obtained results in more details. Configuration of GNS3 is easy to deploy because it shows the network graphically.GNS3 allows to access cisco devices virtually, does not required any physical hardware for simulation. The front end of GNS3 dynagen. Dynagen runs dynampis to create simulation environment. Actually GNS3 access cisco IOSs on your windows or linux systems to design network topology virtually without hardware.

A. Throughput

Fig.3 shows the throughput of IPv4 and IPv6 networks. Initially, both networks are sound and safe, however when the number of packets increases they effects throughput directly. Initially both networks i.e. IPv4 and IPv6 are smooth but as soon as packet size increases simultaneously traffic will increased. It was observed that consequent the network to stability due to jumbo frame takes the loads from the network. It is clearly seen that IPv4 and IPv6 provides better performance by the utilization of hybrid network.

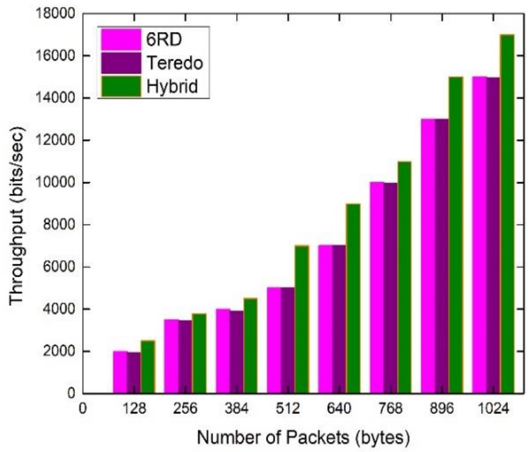


Fig. 3. Throughput of 6RD, Teredo and Hybrid

B. Packet Delivery Ratio (PDR)

Hybrid network shows better performance in the packet delivery ratio, because the insertion of jumbo frames in the network for high payload carried out the network traffic to the destination. It is observed that with the increase of packet size PDR increases. This effect is resulted due to jumbo frames have capability to carry high traffic load to make network performance better Fi. 4 mimics this effect.

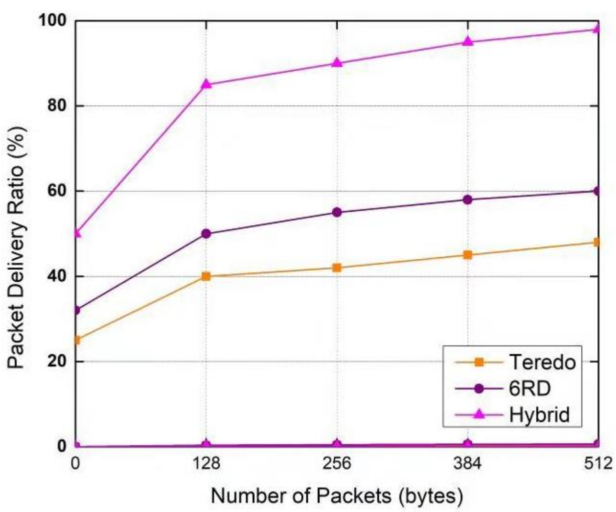


Fig. 4. Packet deliver ratio of 6RD, Teredo and Hybrid

C. Average Delay

Fig.5 shows the average delay with respect to time, IPv6 increases delay at end stages and that effect is carried from many devices. IPv6 shows higher delay as compared to IPv4 from initial time. While passing the traffic from one end to another node, many devices face high payload on every node in the network. Teredo clients devices are located behind the NAT devices and traffic pass through connecting server and encapsulates the incoming the packet and de-capsulate each packet before sending therefore it takes some time.

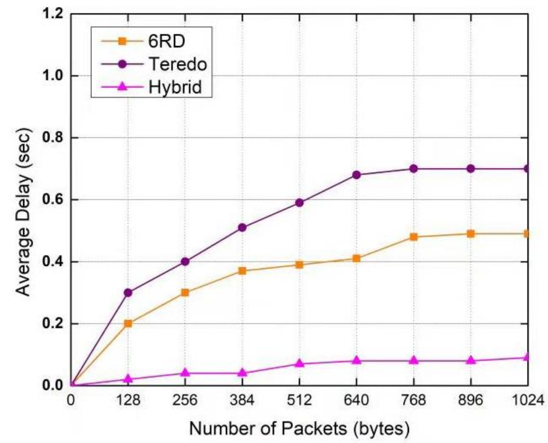


Fig. 5. Average delay of 6RD, Teredo and Hybrid

D. Tunneling Overhead

Fig.6 shows the sum of the overhead caused by engendering tunnels, expunging tunnels, stimulating tunnels, encapsulation, and de-capsulation is called tunneling overhead. Fig. 6 shows the tunneling overhead of 6RD, Teredo and proposed technique. It is measured with respect to packets in bytes. In average overhead it clearly declares that our proposed hybrid technique supersede both the Teredo and 6RD mechanism and hence can be proposed for complex networks to reduce the overhead drastically.

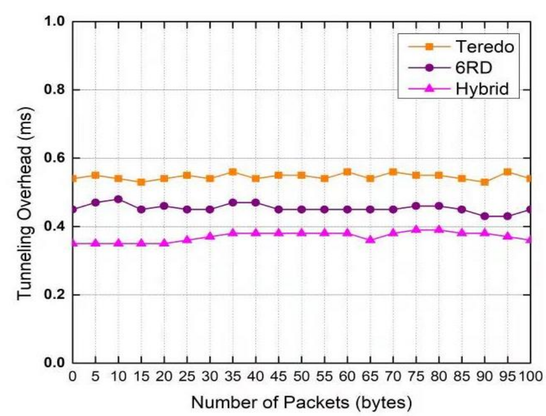


Fig. 6. Tunneling Overhead 6RD, Teredo and Hybrid

VI. Conclusion

This paper provides two main contributions, connectivity of IPv6 to end users and transfer a huge amount of data packets (high payload). Two testbeds are used i.e. 6RD and Teredo to performance tunneling mechanism and to develop a hybrid network. The proposed hybrid network contains characteristic of both techniques. It has been analyzed that our proposed hybrid network has better performance than individual techniques. 6RD & Teredo failed to provide IPv6 connectivity to the individual user. The main advantage of a hybrid network is it can carry out both networks IPv4 and IPv6 together. Both networks performance quite satisfactory and operate easily without any disturbances. Additionally, huge traffic can be transferred across the network with jumbo frame technique.

The proposed hybrid network is compared with 6RD, and Teredo in terms of throughput, packet delivery ratio, average delay and tunneling overhead. From results, it has been observed that the hybrid network has better performance than 6RD and Teredo in term of packet delivery ratio up to 98%, whereas minimum delay occurs at high payload. Collectively, the hybrid network gives a huge improvement in different metrics. For future work, this research can be extended in the perspective of security issues and different security algorithms can be developed to prevent the hybrid network from different attacks.

References

[1] R. Tadayoni, and A. Henten, 'From IPv4 to IPv6: Lost in translation?', In Telematics and Informatics, vol. 33, year 2016 Issue 2, pp 650-659.

[2] L. Smith, M. Jacobi and S. Al-Khayatt, "Evaluation of IPv6 transition mechanisms using QoS service policies", 2018 11th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP), Budapest, Hungary, 2018, pp. 1-5.

[3] M. Nikkhah, "Maintaining the progress of IPv6 adoption", In Computer Networks, vol 102, 2016, pp 50-69.

[4] N. Zhang, M. A. Sirbu and J. M. Peha, "A comparison of migration and multihoming support in IPv6 and XIA", 2017 International Symposium on Networks, Computers and Communications (ISNCC), Marrakech, 2017, pp. 1-8.

[5] Kamaldeep, M. Malik and M. Dutta, "Implementation of single-packet hybrid IP traceback for IPv4 and IPv6 networks", IET Information Security, vol. 12, no. 1,year 2018 pp. 1-6.

[6] D. R. Al-Ani, A. R. Al-Ani, "The Performance of IPv4 and IPv6 in Terms of Routing Protocols using GNS 3 Simulator", Procedia Computer Science, Vol 130, 2018, Pages 1051-1056.

[7] J. Beeharry and B. Nowbutsing, "Forecasting IPv4 exhaustion and IPv6 migration", 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balaclava, 2016, pp. 336-340.

[8] Y. Sookun, and V. Bassoo, "Performance analysis of IPv4/IPv6 transition techniques", 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balaclava, year 2016 pp. 188-193.

[9] Hamarsheh, and Ala, "Deploying IPv4-only Connectivity across Local IPv6-only Access Networks", IETE Technical Review Taylor & Francis year 2018, pp. 1-14.

[10] F. Siddika, and M. A. Hossen , and S. Saha, "Transition from IPv4 to IPv6 in Bangladesh: The competent and enhanced way to follow", 2017 International Conference on Networking, Systems and Security (NSysS), Dhaka, year 2017, pp. 174-179.

[11] S. Narayan and P. R. Lutui, "Network Performance Evaluation of Jumbo Frames on a Network", 6th International Conference on Emerging Trends in Engineering and Technology, Nagpur, year 2013, pp. 69-72.

[12] Z. Qin, "Seamless converging system for IPv4/IPv6 transition", 2017 9th International Conference on Advanced Infocomm Technology (ICAIT), Chengdu, 2017, pp. 110-113.

[13] S. Aravind and G. Padmavathi, "Migration to IPv6 from IPV4 by dual stack and tunneling techniques", International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai, year 2015 , pp. 107-111.

[14] M. F. Suleiman and J. Cordry, "Analysis of organizations IPv6 deployment strategies in Nigeria and evaluating suitable transition mechanisms," *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, 2017, pp. 695-704.

[15] S. R. Javid and S. K. Dubey, "Implementing OSPFv3 in IPv6 network," *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Noida, 2016, pp. 580-584.

[16] C. Udeagha, R. Martin, D. Peck, A. Youton, A. Marshall and J. Clarke, "Migrating from IPV4 to IPV6 in Jamaica", SoutheastCon 2018, St. Petersburg, FL, USA, 2018, pp. 1-8.

[17] D. Chauhan, and S. Sharma, "Performance Evaluation of Different Routing Protocols in IPv4 and IPv6 Networks on the Basis of Packet Sizes", In Procedia Computer Science, Vol 46, year 2015, pp. 1072-1078.

[18] G. Lencse, and Y. Kadobayashi, "Methodology for the identification of potential security issues of different IPv6 transition technologies: Threat analysis of DNS64 and state full NAT64", Computers & Security, Vol 77, Year 2018, pp 397-411.

[19] T. Saraj, and M. Yousaf, S. Akbar, A. Qayyum, and M. Tufail, "ISP Independent Architecture (IIA) for IPv6 Packet Traversing and Inter-connectivity over Hybrid (IPv4/IPv6) Internet", Procedia Computer Science, Vol 32, year 2014, pp. 973-978.

An Empirical Study of App Permissions: A User Protection Motivation Behaviour

Ari Kusyanti
Department of Information Technology
Universitas Brawijaya
Malang, Indonesia

Harin Puspa Ayu Catherina
Department of Information System
Universitas Brawijaya
Malang, Indonesia

***Abstract*—Smartphone is one of the telecommunications media that can be used anytime and anywhere. To be able to support the activity of its users, smartphone users install the application on their smartphone. When installing an application, there are permissions provided by the application about data that will be collected. However, many users choose to ignore and do not read the application permission since it is too long or difficult to understand, hence they accept the apps permissions without thinking and consequently leads to security problems. This study aims to determine the factors that affect users in reading apps permissions that have been provided by an application before they install the application. Data were collected from 292 respondents who were active in using smartphones. The data analysis method used is Structural Equation Modeling (SEM). The results of the study show that the factors that influence the user in reading the app permissions before they install the application are coping self-efficacy and personal responsibility.**

Keywords*—Protection motivation theory (pmt); application permission; smartphone; structural equation modeling (SEM)

I. Introduction

Nowadays, smartphone is one of the telecommunications media in the spotlight because it has the sophistication in various things and its effective and efficient functions that can be used anytime and anywhere [1]. A report from the Statista stated that the number of smartphone users in Indonesia in 2018 reached 70.22 million active users [2]. In fact, a few years ago the phone could only be used for the purposes of SMS and placing call only, but nowadays it has evolved into a smartphone that can be used for various purposes such as: social media, games and any other activities.

When users start installing an application, there will be an app permission displayed various data that will be accessed by the application as a perquisite. For example, an application that need to use the camera, microphone, internet, and other resources on smartphone, then Android apps will ask for permission. Apps will be installed if the user accept the apps agreement. To accept the apps permissions without thinking can have consequences, such as identity theft, disclosure of sensitive information without users noticing it, etc. Based on the initial survey that has been done in this study, users are less likely to read what is stated in the app permissions. It was found that the reasons why most users do not read the app permissions are: (1) they are too lazy to read the app permissions since it is too long (2) they think the app permissions is not important (3) reading the app permission is wasting time. The problem raises since some applications are requesting to retrieve some of the information contained on the smartphone that is mostly related to the users' sensitive information. Users are unaware of the problem that might occur in this situation.

As an example, the PlaceRaider app shows the danger of accepting permissions without reading before installing it. This app can take users' photos without permission and recreate room, hence users are vulnerable to be spied on. Another case in 2013, MouaBad malware allowed attackers to place an expensive calls without users noticing. Another malware called FireLeaker allowed the device's system database file to be accessed by attackers. They retrieved data such as phone numbers and other sensitive information and silently uploaded it to a web server managed by the attackers [3].

Based on the aforementioned reason, this study examines what factors that affect smartphone users to read app permissions by using the Protection Motivation Theory (PMT) model. The Protection Motivation Theory (PMT) model used in this study adapts from a study of Tsai et al. (2016) entitled "Understanding of online safety behaviors: A protection motivation theory perspective" which used nine variables to determine the factors that can influence the behavior of users to be able to apply protection against online security threats.

This paper is organized as follow. In section 2 provides the theoretical framework and the development of the hypotheses. Section 3 discuss the methodology and research design. Section 4 provides the results of empirical study while Section 5 includes a thorough discussion of the empirical findings. Finally, the conclusion is presented in Section 6.

II. Theoretical Framework And Hypotheses

A. *Protection Motivation Theory (PMT)*

Protection Motivation Theory (PMT) is used in predicting individual intentions to take protective action [4]. The factors in the research model namely response efficacy, threat severity, prior experience with safety hazard, coping self-efficacy, habit strength, perceived security support, personal responsibility, response cost, threat susceptibility and security intention.

Response efficacy is defined as the degree of user's belief regarding recommended protective behaviour whether it will be effective in terms of preventing or reducing the dangers or threats. When applying a protective behavior, a user must be able to ensure that the protective behavior he or she is performing is effective in protecting him or her from any danger or threat that might occur.

Threat severity is defined as the extent of the consequences of a threat or a hazard caused by the absence of protective behaviour. When a user does not consider that a threat is serious, then he or she will not apply protective behavior to protect him or her.

Prior experience with safety hazards is defined as the degree of user's previous experiences related to protective behaviour. When a user previously has experience in dealing with threats that occur online then he or she will apply protective behaviour.

Coping self-efficacy is defined as the degree of perceived ability and comfort with respect to user's behavior in conducting online protection. When a user is accustomed and has the ability to deal with online threats then he or she will apply the protective behaviour.

Habit strength is defined as the degree of how strong the habit of a user's ability in applying protective behavior to prevent online threats. When a user has a strong habit of applying protective behavior then he or she will continue to apply the protective behavior

Perceived security support is defined as the degree of the support from others related to applying online protective behavior. When a user is heavily influenced and can feel the support of others in applying protective behaviour, then he or she will apply protective behaviour.

Personal responsibility is defined as the degree of user's beliefs in implementing protective behavior to protect him or her. When a user has a strong belief that he or she can protect him or her from online threats then he or she will apply the protective behavior.

Response cost is defined as the degree of how much time and effort that has to spend to protect the user from online threats. A user will not implement protective behavior when he or she thinks that it takes too much effort in term of time and money.

Threat susceptibility is defined as the degree of vulnerabilities from online threats that may happen to user. A user will implement protective behavior when he or she realizes that he or she is vulnerable to online threats.

Security intention is defined as the degree of how much a user intends to apply protective behaviour. When the user feels that such protective behaviour is important then the user will likely continue to apply recommended protective behaviour.

B. Hypotheses Development

In a study conducted by [4] suggests that the response efficacy evaluates how effective the recommended protective behaviour in reducing a threat. When it comes to enforcing protective behaviour, users should ensure that protective behaviours undertaken will be effective in protecting them from threats. When a user installs an application, the user should read the app permissions provided before they install the application, because by reading the permissions is an effective protection measure to protect the user's smartphone from the perils of hackers who will take the data on their smartphone. Users who are aware that applying this protective behaviour is an effective step in reducing a threat will tend to have an intention to apply that protective behaviour. From this statement, it can be drawn hypothesis as follows:

1) Response Efficacy have a positive effect on Security Intention

Threat severity is used to measure how severe a threat can occur when protective behavior is not applied [4]. Users who consider that the impact of a threat is severe will tend to have the intention to apply protective behaviour. Currently, the use of smartphones can attract the attention of hackers to be able to break the security of the smartphone. One of them by taking data from the smartphone through a slot that is inserted through the app permissions provided when the user will install an application on the smartphone. In this study, this threat involves how severe the consequences of the occurrence of a threat when a user does not read the app permissions prior to installation. Based on the foregone review, the following hypothesis is developed:

2) Threat Severity have a positive effect on Security Intention

Based on study of [5] found that there was a significant relationship between a user's previous experience of a user's intention to apply protective behaviour. When a user has prior experience with online threats then he or she will tend to apply protective behavior. For example, on the use of smartphones when data on the smartphones had been accessed and misused by unauthorized party then a user tends to protect him or her by reading the app permissions prior to installation. Based on this the following hypothesis is developed:

3) Prior Experience with Safety Hazard have a positive effect on Security Intention

According to [5], there is a positive relationship between coping self-efficacy towards user intentions in applying protective behavior. Users who have the ability to protect the security of their data online will tend to implement protective behavior. In this study when a user has the ability to understand the importance of reading an app permissions prior to installation then he or she will not ignore the app permissions. Based on the statement, the hypothesis is drawn as follows:

4) Coping Self-Efficacy have a positive effect on Security Intention

Habit strength is used to measure how strong the habits of a user in applying protective behavior towards threats that may occur online [4]. In this case, user who has a habit and has been accustomed to protect him or her by reading the app permissions prior to installation will not ignore and will read it. According to the explanation, the hypothesis is shown as follows:

5) Habit Strength have a positive effect on Security Intention

As stated in [6] there is a positive relationship between perceived security supports to user's intention in applying protective behavior. When a user feels that he or she is supported by others in applying protective behavior then he or she will tend to have the intention to apply that protective behavior. When a user has the support from others to read the app permissions before installing an application then he or she will tend to read it. It will also allow user to have the intention to apply protective behavior to avoid potential threats when he or she ignores the app permissions. Based on the discussion the hypothesis can be drawn as the following:

6) Perceived Security Support have a positive effect on Security Intention

In this study, personal responsibility is used to measure user's self-confidence level in implementing protective behavior. A user who knows that by ignoring the app permissions prior to application installation can cause a threat and pose a risk that could harm him or her, then the user will tend to apply the protective behavior by reading the app permissions and understand the meaning of each permission requested by the application. This can prevent the occurrence of threats that may pose a risk to the user. Based on the explanation above, the hypothesis is drawn as follows:

7) Personal Responsibility have a positive effect on Security Intention

Pursuant to [7] found that response cost has a significant relationship to user intentions in applying protective behavior. A user who feels that by applying protective behavior is a waste of time and spent a lot of effort will tend not to apply protective behavior. If a user thinks that by reading the app permissions prior to application installation is something that takes a lot of effort and spends a lot of time then he or she will tend to not read the permissions. Furthermore, even if he or she knows that by reading the app permissions constitute a protective behavior to protect him or her against the dangers and threats that occur online, he or she will ignore it. From this statement, the following hypothesis is developed:

8) Response Cost have a positive effect on Security Intention

In a study conducted by [8] found that threat susceptibility has a significant effect on user intentions in applying protective behavior. A user who feels that he or she is highly vulnerable to any possible threats that may occur online will tend to apply protective behavior to protect himself or herself against online threats. When a user is aware that he or she is vulnerable to a threat by not reading the app permissions prior to application installation then he or she will tend to read the app permissions and continue to apply protective behavior. According to the review above, it can be drawn hypothesis as follows:

9) Threat Susceptibility have a positive effect on Security Intention

Based on the hypothesis that has been formulated above, Fig. 1 depicts the research model used.

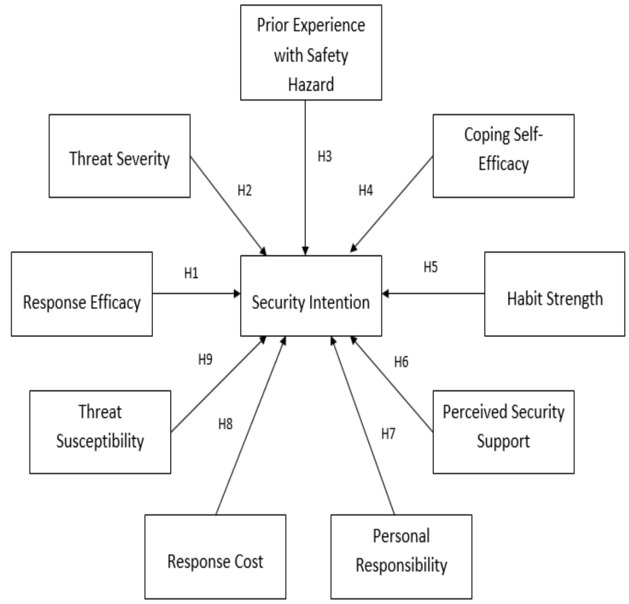


Fig. 1. Research Model.

III. METHODOLOGY AND RESEARCH DESIGN

A. Measurement Development

The questionnaire used in this study consisted of two parts. The first part contains respondents' demographic data, and the second part deals with their perceptions of application permissions. All questions in this questionnaire except demographic questions, all are based on a 5-point Likert scale, coded as, 5: strongly agree, 4: agree, 3: neutral, 2: disagree, 1: strongly disagree for measuring response efficacy, threat severity, prior experience with safety hazard, coping self-efficacy, habit strength, perceived security support, personal responsibility, response cost, threat susceptibility and security intention.

The original questionnaire was in English and was not suitable for the targeted subject. Therefore a translation approach is used to ensure that the original meaning will be maintained in Indonesian version. The English version of this instrument was first translated into Indonesian by one author, and then an independent translator translated the questionnaire into English. The original English version and the translation version is then compared, revised and corrected by three experts.

The preliminary analysis was pilot study by using Cronbach's Alpha. All the multi-item scales met the cut-off criteria of 0.6 as suggested by [9]. The value of Cronbach's Alpha for each variable in this study can be seen on Table I.

TABLE I. CRONBACH'S ALPHA VALUE

Factor	Value
Response Efficacy (RE)	0.775
Threat Severity (TS)	0.735
Prior Experience with Safety Hazard (PEW)	0.824
Coping Self-Efficacy (CSE)	0.905
Habit Strength (HS)	0.727
Perceived Security Support (PSS)	0.745
Personal Responsibility (PR)	0.824
Response Cost (RC)	0.675
Threat Susceptibility (TSUS)	0.721
Security Intention (SIN)	0.882

IV. DATA ANALYSIS AND RESULT

A. Mahalanobis Distance

The outlier test is used to find data that has extreme value by finding the value of Mahalanobis Distance. The result of Mahalanobis Distance equals to 45.13 hence data having Mahalanobis Distance above 45.13 must be removed from data processing. In this study there were 12 data that must be removed. Data that can be used in the next analysis were 280.

B. Kaiser--Meyer-Olkin and Barlett Test of Spericity

Kaiser-Meyer-Olkin Measure (KMO) was used to determine sampling adequacy [10]. The value of KMO in this study was 0.845 (>0.5) which means that factor analysis is feasible to do.

C. Normality Test

Normality test was used to determine whether the data to be used is normally distributed or not [10]. In this study, the normality test results of 0.068 (>0.5) which indicated that the data used in this study was normally distributed.

D. Homogenity Test

Levene test in this research was used to examine homogeneity of the data. This test was used to evaluate the similarities of variance throughout the data, with an assumption that a variance of a variable must be stable in all levels. Using Levene test would have the same variance if the Sig. value is > 0.05 [10]. In this research, all variables were considered homogeneous.

E. Measurement Model Fit

Measurement model fit was useful to determine the manifest variable (indicator) actually has a relationship with the latent variable (construct). This test was done by using Confirmatory Factor Analysis (CFA) method. Measurement model fit test results can be seen in Table II. Based on Table II, all values in this study met the specified criteria. Therefore the analysis can be continued at the structural model fit stage.

TABLE II. GOODNESS OF FIT INDEX VALUE

Index	Criteria	Value	Result
Chi-Square	> 0.05	807.91	Good
CMIN/DF	1<CMIN/DF< 3	1.616	Good
GFI	> 0.8	0.810	Good
RMSEA	< 0.08	0.059	Acceptable Fit

TABLE III. HYPOTHESES RESULTS

	Hypotheses	P < 0.05	Result
H1	SIN ← RE	.601	Rejected
H2	SIN ← TS	.347	Rejected
H3	SIN ← PEW	.331	Rejected
H4	SIN ← CSE	***	Accepted
H5	SIN ← HS	.371	Rejected
H6	SIN ← PSS	.785	Rejected
H7	SIN ← PR	.036	Accepted
H8	SIN ← RC	.108	Rejected
H9	SIN ← TSUS	.824	Rejected

F. Structural Model Fit

Structural model fit was performed to evaluate the relationship between variables that have a causal relationship or mutual relationship influence. This test was done by using Path Analysis method and the result showed in Table III.

Based on the results of structural model from 9 hypotheses that have been tested, there are 7 rejected hypothesis and 2 accepted hypothesis.

The impact of coping self-efficacy (p=***) and personal responsibility (p=0.036) on security intention were significant at p=0.05. Thus, H4 and H7 can be accepted. Meanwhile, respond efficacy, threat susceptibility, prior experience with safety hazard, habit strength, perceived security support, response cost and threat susceptibility have no significant impact on the intention to read the application permission, and thus H1, H2, H3, H5, H6, H8 and H9 were rejected.

V. DISCUSSION

A. Discussion on Hypothesis 1

In testing hypothesis H1, the results failed to support the proposition. Respondents feel that they are not convinced by reading the app permissions prior to application installation is an effective way in preventing or reducing the dangers and threats that may occur. They assume that there will be no danger or threats that occur even if they do not read the permissions of an application. Thherefore they have no intention to read the permissions of an application before they install the application. Therefore, in this study response efficacy (RE) had no significant effect on security intention (SIN).

The finding may further imply [11] which suggested that when users feel that the protective behavior to be applied is not an effective way of protecting them from the dangers and online threats then they will tend to have no intention in applying such protective behavior.

B. Discussion on Hypothesis 2

Hypothesis 2 was rejected. Based on the results of hypothesis testing 2, it can be concluded that respondents assume that the danger or threat caused by not reading the app permissions prior to application installation is not a severe threat. They also assume that even if they do not read the app permissions before they install the application there will be no serious threats that will occur. Thus, it makes them have no intention of reading the app permissions before they install the application. It established that in this study threat severity (TS) did not have significant influence on security intention (SIN).

The results of this study together with the results of research conducted by [4] who argued that when a user feels that a threat does not have a negative impact on the user then the user will tend to not have any intention in applying protective behavior.

C. Discussion on Hypothesis 3

With regards to H3, the results supported a negative relationship which indicates that respondents have no prior experience related to the hazards or threats caused by not reading the app permissions before they install the application. In previous experience, they feel there is no danger or threat that occurs because they do not read the permissions of an application. Thus, they have no intention of reading the app permissions as a protective behaviour. Therefore, in this study prior experience with safety hazard (PE) had no significant effect on security intention (SIN).

Prior literature [12] suggested that when users have no prior experience that could adversely affect their use of online applications, they will tend to have no intention of implementing protective behavior.

D. Discussion on Hypothesis 4

The results of hypothesis H4 supported the proposition that that respondents assume that when they have the ability to know about how severe a danger or threat that might occur when they choose not to read the app permissions before they install the application, it will make them have the intention to continue reading the app permissions as an act to protect themselves online. A significant direct association was found between coping-self efficacy (CSE) and security intention (SIN).

Similarly, previous literature [11] argued that when users have the ability to know and use an application and they have a sense of comfort in using it then the user will tend to continue to have intention in using the application.

E. Discussion on Hypothesis 5

In testing hypothesis H5, the results revealed a negative relationship. Respondents feel they do not have a habit to understand or read the app permissions before they install the application, even they tend to ignore it. It proved that most respondents in this study did not have a habit to read the app permissions before they installed the application. Thus, reading the app permissions is not considered as a protective behaviour. It indicated that in this study habit strength (HS) did not have significant influence towards security intention (SIN).

The finding is supported by [13] who suggested that when users do not have a strong habit of applying protective behavior then the user will tend to have no intention in implementing the recommended protective behavior.

F. Discussion on Hypothesis 6

Hypothesis 6 was accepted. Based on the results of hypothesis testing 6, it can be concluded that respondents do not have support from other parties related to the hazards or threats caused by not reading the app permissions before they install the application. If users do not have support from other parties, they will tend to feel that there is no danger or threat that occurs because they do not read the app permissions. This is because there is no single party around those who are exposed to such dangers or threats. Therefore, respondents have no intention in reading the app permission as a protective behavior. This revealed that in this study perceived security support (PSS) did not have significant influence on security intention (SIN).

The results of this study are similar to those of [4] who suggested that when users feel that no party around them can adversely affect or harm them on the use of an online application then they will tend to have no intention in implementing the protective behavior.

G. Discussion on Hypothesis 7

As it has been hypothesized, Hypothesis H7 supported the proposition that respondents have high confidence and responsibility in protecting themselves against the dangers or threats caused by reading the app permissions before they install the application. Respondents have a sense of confidence and responsibility arising from their own consciousness that by reading the app permissions is considered an act to protect themselves from possible dangers or threats. Thus, it encourages them to have an intention to read the app permissions prior to application installation. This study posited a significant relationship between coping-personal responsibility (PR) and influence on security intention (SIN).

This result confirmed previous literature [4] who argued that when a user has a high sense of responsibility in protecting themselves against threats or dangers online it will encourage the user to have intentions in implementing protective behavior.

H. Discussion on Hypothesis 8

Hypothesis 8 was not supported. Respondents consider having to read the app permissions before they install the application is a very time-consuming and requires a lot of effort. Furthermore, it is not guaranteed that they will be suffered from dangers or threats that may occur. They also find it uncomfortable to read the app permissions hence they have no intention of reading the app permissions prior to application installation. This study posited a negative relationship between response cost (RC) and security intention (SIN).

This is consistent with previous findings [14] which suggested that when users feel they need to take a lot of effort in using an application then they will tend to have no intention in using the application.

I. Discussion on Hypothesis 9

Hypothesis 9 was unaccepted. From the test result of Hypothesis 9, it can be concluded that respondents assume when they decide not to read the app permissions before they install the application there will be no high vulnerability to threats or dangers that may occur and harm them. They also assume that even if they do not read the app permissions, they will not be vulnerable to threats, hence they have no intention to read the app permissions before installing an application. The result found a significant relationship between treat susceptibility (TSUS) and influence on security intention (SIN).

The finding of this study with regard to threat susceptibility is consistent with previous studies [4] who argued that when users feel that they will not be vulnerable to threats caused by the use of an application then they will tend to not apply protective behaviour.

VI. Conclusions

This study aims at contributing in this viewpoint by determining the factors that influence users to read the app permissions before they install an application. Based on the results of the study, there are 2 factors that are found, namely: coping self-efficacy and personal responsibility. With regards to coping self-efficacy, respondents will read the app permissions prior to application installation as an effective way to prevent or reduce the dangers and threats that may occur since they have the ability to understand the importance of reading an app permissions as a protective behaviour. Meanwhile, concerning personal responsibility, respondents have high confidence and responsibility in protecting themselves against dangers or threats by reading the app permissions prior to installation. Additionally, the result of this study can raise users' awareness and inform them in term of protecting themselves by reading the app permissions before installing an application.

References

[1] Gary B. S., Thomas J. C., & Misty E. V., 2007. Discovering Computers: Fundamental 3rd.

[2] Statista. 2018. Number of smartphone users in Indonesia from 2011 to 2022 (in millions)* https://www.statista.com/statistics/266729/smartphone-users-in-indonesia/

[3] Benedetti, F. F. 2014. Stop ignoring Android app permissions. https://en.softonic.com/articles/stop-ignoring-android-app-permissions

[4] Tsai et al. 2016. Understanding of online safety behaviors: A protection motivation theory perspective. *Computers & Security* Journal. Volume 59, June 2016, Pages 138-150

[5] Lee D, LaRose R, Rifon NJ. 2008. Keeping our network safe: a model of online protection behaviour. Behav Inf Technol;27(5):445–54. doi:10.1080/01449290600879344.

[6] Luarn P, Lin H-H. 2005. Toward an understanding of the behavioral intention to use mobile banking. Comput Human Behav;21(6):873–91. doi:10.1016/j.chb.2004.03.003.

[7] Liang H, Xue Y. 2010. Understanding security behaviors in personal computer usage: a threat avoidance perspective. J Assoc Inf Syst;11(7):394–413.

[8] Zahedi FM, Abbasi A, Chen Y. 2015. Fake-website detection tools: identifying elements that promote individuals' use and enhance their performance. J Assoc Inf Syst;16:448–84

[9] Hair Joseph F. Jr., Black William C., Babin Barry J., Anderson Rolph E. & Tatham Ronald L. 2006. Multivariate Data Analysis, 6th edn, Pearson Prentice Hall, Pearson Education, Inc., Upper Saddle River, New Jersey 07458.

[10] Field, A., 2009. Discovering statistics using spss. 3rd ed. [e-book]. Sage Publications. Web Page: http://fac.ksu.edu.sa/sites/default/files/ktb_lktrwny_shml_fy_lhs.pdf

[11] Pahnila, S., Siponen, M., and Mahmood, A. 2007. Which Factors Explain Employees' Adherence to Information Security Policies? An Empirical Study. Pacific Asia Conference on Information Systems (PACIS)

[12] Hsieh, J. Y., & Liao, P. W. 2011. Antecedents and Moderators of Online Shopping Behaviour in Undergraduate Students. Social Behavior and Personality.

[13] Shin, K-h., Kim, G., and Lee, C., Y., 2014. A Study on Behavioral Intention and Use Behavior of Sharp (#) Mail Initial Experienced Users. Contemporary Engineering Sciences, Vol. 7, 2014, no. 30, 1657 – 1663

[14] Jeong, B. K., and Yoon, T. E. 2013. An Empirical Investigation on Consumer Acceptance of Mobile Banking Services. Business and Management Research ISSN 1927-6001 (Print) ISSN 1927-601X (Online)

Automated Extraction of Large Scale Scanned Document Images using Google Vision OCR in Apache Hadoop Environment

Rifiana Arief, Achmad Benny Mutiara, Tubagus Maulana Kusuma, Hustinawaty
Information Technology
Gunadarma University
Jakarta, Indonesia

***Abstract*—This Digitalization of documents is now being done in all fields to reduce paper usage. The availability of modern technology in the form of scanners and cameras supports the growth of multimedia data, especially documents stored in the form of image files. Searching a particular text in a large-scale scanned document images is a difficult task if the document is in the form of images where the text has not been extracted. In this research, text extraction method of large-scale scanned document images using Google Vision OCR on the Hadoop architecture is proposed. The object of research is student thesis documents, which includes the cover page, the approval page, and abstract. All documents are stored in the university's digital library. Extraction process begins with preparing the input folder that contains image documents (in JPEG format) in HDFS Apache Hadoop and followed by reading the image document. The image document is then extracted using Google Vision OCR in order to obtain text document (in TXT format) and the result is saved to output folder in Hadoop Distributed File System (HDFS). The same process is repeated for the entire documents in the folder. Test results have shown that the proposed methods were able to extract all test documents successfully. The recognition process achieved 100% accuracy and the extraction time is twice as fast as manual extraction. Google Vision OCR also shows better extraction performance compared to other OCR tools. The proposed automated extraction systems can recognize text in a large-scale image document accurately and can be operated in a real-time environment.**

Keywords—Automation; extraction; google vision OCR; hadoop; scanned document images

I. Introduction

Document digitization provides an effective way to process, maintain and transfer all types of information from printed form to digital form. The advancement of current information technology and the increased volume of printed documents in many applications, making digitalization of documents increasingly important to reduce paper-based physical documents. This is motivated by the emergence of several issues in the management of physical storage in the form of the risk of damage or loss of paper-based documents and the increasing pile of paper documents that require large storage.

Users in various institutions, such as government, education, medical, commerce and entertainment as well as private companies, have retained documents in electronic form and at the same time require a fast access service to the desired information[8]. The biggest challenges of large-scale digital document growth are scalability, data consistency, data completeness, time, and security (Chen and Zhang, 2014). Various algorithm has been developed continuously in order to capture, store, search, share, analyze, and visualize data, as well as to anticipate the increased of data volume by increasing the capacity using parallel processing (Chen and Zhang, 2014). Searching a particular text in a large-scale scanned document images is a difficult task if the document is in the form of images where the text has not been extracted.

Text extraction from images can be defined as the work of extracting text objects from a set of images. The results of text extraction can also be used as image search keywords, document search, content-based image search, video content analysis, text-based video search, location search words on documents and others [1]. Text extraction is a challenging task because there are variations of text size, font, style, orientation and alignment to a complex background. Text extraction process from scanned document images includes pre-processing, detection, localization, extraction, enrichment and text recognition. The image to be extracted can be a gray or colored image, in a compressed / uncompressed format. Text detection aims to find the presence of text in images and localization of text aims to find the location of the text and to create boundary boxes of text. The text is then extracted by separating the text from the background image and enriched to improve the quality of the extracted text in order to be recognized. Following the extraction process, the extracted text is recognized using Optical Character Recognition (OCR) [2].

OCR is a technology for recognizing text from images automatically. OCR supports various types of image formats such as JPG, PNG, BMP, GIF, TIFF and PDF files. OCR involves analyzing the captured or scanned images and then translating the image into an editable text format. The text contained in the scanned document image can be easily extracted with the help of an OCR tools. Various OCR applications are available and can be used to extract text on images. Many OCR tools available include Online OCR, Free Online OCR, OCR Convert, Convert image to text.net, Free OCR, i2OCR, Free OCR to Word Convert, Google Docs [6]. The reliability of Google Vision OCR has been shown to extract and recognize text from document images very well compared to other OCRs [5]. The extraction process of the excessive number of data is automatically performed in Hadoop environment using big data architecture.

Big Data is defined as high volume, high velocity, and/or high variety data sources, which requires new process paradigm to explore the information attached to it, to develop decision-making process, and optimization process. Based on this definition, Big Data is not characterized by specific size metrics, but completion in processing such data because the character (size, velocity or variety) is difficult with conventional processing approach. The Big Data potential is underlined in its definition; but the realization of such potential depends on developing traditional methods or developing new methods capable of handling such large data [4]. Hadoop is a platform for dealing with Big Data and provides problem solving with the ability to analyze large data. Hadoop is an open source based tool that enables distributed processing of large data with multiple clusters to accommodate services. Hadoop is designed to handle data from one service (server) to thousands of machines with high fault tolerance. Parallelization is used for the cost efficiency and processing time required. Big Data includes large-scale, diverse and complex data requiring new architectures, techniques, algorithms and analysis to manage the data and extract the hidden values and knowledge from the data set [3].

In 2013, Tae Ho Hong et. al. developed image-based or pdf-based ebook conversion system to facilitate the search for a words contained in the ebook [7]. To recognize text characters in image files using Tesseract OCR and this conversion process includes large data and uses Map Reduce Hadoop with cluster system so that the conversion process can be successfully done as well as minimizing the processing time. In this research, automated text extraction from large-scale scanned document images based on Google Vision OCR is proposed. The source of documents is stored in multiple folders with different file size using big data technology that is based on Apache Hadoop. The performance measurement of the automated extraction process will be based on the accuracy and the speed of extraction. Extracted results are stored in HDFS to be further analyzed for other purposes.

II. Proposed Methods

Object data used in this research are scanned document images of student’s thesis stored in the Gunadarma University library, which includes cover page, approval page, and abstract page as illustrated in Fig.1. The number of documents used in this research are 182,532 with data size of 33.4 GB. All documents are in the JPEG image formats.

The extraction stage aims to extract, to recognize and to get the text contained in the document images. The process starts with reading the input documents using Google Vision OCR and producing the output in the form of text documents. The output text document is stored in HDFS. The extraction process is shown in Fig. 2. The example of automated extraction process is demonstrated in Fig 3.

Fig. 1. Image Document Text in Dataset, (Source: Gunadarma University Digital Library).

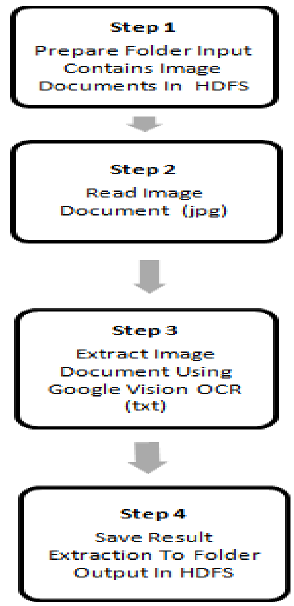


Fig. 2. Method in Extraction Phase.

The automated extraction process for detecting and recognizing text contained in the dataset documents using Google Vision OCR with single node Hadoop is presented in Fig.4. Prior to the extraction, the preprocessing must be done, which includes the preparation of input folder (document images) and to ensure that the document image folder contains document images ready to be extracted. Then, a folder must be created to stored extraction result (document text). All files contained in the input folder is read to get the entire filename, which will be included in the file list. The first document image file is extracted using Google Vision OCR and the fully extracted results is obtained. The extraction results needed are only in the form of text so that objects other than text will be removed or discarded and just take the text as a description then the contents of the description will be saved according to the file name of the scanned document, ie. cov1.jpg becomes cov1.jpg.txt. Following the above process, the results are stored in the HDFS output folder. The extraction process continues to the next document image until the entire contents of the completed folder is processed.

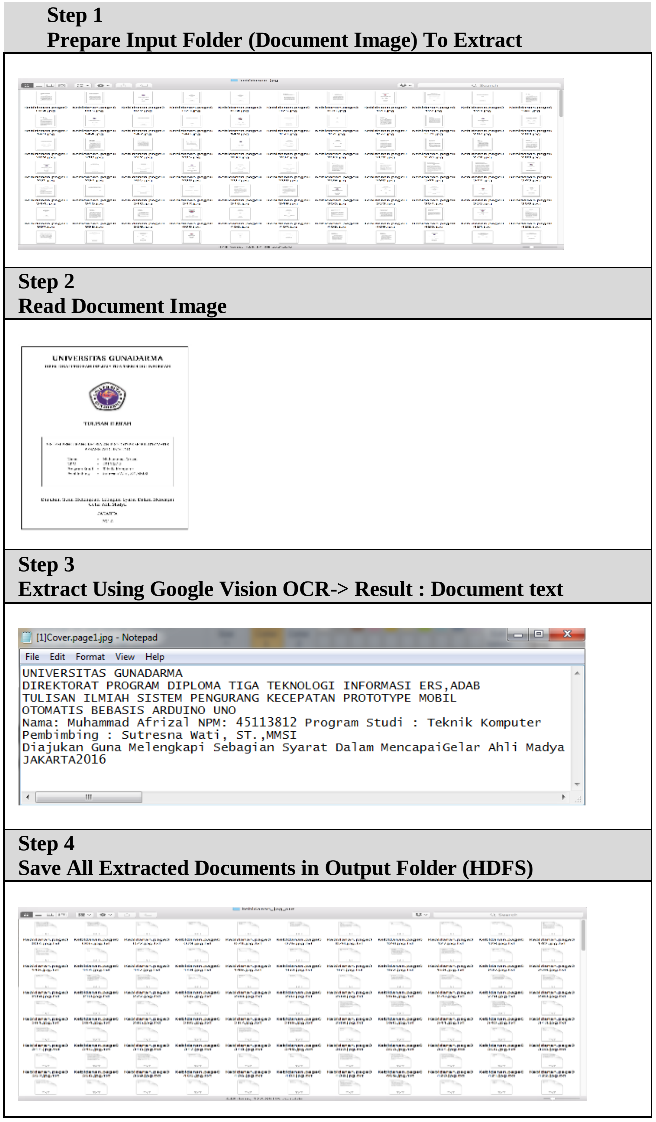


Fig. 3. Example of Automated Extraction Process.

III. Result and Discussion

Performance of the extraction process using Google Vision OCR were tested on three types of document images, namely cover document, approval document and abstract document. The details of the results are shown in Fig. 5, Fig. 6 and Fig.7.

All of the test results from the extraction process are summarized in Table 1. As shown in Table 1, the results of the extraction test using Google Vision OCR on cover documents, approval documents and abstracts have demonstrated successful recognition of the text contents. Limitation on the extraction of cover page has been found during the extraction process where text character in the logo image could not be extracted properly. Therefore logo image was excluded in the process. Limitations on the extraction of approval page has been identified where text characater overwritten by signature image could still be read but sometimes could not be accurately recognized. Although there are some limitations found, the overall performance of the extraction process using Google Vision OCR has shown good results.

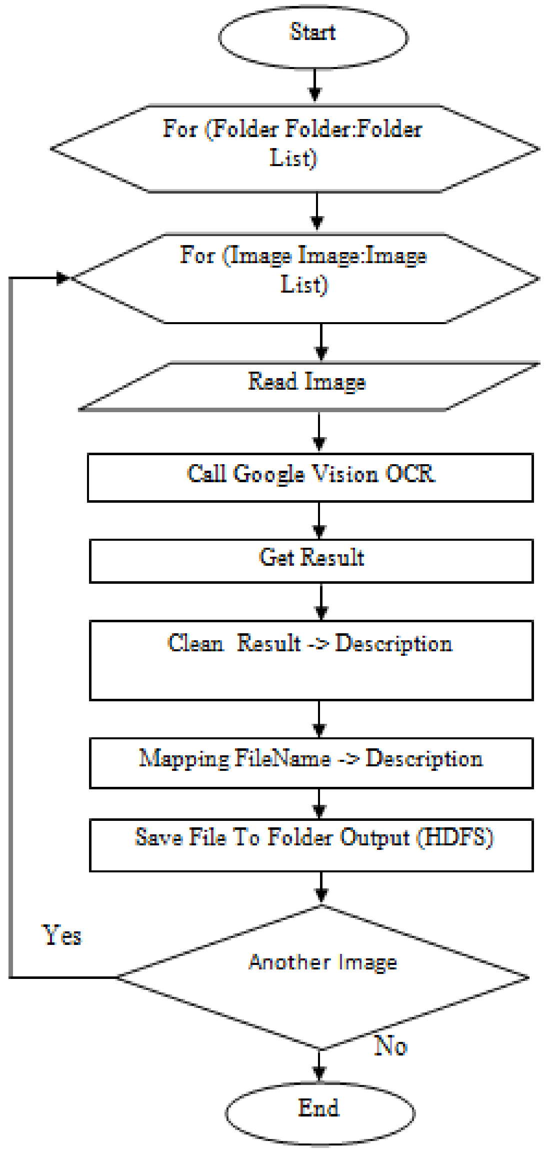


Fig. 4. Extraction Process using Google Vision OCR in Hadoop and Extraction Result Save in HDFS.

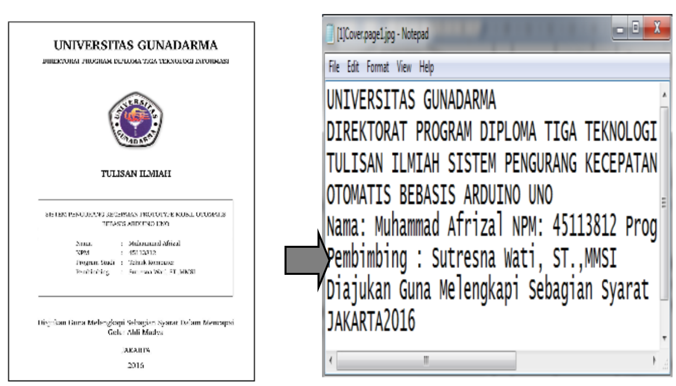


Fig. 5. The Extraction Result of Cover Document Image.

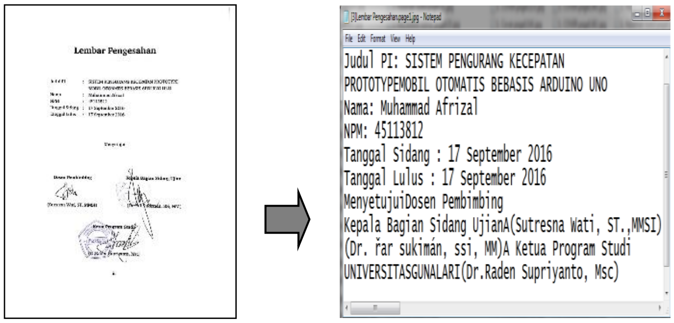


Fig. 6. The Extraction Result of Approval Document Image.

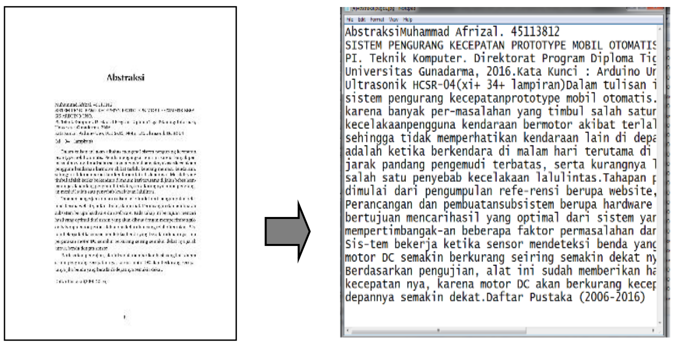


Fig. 7. The Extraction Result of Abstract Document Image.

TABLE I. SUMMARY OF EXTRACTION TEST RESULTS FOR DIFFERENT TYPES OF DOCUMENTS USING GOOGLE VISION OCR

No	Document Type	Document Properties	Results
1	Document Cover	Text and logo image with text character inside the logo	- The accuracy of text extraction is very good - The text on the logo image is readable but not correct, should be removed (not extracted)
2	Document Approval	Text and signature image overwritten text character	- The accuracy of text extraction is very good - The accuracy of the name text is sometimes incorrect because of the signature overwritten the text
3	Document Abstract	Text only	- The accuracy of text extraction is excellent with no errors occurred.

For benchmarking purpose, text extraction using Free Online OCR tools and Onlineocr.net has been conducted. The results of these extraction process are shown in Table 2.

TABLE II. THE EXTRACTION RESULTS OF DOCUMENT COVER, APPROVAL AND ABSTRACT WITH FREE ONLINE OCR AND ONLINEOCR.NET

No	Document	FreeOnlineOCR	Online ocr.net
1	 Cover		
2	 Approval		
3	 Abstract		

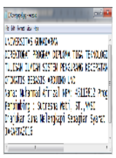
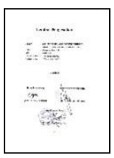
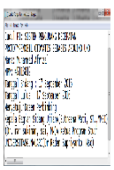
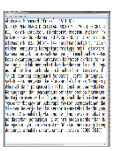
TABLE III. COMPARISON OF MANUAL EXTRACTION PROCESS OF 1 DATA GOOGLE VISION OCR EXTRACTION RESULTS WITH OTHER OCR TOOLS (FREE ONLINE AND ONLINEOCR.NET)

Doc	Google Vision OCR		Free Online		OnlineOCR. net	
	Accu racy	Time	Accu racy	Time	Accu racy	Time
Cover	97%	5 Sec	85%	41 Sec	95%	7 Sec
Agreement	95%	5 Sec	80%	54 Sec	90%	8 Sec
Abstract	97%	5 Sec	90%	46 Sec	93%	6 Sec

The performance comparison in terms of accuracy and recognition time of Google Vision OCR, Free Online, and OnlineOCR.net in extracting single set of document (cover, approval, and abstract) manually is presented in Table 3. Accuracy is calculated by comparing the number of words in a scanned document that can be recognized properly with the total number of words contained in a scanned document, then multiplied by 100%. As shown in the table, the accuracy of Google Vision OCR is the highest, followed by OnlineOCR.net, and then by the FreeOnlineOCR as the lowest. In terms of execution time, Google Vision OCR performed the fastest, followed by Online ocr.net and then by the FreeOnlineOCR as the slowest.

Total extraction time for the entire document (Cover, Approval and Abstract), which is 182,532 files is 15,215 minutes if it was done manually and 7,301 minutes if it was done automatically. In comparing the performance of Google Vision OCR while running manually and automatically, it is shown that the manual process took about 5 seconds to extract while the automated extraction process took about 2.4 seconds to extract a single document. The details are presented in Table 4.

TABLE IV. TIME OF EXTRACTION PROCESS FOR MANUAL AND AUTOMATIC USING GOOGLE VISION OCR FOR ALL DOCUMENTS ON SINGLE COMPUTER

Document Image (jpg)	Extracted Document (txt)	Number of Document	Time to Extract Manual (Minutes)	Time to Extract Automatic (Minutes)
		64,249	5,354	2,570
		59,932	4,999	2,397
		58,351	4,862	2,334
TOTAL		182,532	15,215	7,301

IV. CONCLUSION

This paper presents the implementation of Google Vision OCR to recognize text content in a document image and introduces an automated extraction framework to a large-scale text document image collection in Hadoop architecture. The Google Vision OCR was selected, because it has proven an excellent accuracy compared to other tools. Based on the results of the automated extraction test, the average automated extraction process in Hadoop environment using single computer is approximately 2 times faster than manual extraction time. All documents in the input folder were successfully extracted, while at the same time the text recognition reached almost 100% accuracy.

Along with the growth of significant data, the future work will be to build an automated extraction system by implementing the automated extraction process using multiple computers in parallel so as to reduce the time required as well as the workload of the computer if using only a single computer. The used of larger datasets with different characteristics will also be considered to see the performance of the proposed system in handling various types of documents. Other types of documents such as goverment or private data source agencies or documents from the Internet, where scanned document images can be extracted with OCR might also be included. Real-time document retrieval and classification based on text content might also be considered for future work.

ACKNOWLEDGMENT

We would like to thank Gunadarma University Library for their permission to use the students's thesis data.

REFERENCES

[1] C.P. Sumathi, T. Santhanam and G.Gayathri Devi. “A Survey On Various Approaches Of Text Extraction In Images”. International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.4, August. 2012. Pp 27- 42. 2012

[2] Divya gera and Neelu Jain. “Comparison of Text Extraction Techniques- A Review”. International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2015. Pp 621-626. 2015

[3] Harshawardhan S. Bhosale, Devendra P. Gadekar. 2014. “ A Review Paper on Big Data and Hadoop”. International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 pp 1-7

[4] Beyer, M.A. and Laney, D. “The Importance of “Big Data: A Definition”. In Gartner. 2012. https://www.gartner.com/doc/2057415.

[5] H. Choudhary, “Comparing the Top Computer Vision APIs for OCR, in Article Computer Vision, 2017

[6] S.Vijayarani and A.Sakila. “Performance Comparison Of OCR Tools” . *International Journal of UbiComp (IJU),* Vol.6, No.3, pp.19-30, 2015.

[7] Tae Ho Hong, Chang Ho Yun, Jong Won Park, Hak Geon Lee, Hae Sun Jung and Yong Woo Lee, “Big Data Processing with MapReduce for E-Book”. International Journal of Multimedia and Ubiquitous Engineering Vol. 8, No. 1, January, 2013 pp. 151-162.

[8] Yoganand. C.S, Praveen.N, Saranya.N, Ganesh Karthikeyan V., “Survey on Document Classification based on Keyword and Key Phrase Extraction using Various Algorithms“. International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 2, February – 2014. Pp 1804-1808. 2014.

Optimal Overcurrent Relays Coordination using an Improved Grey Wolf Optimizer

Noor Zaihah Jamal[1], Mohd Herwan Sulaiman[2], Omar Aliman[3], Zuriani Mustaffa[4]
Faculty of Engineering Technology[1],
Faculty of Electrical & Electronics Engineering[2,3],
Faculty of Computer Systems and Software Engineering[4]
Universiti Malaysia Pahang
26300 Gambang, Pahang

***Abstract*—Recently, nature inspired algorithms (NIA) have been implemented to various fields of optimization problems. In this paper, the implementation of NIA is reported to solve the overcurrent relay coordination problem. The purpose is to find the optimal value of the Time Multiplier Setting (TMS) and Plug Setting (PS) in order to minimize the primary relays' operating time at the near end fault. The optimization is performed using the Improved Grey Wolf Optimization (IGWO) algorithm. Some modifications to the original GWO have been made to improve the candidate's exploration ability. Comprehensive simulation studies have been performed to demonstrate the reliability and efficiency of the proposed modification technique compared to the conventional GWO and some well-known algorithms. The generated results have confirmed the proposed IGWO is able to optimize the objective function of the overcurrent relay coordination problem.**

***Keywords*—Time multiplier setting (TMS); plug setting (PS; grey wolf optimization algorithm (GWO); overcurrent relay coordination**

I. Introduction

The electricity demand is keep increasing from year to year to accommodate the grown of the human population. In order to provide the best services, the old power system must be improved and transformed to be more compatible. Complex electrical power networking systems comprise with switchgears, transformers, ring main units and motors. All the equipment is located at different voltage rating which needs to be protected in to ensure that any fault occurrences are under control and does not affect the healthy portion of the system. To ensure the flexibility of the system to withstand any abnormal condition, the numbers of protective devices must be well arranged and coordinated.

The overcurrent relay coordination problem has been recognised as a constrained optimization problem [1-4]. Optimization of the overcurrent relay operating time (Top) is certified by two parameters which are Time Multiplier Setting (TMS) and Plug Setting (PS). These two parameters are formulated as Mix Integer Non-Linear Programming (MINLP) problem. No matter how details the progress undergoes during design stage, it is impossible to build a system without failure with external cause [5]. However, the huge catastrophe could be reduced with good and well-coordinated protection scheme. The good protection scheme should comprehend the requirements of sensitivity, speed, reliability and last but not least selectivity. Moreover, in this modern complicated electrical networking system, more numbers of relays should be coordinated.

During decade back, the implementation of analytical and graphical approach as in [6, 7] has been used to coordinate the overcurrent relay. The improvements of the technique have been done in [8], derivation of new non-standard tripping characteristic. In [9], a new method for repairing and inspecting curve crossing between primary and back-up relay has been developed. Meanwhile, to tackle sympathy trips threats to the system additional constraint has been introduced in [10].

Modern techniques by nature inspired have been introduced which started with Genetic Algorithm (GA) [11-14]. GA has becoming a most popular algorithm in this area in early 90s. Improvement to this algorithm have been made in [15] called Continuous Genetic Algorithm (CGA) where CGA has been proven to be faster in result generated compared to binary GA, since the chromosome in CGA does not need to be decoded. Ref. [16] has developed an improvement method to solve the mis-coordination problem which updated the weighting factors during simulation called fuzzy based Genetic Algorithm method. Next evolution of bio-nature inspired technique is introduced in [17-19] called as Particle Swarm Optimization (PSO). The PSO has been proven to provide better result compared to conventional GA and modern GA. The revolution of the algorithm is continued by Differential Evolution (DE) and Modified Differential Evolution (MDE) method as in [20-22], and Invasive weed optimization [23]. In order to generate better performance of MDE, hybrid method has been developed in [3, 24-26]. Cuckoo Search Algorithm is developed in [27]. Electromagnetic Field Optimization (EFO) method in [28] and Improved GSO has been introduced in [1]. All of these algorithms are developed to search for the best overcurrent relays setting. Hybridization of some methods such as PSO-TVAC [29], GA-NLP [30], Fuzzy based-GA [31] and Hybrid PSO [32]. are also developed to improve the generated optimum results.

Recently, a new reliable and robust algorithm have been introduced known as Grey Wolf Optimization (GWO) technique. This GWO algorithm have been implemented in [33] in biomedical engineering field, optimal reactive power dispatch problem [34] and combined economic emission dispatch problems [35]. GWO has been introduced by [36] which is inspired by hunting behavior of a group of wolves. Some amendments have been applied to the conventional GWO in order to improve the exploration rate of the searching agents. The conventional GWO has been identified having low convergence speed and in most cases being trapped in local optimal. The recommended improvement has increased the number of searching agents instead of role as followers to the first three best agents. The objective of this paper is to pick the best TMS and PS value in order to minimize the objective function.

This paper is organized as follows; section II presents overcurrent relay coordination problem formulation. Explanation of the conventional GWO is presented in section III. Section IV explain on the improvement of the GWO algorithm. Results and analysis is presented in Section V. Finally, section VI concludes the achievements of the proposed algorithm.

II. Problem Formulation

The coordination problem of overcurrent relays is formulated as an optimization problem. To optimize the nonlinear objective function, various nonlinear inequality constraints shall be satisfied.

A. Objective Function

The objective of the coordination problem is minimization of the primary relays' total operating time and remain the primary - backup pair relays coordinated with fulfilled the 0.2s – 0.5s coordination time interval (CTI). The minimization of the relay's operating time is close related to the optimization of the value of TMS and PS. The objective function is.

$$\min F = \sum_{i=1}^{n} \omega_i T_i \tag{1}$$

Where ω_i is the weight of relay R_i and n is the number of relays inside the system. While T_i is the operating time of primary relay. Generally the value of ω_i is set as one [30, 37], hence (1) becomes:

$$\min F = \sum_{i=1}^{n} T_i \tag{2}$$

The relay operating time is define by IEC standard [38] as

$$T_i = TMS_i \times \frac{k}{\left(\dfrac{I_{sc}}{PS_i}\right)^{\alpha} - 1} \tag{3}$$

Where PS_i is plug setting for relay R_i, TMS_i is time multiplier setting for relay R_i, I_{sc} is short circuit current which seen by relay R_i

B. Constraints

The objective function is possible to be achieved if relay parameters contraints and coordination constraints are fullfilled.

The relay parameters constraints are *TMS* and *PS* boundaries

$$PS_i^{\min} \le PS_i \le PS_i^{\max} \tag{4}$$

The boundary of the PS can be calculated as

$$PS_{\min} = 1.25 \times I_n \tag{5}$$

$$PS_{\max} = \frac{2}{3} \times I_{f\,\min} \tag{6}$$

Where I_n is the normal current rating which protected by the relay R_i. $I_{f\,min}$ is the minimum value of current which is detected as fault by relay R_i

The boundary of TMS is given as

$$TMS_i^{\min} \le TMS_i \le TMS_i^{\max} \tag{7}$$

The TMS value is the time delay that varies from 0.1 to 1.1 [3, 4]. Where TMS_{min} is minimum limit and TMS_{max} is maximum limit value of TMS for relay R_i.

The coordination constraints is in between Back-up and Primary relay. The selectivity should fullfilled the time interval required. The primary relay should reacted in advanced during fault occurences as compared to back-up relay and not vise versa to escape any sympathy trips

$$CTI = T_{bc} - T_{pr} \tag{8}$$

Where T_{pr} is primary relay time operating, T_{bc} is the back-up relay time operating and CTI varies between 0.2s – 0.5s [3].

III. Grey Wolf Optimization Algorithm

This section presents an overview of the conventional grey wolf optimization algorithm. Details on the GWO can be found in [36]. Then, in the next section the improvement to the proposed algorithm will be presented.

A. Conventional Grey Wolf Optimization Algorithm

The Grey Wolf Optimizer is derived by leadership hierarchy and hunting of grey wolf. The dominant social hierarchy of grey wolf have an average group of 5-12 members. The first tier called alpha (α) which dominating the group and responsible for decisions making as a leader. The dominant alpha is selected based on ability to manage their group members well.

The next tier is called beta (β) role as assistance to alpha in order to enforce any instruction or command by the leader. Beta could be the next leader with good discipline criteria which can be either male or female.

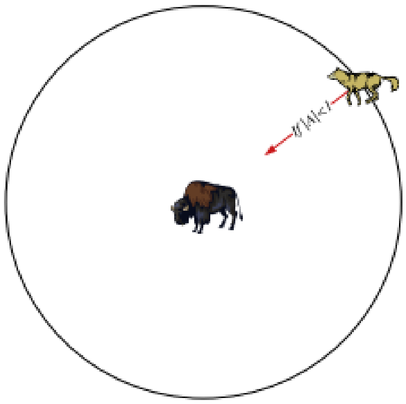


Fig. 1. Attacking the Prey [36].

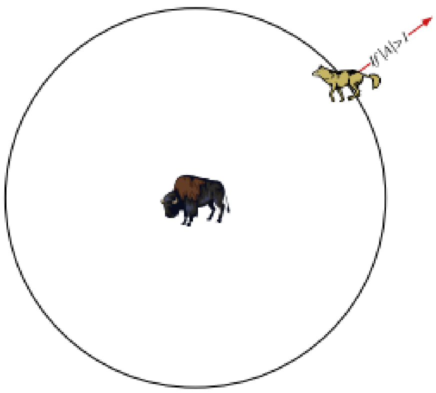


Fig. 2. Diverge from the Prey [36].

Delta (δ) is once used to be beta and alpha would be placed on the third-tier roles as hunters, caretakers to the younger members, sentinels and scouts. Hunters help foods delivering to the group members. Caretakers take care of the weak, ill and wounded young members. Sentinels control the security of the members and guarantee their territory safety and scouts role as territory marker to monitor the boundaries and discover any dangers ahead.

The bottom ranking is Omega (ω). Omega appears to be a balance to the nature bio-chain of the grey wolf. Even though their existence is not really appreciated by the other members of the group but still their role as a babysitter to the group can be acceptable. They are last wolves that are permitted to eat the prey.

In grey wolf community, the hunting activity is categorized by three phases as follows:

- Tracking: trace the location of the prey.
- Encircling: trap the prey in a circle.
- Attacking: move towards the prey by fulfilling the terms.

Alpha will lead during the hunting activities as the best solution, followed by Beta as second best and Delta as the third best. Omega will update positions as remaining solution by considering the position of the first, second and third best of the group.

For mathematical encircling activity behaviour modelling, below equation is considered [36]:

$$\vec{D} = \left|\vec{C} \cdot \vec{X}_p(t) - \vec{X}(t)\right| \tag{9}$$

$$\vec{X}(t+1) = \vec{X}_p(t) - \vec{A} \cdot \vec{D} \tag{10}$$

Where $\vec{X}_p$ is the position of the prey, $\vec{X}$ is the grey wolf position vector, $\vec{C}$ and $\vec{A}$ are vector's coefficient and t is the present iteration. The formulation of the vector's coefficient are as following equation [36]:

$$\vec{A} = 2\vec{a} \cdot \vec{r}_1 - \vec{a} \tag{11}$$

$$\vec{C} = 2 \cdot \vec{r}_2 \tag{12}$$

According to grey wolf hunting behavior, they will re-positioning their current location according to the position of the prey. The value of vector $\vec{A}$ and $\vec{C}$ will be the updated position with respect to the current position of the wolf which means, adjusting the value of $\vec{A}$ and $\vec{C}$ can placed the wolf to the different places. where $\vec{a}$ are linearly reduced from 2 to 0 over the iterations course, $\vec{r}_1$ and $\vec{r}_2$ are random vectors within (0,1). The random value of $\vec{r}_1$ and $\vec{r}_2$ allows agents to move to any position around the prey in random location by using eq. (12) and (13).

It is tough to locate the prey's location furthermore in an open search area. For mathematical hunting activity modelling purposes, the alpha, beta and delta are assumed to have knowledge on the prey's location based on their bio-nature capabilities. Therefore, the first solution of α, β and δ force the remaining search agents (including ω) to update their locations by referring according to the location of the best search agents [36].

The following formulas are obtained.

$$\vec{D}_i = \left|\vec{C}_n \cdot \vec{X}_i - \vec{X}\right| \tag{13}$$

$$\vec{X}_n = \vec{X}_i - \vec{A}_n \cdot \left(\vec{X}_i\right) \tag{14}$$

Where *i* indicate the search agent of *α, β and δ* and *n= 1, 2, 3*..

$$\vec{X}(t+1) = \frac{\vec{X}_m + \vec{X}_{m+1} + \vec{X}_{m+2}}{3} \tag{15}$$

Where *t* is the present iteration and *m=1* which indicate the updated position of the *α, β and δ*.

The random position within the search area is updated according to the first three best solutions. The estimated position of the prey by alpha, beta and delta will then be a guide to omegas to update their positions.

The last stage of hunting is by attacking when the prey is in static position. The decreasing value of $\vec{a}$ is when the wolves are approaching the prey. This will also decrease the value of $\vec{A}$ which $\vec{A}$ is a random value between (-2a, 2a). The wolves are moving forward to attack the prey if $|A| < 1$ as in fig. 1 and fig. 2. The process is repeating for the next iteration until the termination criterion is justified.

B. Improved Grey Wolf Optimization (IGWO)

The most challenging task in bio-nature population is to avoid the searching agents from trapping inside the local optimal. The end result of the objective function is influenced by this trapping problem and only near optimal solution is generated. The converging towards global optimal could be segregated in two different conditions. At the first place, the searching agents should be motivated to disperse throughout the wide range of searching space to find out the potential prey instead of crowding around the consistent local optimal. This stage also called as exploration stage. In the next stage which called exploitation stage, where the searching agents should be able to manipulate the knowledge of the potential prey to converge towards the global optimal value. In GWO, fine tuning of the parameters $\vec{a}$ and $\vec{A}$ could balance these two stages.

From the eq. (12), the coefficient vector of $\vec{A}$ is influence by component $\vec{a}$ with the formulation as follows [36]:

$$\vec{a} = 2 - 1 * \left(\frac{2}{\max iter} \right) \tag{16}$$

Some recommendation by researchers' that the exploration stage motivates the searching agents to update their position stochastically and abruptly. This situation has improved the variety of the solution and resulted to increase exploration wisdom in the search space.

But on the other hand, the exploitation is focusing on improving the solution's quality by searching locally around the promising area. In this stage, the search agents are obliged to search locally.

In general, the probability of the local optimal trapped could be avoided with the wisdom of explorations by the searching candidates. In conventional GWO, tracking or hunting activity is only considered the knowledge of the alpha, beta and delta whereas the rest wolves are obliged to follow them including omega.

In order to increase the exploration wisdom of the search agents, some modification to the conventional GWO algorithm has been recommended. Improved GWO (IGWO) algorithm proposed that omega should be considered as a searching agent instead of obliged to follow the first three best candidates. The increasing of the numbers of searching agents improve the search ability of the grey wolves in a wide range of search space. This improvement motivates the search agents to be scattered during exploration stage. In other words, that the wide range of the search space could be explored in further by the increasing of the search agents. The hunting activity could be more efficient and time saving. The mathematical modelling of the IGWO hunting agents are as follows:

$$\vec{D}_\alpha = \left| \vec{C}_1 \cdot \vec{X}_\alpha - \vec{X} \right| \tag{17}$$

$$\vec{D}_\beta = \left| \vec{C}_2 \cdot \vec{X}_\beta - \vec{X} \right| \tag{18}$$

$$\vec{D}_\delta = \left| \vec{C}_3 \cdot \vec{X}_\delta - \vec{X} \right| \tag{19}$$

$$\vec{D}_\omega = \left| \vec{C}_4 \cdot \vec{X}_\omega - \vec{X} \right| \tag{20}$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot \left(\vec{X}_\alpha \right) \tag{21}$$

$$\vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot \left(\vec{X}_\beta \right) \tag{22}$$

$$\vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot \left(\vec{X}_\delta \right) \tag{23}$$

$$\vec{X}_4 = \vec{X}_\omega - \vec{A}_4 \cdot \left(\vec{X}_\omega \right) \tag{24}$$

$$\vec{X}(t+1) = \frac{\vec{X}1 + \vec{X}2 + \vec{X}3 + \vec{X}4}{4} \tag{25}$$

In [39], it is argued that, too much exploration will have resulted to too much randomness and probably generates bad results. However, this argument could be counteracted by the increased numbers of the active exploration agents. The flow chart of the application of IGWO to relay coordination problem as fig. 3. The pseudocode of the IGWO as in fig. 4

```
Initialized the search agents' population Xi (i=1, 2…..,n)
Initialized the a, A and C
Calculate the fitness of each search agent
Xα = the best search agent
Xβ = the second best search agent
Xδ =the third best search agent
while (t < max iteration no.)
        for each search agent
           Update the position of current search agent by (26)
        end for
        Update a, A and C
        Calculate the fitness of the search agents
        Update Xα, Xβ, Xδ and Xω
        t = t+1
end while
return Xα
```

Fig. 3. Pseudo Code of IGWO Algorithm.

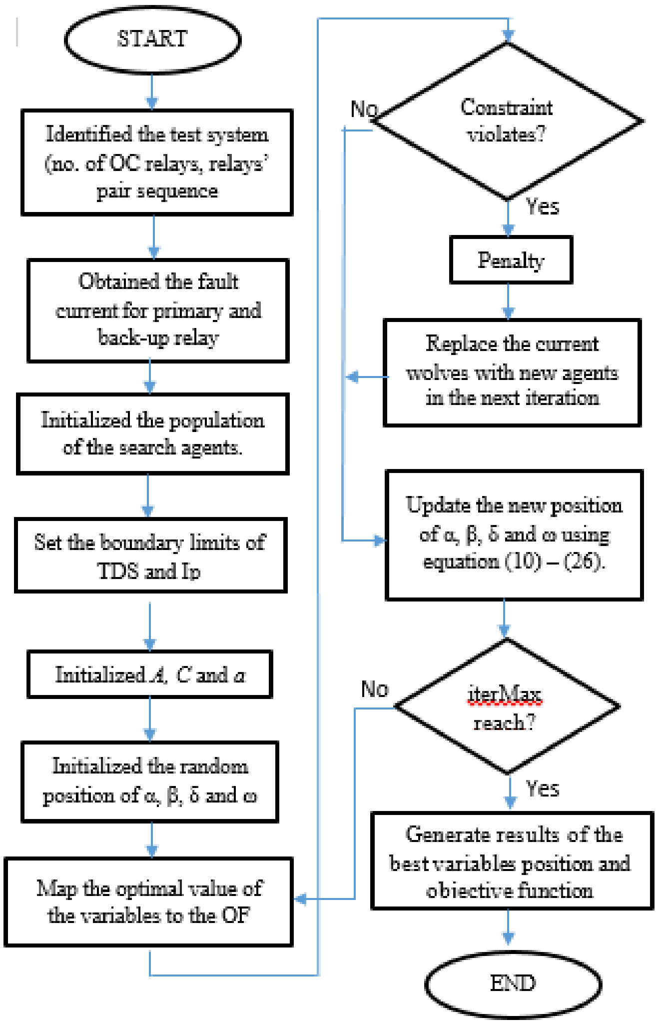


Fig. 4. Implementation of IGWO to Coordination Overcurrent Relay Problem.

IV. Results and Discussions

Simulations have been performed to three different IEEE test cases (three-bus, eight-bus and 15-bus test system) to test the efficiency of the GWO and IGWO techniques. The simulations are using MATLAB software and executed on an intel core i5-6200U CPU, 2.3GHz with 8GB RAM. The implemented value of CTI is 0.2 to 0.5s. The constant values used are according to IEC standard [38] and implemented normal inverse characteristic to all of the test case where with k = 0.14 and α = 0.02

A. Case I

The system consists of three busbar (B_1, B_2 and B_3), six overcurrent relay (R_1, R_2,…R_6), three ring lines and powered by three generators with 69kV system voltage. The TMS and PS are considered as variables which bound from X_1 to X_6 and X_7 to X_{12} respectively.

The results are presented in MINLP with continuous TMS and PS models for this case study. The search agents are 30 and iteration no. implemented is 1000.

TABLE I. COMPARISON OF IGWO WITH OTHERS TECHNIQUE FOR CASE I

Method	**Objective function (s)**
Modified PSO [2]	1.9258
MINLP [4]	1.727
Seeker Algorithm [4]	1.599
GWO	1.5124
IGWO	1.4789

TABLE II. COMPARISON OF IGWO WITH OTHERS TECHNIQUE FOR CASE I

Relay no.	**CT**	**GWO**		**IGWO**	
		TMS	**PS**	**TMS**	**PS**
1	300	0.1000	3.0	0.1000	1.5000
2	200	0.1001	1.5	0.1000	2.6166
3	200	0.1000	3.0	0.1001	2.9770
4	300	0.1000	3.0	0.1000	1.5858
5	200	0.1001	1.5	0.1000	2.8169
6	400	0.1000	1.5	0.1000	1.5009
Result (s)		1.5124		1.4789	

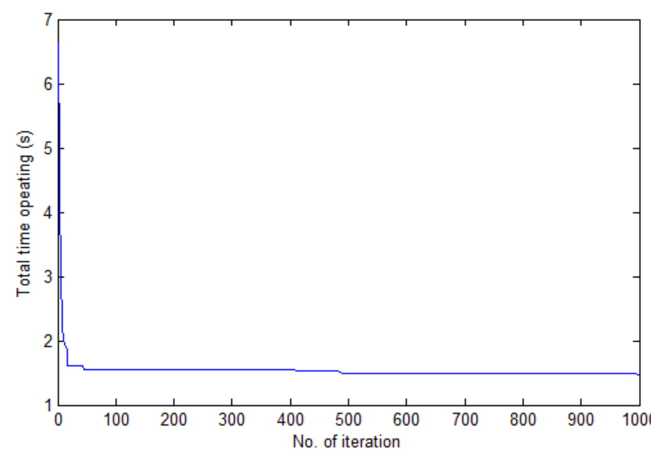


Fig. 5. Best Result for 1000 Iteration.

In [40], the details of this test case can be obtained. The TMS values is bound from 0.1s to 1.1s [3, 4] and the PS values bound from 1.5 to 5[4]. The CTI value of 0.3s is applied to this three bus test case.

Table I shows the comparative results of the IGWO with the modified PSO, MINLP, Seeker Algorithm and conventional GWO. The optimized result of conventional GWO and IGWO are shown in Table II. From table II, it can be seen that the IGWO performs better solution with 0.0335s faster than GWO. This has proven that improvement of GWO performs the best way compared to the others technique applied before. Fig. 5 shows the generated best solution for 1000 iteration with 30 agents.

The best result in fig.5 has shown the efficiency of the IGWO in 30 free running conditions while in fig. 6, the convergence of the mean and best result is presented.

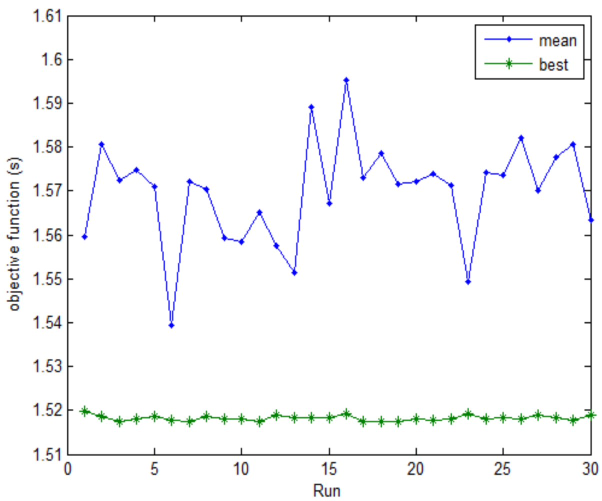


Fig. 6. Performance of IGWO for 30 Free Running.

B. Case II

The case 2 consist of 14 overcurrent relays (R_1, R_2,…R_{14}), seven ring lines to connect six busbars (B_1, B_2…..B_6) as in fig. 7. The bound of TMS value from X_1 to X_{14} and bound of PS is from X_{15} to X_{28}. The dimension of variables is 28 with constraints of 20.

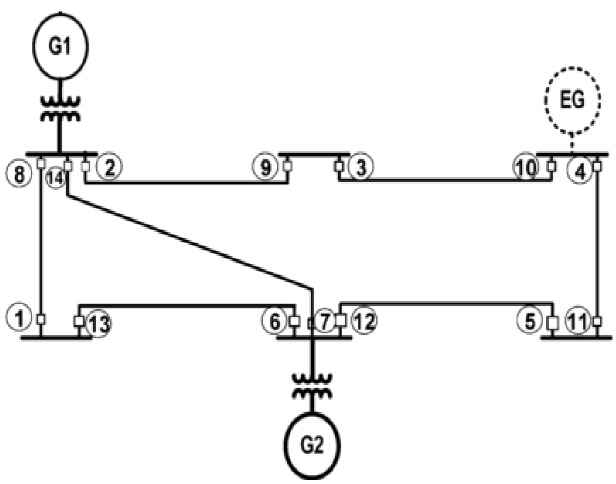


Fig. 7. Test Case 2 with IEEE 8 Bus System.

The TMS values are varies in between 0.1s to 1.1s and the PS values are in between 1.5 to 5. Both TMS and PS are continuous models. The current transformer ratio of each relays are as stated in table III. The details of this test system can be obtained from [19].

The comparative results of the IGWO with GA-NLP, CSA, Seeker Algorithm and conventional GWO are tabulated in Table III.

Table IV shows the optimized value of TMS and PS for conventional GWO and IGWO. Based on the obtained results, the IGWO has decreased the total operating time of the primary relays around 0.0114s compared to conventional GWO.

TABLE III. COMPARISON OF IGWO WITH OTHERS TECHNIQUE FOR CASE II

Method	Objective function (s)
GA-NLP [30]	10.9499
CSA [41]	9.7628
Seeker Algorithm [4]	8.4270
GWO	8.0979
IGWO	8.0865

TABLE IV. OPTIMUM SETTING FOR CASE II

Relay no.	CT	GWO		IGWO	
		TMS	PS	TMS	PS
1	1200	0.1001	1.5014	0.1000	1.5000
2	1200	0.1101	4.1984	0.1000	4.4101
3	800	0.1759	4.9241	0.1722	4.9957
4	1200	0.4346	5.0000	0.435	5.0000
5	1200	0.1015	1.5158	0.1001	1.5082
6	1200	0.1008	1.5214	0.1015	1.5048
7	800	0.1003	1.5801	0.1055	1.9719
8	1200	0.1010	4.3769	0.1004	4.4778
9	800	0.1016	4.6648	0.1000	4.7092
10	1200	0.1008	4.1119	0.1001	4.2285
11	1200	0.1005	1.8435	0.1015	1.6762
12	1200	0.1000	2.1878	0.1011	2.2784
13	1200	0.1001	1.6084	0.1052	1.5139
14	800	0.1046	2.0965	0.1015	2.0523
Result (s)		**8.0979**		**8.0865**	

Table IV shows that IGWO has outperform others optimization algorithm for this test case. This proves that IGWO has better efficiency towards conventional GWO and other identified algorithm.

C. Case III

In this case, the proposed method is applied to IEEE 15 bus test system. The system's single line diagram is as in Fig. 8. The system details on three phase short circuit data can be found in [2]. This system is powered by highly distributed generation network with 15 bus consists of 42 relays and connected by 21 lines.

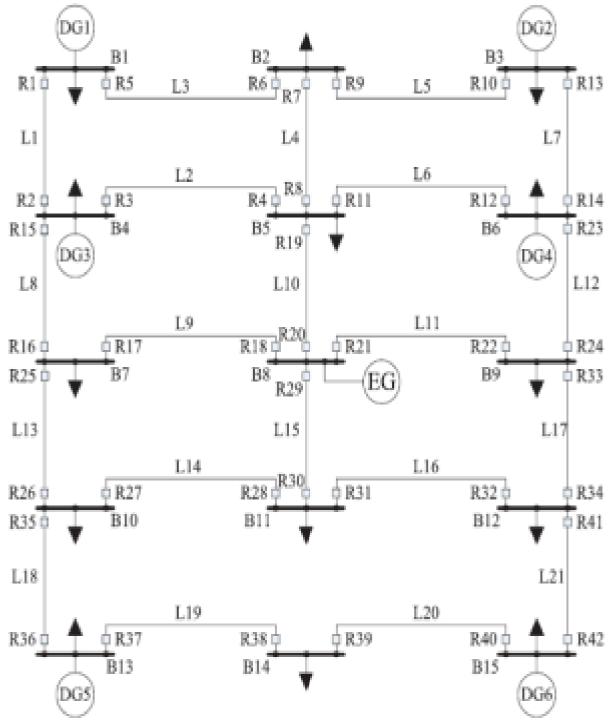


Fig. 8. IEEE 15 Bus Test System.

TABLE V. COMPARISON OF IGWO WITH OTHERS TECHNIQUE FOR CASE III

Method	Objective function (s)
GA-NLP [30]	19.5843
CSA [41]	19.5521
PSO-LP [25]	15.0020
GWO	12.9637
IGWO	12.6446

There are 84 variables with 82 coordination constraints. The TMS bound from X_1 to X_{42} and PS bound from X_{43} to X_{84}. The normal inverse type characteristic is selected. The TMS values are in between 0.1s to 1.1s and the PS value is in between 1.5 to 5. The CTI value is assumed as 0.2s.

Table V shows the comparative results of the IGWO with GA-NLP, CSA, PSO-LP and conventional GWO. From the generated results, it has confirmed the robustness of IGWO in order to solve the optimization problem of overcurrent relay coordination.

TABLE VI. PERFORMANCE COMPARISON FOR GWO AND IGWO FOR CASE III

Criteria	GWO	IGWO
Best	12.9637	12.6446
Worst	82.0985	81.6864
Mean	17.7500	17.6599

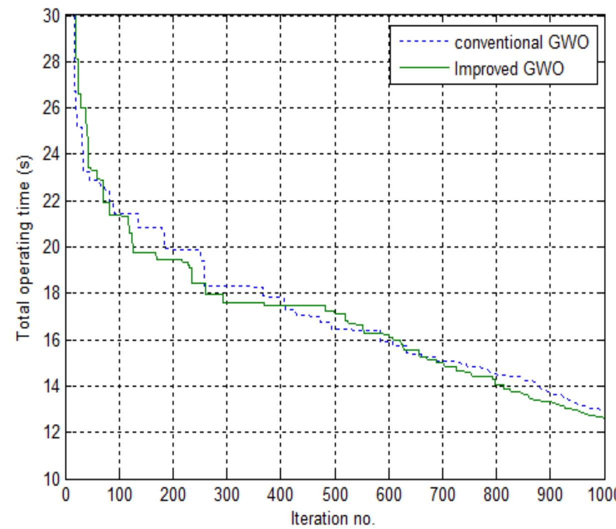


Fig. 9. Conventional Versus Improved GWO for Case III

TABLE VII. OPTIMUM SETTING FOR 15 BUS TEST CASE

Relay no.	GWO		IGWO	
	TMS	PS	TMS	PS
1	0.1013	2.7320	0.1001	2.8114
2	0.1001	1.7947	0.1000	1.5104
3	0.1002	1.5621	0.1048	1.7950
4	0.1018	2.6221	0.1002	3.0701
5	0.1083	1.5487	0.1003	1.5191
6	0.1002	1.5042	0.1004	1.5367
7	0.1013	2.1928	0.1000	2.4252
8	0.1001	2.5192	0.1012	2.3699
9	0.1004	2.0425	0.1030	2.0256
10	0.1003	1.7860	0.1000	1.5244
11	0.1000	2.5195	0.1002	2.5299
12	0.1012	2.8016	0.1008	2.8068
13	0.1003	2.7715	0.1004	2.4412
14	0.1033	1.6329	0.1001	1.6608
15	0.1011	1.6711	0.1000	1.5666
16	0.1007	1.5035	0.1001	1.7755
17	0.1003	1.9094	0.1008	1.5505
18	0.1003	2.0602	0.1130	1.6783
19	0.1001	2.1253	0.1003	2.1283
20	0.1002	1.6312	0.1040	1.5726
21	0.1001	2.4752	0.1002	2.0357

Table VI and VIII provide the optimum setting of TMS and PS respectively. From the tabulated results, it indicates that IGWO has outperformed about 0.3191s faster than the results of GWO. The characteristic of conventional GWO versus IGWO as in fig. 9 which shows that the IGWO has given improved result. Table VII shows the performance comparative results in between GWO and IGWO.

TABLE VIII. OPTIMUM SETTING FOR 15 BUS TEST CASE , CONTINUED

Relay no.	GWO		IGWO	
	TMS	PS	TMS	PS
22	0.1007	1.5455	0.1001	1.5461
23	0.1016	1.5723	0.1000	1.5502
24	0.1000	1.5245	0.1002	1.6656
25	0.1001	1.5056	0.1002	1.5019
26	0.1002	3.1433	0.1002	2.9163
27	0.1057	1.9240	0.1002	2.1045
28	0.1007	1.5607	0.1002	1.5029
29	0.1071	1.6619	0.1017	1.6389
30	0.1001	2.1626	0.1006	2.1199
31	0.1119	1.8264	0.1000	2.0376
32	0.1005	2.9962	0.1008	3.0167
33	0.1000	1.5108	0.1012	1.5459
34	0.1006	1.9720	0.1006	1.5149
35	0.1000	2.0547	0.1002	1.9585
36	0.1003	3.3072	0.1002	3.3271
37	0.1034	2.0512	0.1016	1.7971
38	0.1009	2.4801	0.1002	2.5855
39	0.1086	2.3918	0.1004	2.7681
40	0.1001	2.6924	0.1010	2.5714
41	0.1007	1.5155	0.1003	1.7723
42	0.1001	3.6070	0.1001	3.4527
Result (s)	**12.9637**		**12.6446**	

V. CONCLUSION

This paper proposed IGWO algorithm for optimal coordination setting of the overcurrent relays problem. Some modification has been recommended to improve the exploration ability of the grey wolves. This exploration ability has been proven to improve the conventional GWO convergence characteristic. Three test cases are utilized to confirm the reliability of the IGWO. Comparison results between IGWO, conventional GWO and with other identified algorithm such GA-NLP and CSA indicated that IGWO has improved the convergence performance when applied to the optimization problem of overcurrent relays coordination. In addition, proposed modification has counteracted the argument of randomness exploration activity. As the conclusion, the IGWO is appears to be an efficient and robust optimization algorithm for optimal solution of overcurrent relay coordination problem in electrical network system.

ACKNOWLEDGMENT

This work was supported by Universiti Malaysia Pahang under grant no. RDU1803101.

REFERENCES

[1] M. Alipour, S. Teimourzadeh, and H. Seyedi, "Improved group search optimization algorithm for coordination of directional overcurrent relays," Swarm and Evolutionary Computation, vol. 23, pp. 40-49, 8// 2015.

[2] M. M. Mansour, S. F. Mekhamer, and N. El-Kharbawe, "A Modified Particle Swarm Optimizer for the Coordination of Directional Overcurrent Relays," IEEE Transactions on Power Delivery, vol. 22, pp. 1400-1410, 2007.

[3] A. S. Noghabi, J. Sadeh, and H. R. Mashhadi, "Considering Different Network Topologies in Optimal Overcurrent Relay Coordination Using a Hybrid GA," IEEE Transactions on Power Delivery, vol. 24, pp. 1857-1863, 2009.

[4] T. Amraee, "Coordination of Directional Overcurrent Relays Using Seeker Algorithm," IEEE Transactions on Power Delivery, vol. 27, pp. 1415-1422, 2012.

[5] M. H. Hussain, S. R. A. Rahim, and I. Musirin, "Optimal Overcurrent Relay Coordination: A Review," Procedia Engineering, vol. 53, pp. 332-336, // 2013.

[6] C. W. Rogers, R. Ramaswami, and P. F. McGuire, "Graphical coordination program aids system relaying at Consumers Power," IEEE Computer Applications in Power, vol. 3, pp. 30-35, 1990.

[7] "Computer Representation of Overcurrent Relay Characteristics," IEEE Power Engineering Review, vol. 9, pp. 50-51, 1989.

[8] T. Keil and J. Jager, "Advanced Coordination Method for Overcurrent Protection Relays Using Nonstandard Tripping Characteristics," Power Delivery, IEEE Transactions on, vol. 23, pp. 52-57, 2008.

[9] Y. Lu and J.-L. Chung, "Detecting and solving the coordination curve intersection problem of overcurrent relays in subtransmission systems with a new method," Electric Power Systems Research, vol. 95, pp. 19-27, 2// 2013.

[10] D. Birla, R. P. Maheshwari, and H. O. Gupta, "An Approach to Tackle the Threat of Sympathy Trips in Directional Overcurrent Relay Coordination," Power Delivery, IEEE Transactions on, vol. 22, pp. 851-858, 2007.

[11] L. Cheng-Hung and C. Chao-Rong, "Using Genetic Algorithm for Overcurrent Relay Coordination in Industrial Power System," in Intelligent Systems Applications to Power Systems, 2007. ISAP 2007. International Conference on, 2007, pp. 1-5.

[12] C. W. So, K. K. Li, K. T. Lai, and K. Y. Fung, "Application of genetic algorithm for overcurrent relay coordination," in Developments in Power System Protection, Sixth International Conference on (Conf. Publ. No. 434), 1997, pp. 66-69.

[13] C. W. So, K. K. Li, K. T. Lai, and K. Y. Fung, "Application of genetic algorithm to overcurrent relay grading coordination," in Advances in Power System Control, Operation and Management, 1997. APSCOM-97. Fourth International Conference on (Conf. Publ. No. 450), 1997, pp. 283-287 vol.1.

[14] F. Razavi, H. A. Abyaneh, M. Al-Dabbagh, R. Mohammadi, and H. Torkaman, "A new comprehensive genetic algorithm method for optimal overcurrent relays coordination," Electric Power Systems Research, vol. 78, pp. 713-720, 4// 2008.

[15] P. P. Bedekar and S. R. Bhide, "Optimum coordination of overcurrent relay timing using continuous genetic algorithm," Expert Systems with Applications, vol. 38, pp. 11286-11292, 9// 2011.

[16] D. S. Alkaran, M. R. Vatani, M. J. Sanjari, G. B. Gharehpetian, and M. S. Naderi, "Optimal Overcurrent Relay Coordination in Interconnected Networks by Using Fuzzy-Based GA Method," IEEE Transactions on Smart Grid, vol. PP, pp. 1-1, 2016.

[17] J. C. Bansal and K. Deep, "Optimization of directional overcurrent relay times by particle swarm optimization," in Swarm Intelligence Symposium, 2008. SIS 2008. IEEE, 2008, pp. 1-7.

[18] Z. Moravej, M. Jazaeri, and M. Gholamzadeh, "Optimal coordination of distance and over-current relays in series compensated systems based on MAPSO," Energy Conversion and Management, vol. 56, pp. 140-151, 2012.

[19] H. H. Zeineldin, E. F. El-Saadany, and M. M. A. Salama, "Optimal coordination of overcurrent relays using a modified particle swarm optimization," Electric Power Systems Research, vol. 76, pp. 988-995, 7// 2006.

[20] R. Thangaraj, M. Pant, and K. Deep, "Optimal coordination of over-current relays using modified differential evolution algorithms," Engineering Applications of Artificial Intelligence, vol. 23, pp. 820-829, 8// 2010.

[21] R. Thangaraj, M. Pant, and A. Abraham, "New mutation schemes for differential evolution algorithm and their application to the optimization of directional over-current relay settings," Applied Mathematics and Computation, vol. 216, pp. 532-544, 3/15/ 2010.

[22] M. Singh, B. K. Panigrahi, A. R. Abhyankar, and S. Das, "Optimal coordination of directional over-current relays using informative differential evolution algorithm," Journal of Computational Science, vol. 5, pp. 269-276, 3// 2014.

[23] C. A. Castillo, A. Conde, and M. Y. Shih, "Improvement of non-standardized directional overcurrent relay coordination by invasive weed optimization," Electric Power Systems Research, vol. 157, pp. 48-58, 2018/04/01/ 2018.

[24] J. A. Sueiro, E. Diaz-Dorado, E. Míguez, and J. Cidrás, "Coordination of directional overcurrent relay using evolutionary algorithm and linear programming," International Journal of Electrical Power & Energy Systems, vol. 42, pp. 299-305, 11// 2012.

[25] V. A. Papaspiliotopoulos, T. A. Kurashvili, and G. N. Korres, "Optimal coordination of directional overcurrent relays in distribution systems with distributed generation based on a hybrid PSO-LP algorithm," in MedPower 2014, 2014, pp. 1-6.

[26] X. Chunlin, Z. Xiufen, Y. Rongxiang, and W. Chuansheng, "Optimal coordination of protection relays using new hybrid evolutionary algorithm," in Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on, 2008, pp. 823-828.

[27] A. Ahmarinejad, S. M. Hasanpour, M. Babaei, and M. Tabrizian, "Optimal Overcurrent Relays Coordination in Microgrid Using Cuckoo Algorithm," Energy Procedia, vol. 100, pp. 280-286, 11// 2016.

[28] H. R. E. H. Bouchekara, M. Zellagui, and Ma.Abido, "Optimal Coordination of Directional Overcurrent Relays Using a Modified Electromagnetic Field Optimization Algorithm," Applied Soft Computing.

[29] N. Mancer, B. Mahdad, K. Srairi, M. Hamed, and B. Hadji, "Optimal Coordination of Directional Overcurrent Relays Using PSO-TVAC," Energy Procedia, vol. 74, pp. 1239-1247, 8// 2015.

[30] P. P. Bedekar and S. R. Bhide, "Optimum Coordination of Directional Overcurrent Relays Using the Hybrid GA-NLP Approach," IEEE Transactions on Power Delivery, vol. 26, pp. 109-119, 2011.

[31] D. S. Alkaran, M. R. Vatani, M. J. Sanjari, G. B. Gharehpetian, and M. S. Naderi, "Optimal Overcurrent Relay Coordination in Interconnected Networks by Using Fuzzy-Based GA Method," IEEE Transactions on Smart Grid, vol. 9, pp. 3091-3101, 2018.

[32] Y. Damchi, H. R. Mashhadi, J. Sadeh, and M. Bashir, "Optimal coordination of directional overcurrent relays in a microgrid system using a hybrid particle swarm optimization," in Advanced Power System Automation and Protection (APAP), 2011 International Conference on, 2011, pp. 1135-1138.

[33] A. Sahoo and S. Chandra, "Multi-objective Grey Wolf Optimizer for improved cervix lesion classification," Applied Soft Computing, vol. 52, pp. 64-80, 3// 2017.

[34] M. H. Sulaiman, Z. Mustaffa, M. R. Mohamed, and O. Aliman, "Using the gray wolf optimizer for solving optimal reactive power dispatch problem," Applied Soft Computing, vol. 32, pp. 286-292, 7// 2015.

[35] H. Mee Song, M. H. Sulaiman, and M. R. Mohamed, "An Application of Grey Wolf Optimizer for Solving Combined Economic Emission Dispatch Problems," 2014, vol. 7, 2014.

[36] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," Advances in Engineering Software, vol. 69, pp. 46-61, 3// 2014.

[37] V. N. Rajput, K. S. Pandya, and K. Joshi, "Optimal coordination of Directional Overcurrent Relays using hybrid CSA-FFA method," in Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2015 12th International Conference on, 2015, pp. 1-6.

[38] "IEC standard 60255-4, Single input energizing measuring relays with dependent specified time.," in IEC standard 60255-4, ed: IEC publication 60255-4, 1st ed., 1976.

[39] N. Mittal, U. Singh, and B. S. Sohi, "Modified Grey Wolf Optimizer for Global Engineering Optimization," Applied Computational Intelligence and Soft Computing, vol. 2016, p. 16, 2016.

[40] A. J. Urdaneta, R. Nadira, and L. G. Perez Jimenez, "Optimal coordination of directional overcurrent relays in interconnected power systems," IEEE Transactions on Power Delivery, vol. 3, pp. 903-911, 1988.

[41] G. U. Darji, M. J. Patel, V. N. Rajput, and K. S. Pandya, "A tuned cuckoo search algorithm for optimal coordination of Directional Overcurrent Relays," in Power and Advanced Control Engineering (ICPACE), 2015 International Conference on, 2015, pp. 162-167.

Risk Assessment Method for Insider Threats in Cyber Security: A Review

Nurul Akmal Hashim[1], Zaheera Zainal Abidin[2], Nurul Azma Zakaria[4], Rabiah Ahmad[5]
Information Security, Forensic and Networking Research Group (INSFORNET),
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka Malaysia

A.P. Puvanasvaran[3]
Faculty of Manufacturing Engineering
Universiti Teknikal Malaysia Melaka, Melaka Malaysia

***Abstract*—Today's in manufacturing major challenge is to manage large scale of cybersecurity system, which is potentially exposed to a multitude of threats. The utmost risky threats are insider threats. An insider threat arises when a person authorized to perform certain movements in an organization decides to mishandle the trust and harm the organization. Therefore, to overcome these risks, this study evaluates various risk assessment method to assess the impact of insider threats and analyses the current gaps in risk assessment method. Based on the literature search done manually, we compare four methods which are NIST, FRAP, OCTAVE, and CRAMM. The result of the study shows that the most used by an organization is the NIST method. It is because NIST is a method that combines the involvement between human and system in term of collection data. The significance of this study contributes to developing a new method in analyzing the threats that can be used in any organization.**

Keywords—Insider threats; manufacturing; risk assessment; cyber security; threats; risk

I. Introduction

The industrial revolution (IR) 4.0 for the manufacturing area is mostly based on advances in the areas of autonomous robots, big data, augmented reality, cloud computing, internet of thing and cybersecurity [1]. Malaysian as a dependent nation needs to increase the value chain to become a high-quality manufacturing base using technology to make the country more competitive at regional and global levels. Besides that, IR 4.0 encourages companies to use computerization and data exchange in manufacturing technologies that create smart robot where machines are linked to the internet and to a system that can depict the whole production chain[2].

However, nowadays it shows that cybercrimes cases are reported and increased over than 40%. Organizational security professionals are worried about workers with low-security awareness may provide required information accidentally under the trickery hackers [3]. The insider threat is considered as a part of social engineering, which we also call as unintentional insider threat (UIT). It is worth noting that insider threat about intentional leakage has begun to raise the courtesy of researchers recently [3], [4].

The term insider threat refers to threats originating from people who have been given access rights to an IS and misuse their privileges, thus violating the IS security policy of the organization. Criminology research has extensively studied this kind of behavior, even though it does not always lead to committing a crime. In the same way, attacks can be non-malicious while performing the tasks in an organization like carelessness, lack of knowledge, or intentional circumvention of security. Internal Intrusion Detection System (IDS) protect organizations against insider attacks.

Therefore, to reduce and analyze insider threats is by using risk assessment. Risk assessment is the procedure that evaluates the information system and the security characteristics of information like confidentiality, integrity, and availability [5]. The evaluation is based on related information security technology and management criteria. Through risk assessment, we can understand the security situation and take targeted security measures which control the risk within an acceptable range. The basic risk assessment model is shown in Fig.1.

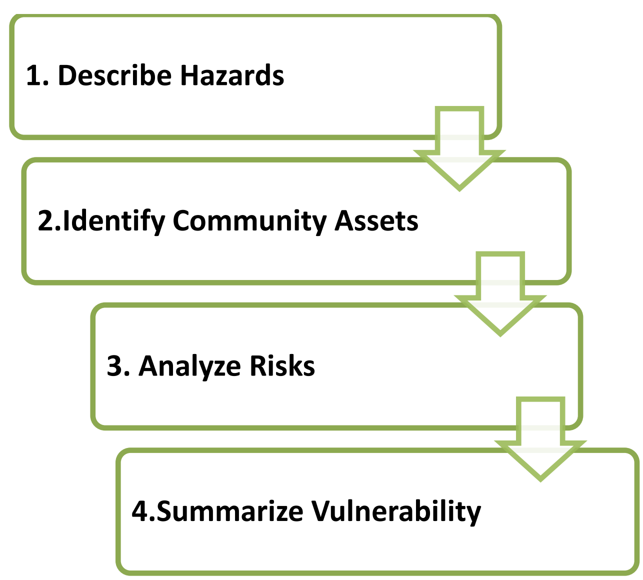


Fig. 1. Risk Assessment Basic Model [6]

Risk assessment considers four factors: hazards, assets, threats, and vulnerabilities. This research focuses on assets, analyzing assets, the relationship between threats and vulnerabilities, and the value of the risk of computing systems[7]. Many security techniques and mechanism have been developed to counter the insider threats such as National Institute of Standards & Technology Special Publication 800-30 (NIST SP 800-30), The Operationally Critical, Threat, Asset and Vulnerability Evaluation (OCTAVE) process, The Facilitated Risk Assessment Process (FRAP), and The Central Risk Analysis and Management Method (CRAMM).

Currently, risk assessment has been applied to almost every aspect of the industry. A risk is defined as the impact on the uncertain target; the impact can be positive or negative [8]. According to Hubbard, risk management includes risk identification, assessment and prioritization, and subsequent reduction, monitoring, and control of negative events [8]. With the joint efforts of scholars and experts, there are several popular risk assessment models that can meet different needs.

Rest of the paper consists of following sections: Section 2 presents the related work that unveils the methods of the risk assessment. Result and Discussion are covered in section 3. Finally, section 4 concludes the paper and discusses future work.

II. Related Works: Review of the Risk Assessment Methods

The studied-on risk assessment method in cybersecurity have been used to identify insider threats will be discussed. Furthermore, an analysis of the related works of the risk assessment method to ease the security condition task is offered.

A. *National Institute of Standards & Technology (NIST)*

The method described in NIST SP800-30 is a combination of quantitative and qualitative. The NIST 800-30 is primarily a model rather than a specialized method [9], [10]. It still contains a complete guide to defining all aspects of an effective risk management plan. It also contains the criteria and processes needed to assess and mitigate risk. It is suitable for better large organizations such as government agencies and large corporations. NIST SP800 supports organizations, CIOs (CIOs), security officers, IT consultants, and anyone who is generally involved with risk management in the organization [11].

The first step in NIST is to identify assets. System characteristics describe the boundaries of the system and the resources and information that make up the system. The characterization system defines the scope of the risk assessment effort, describes the operational authorization (or certification) boundaries, and provides the information necessary to define the risk (eg, hardware, software, system connectivity, and responsible department or support staff). There are two ways to identify an asset [12]. First, system-related information can be applied to describe the IT system and its operating environment. The second method is to use information gathering techniques to solicit information related to the IT system process environment. Common information gathering techniques include questionnaires, live interviews, document review and the use of automated scanning tools. The target asset can be a single or multiple interrelated system. In the latter case, the domain of interest and all interfaces and dependencies must be well defined before applying the method. Fig. 2 below shown a basic NIST step.

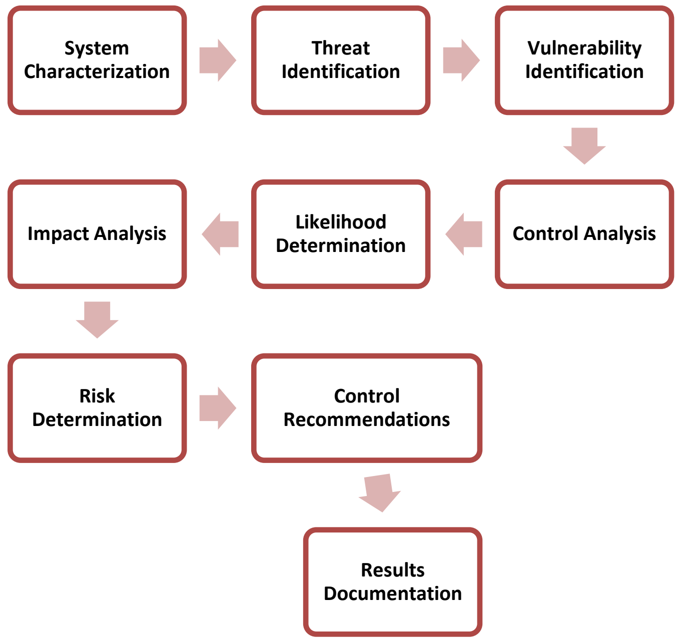


Fig. 2. NIST Basic Model [12]

B. *The Operational Critical, Threat, Asset and Vulnerability Evaluation Process (OCTAVE)*

The OCTAVE method was developed by the Software Engineering Institute (SEI) at Carnegie Mellon University. This approach was established to help organizations identify and assess the risks of information systems, improve their capabilities and protect themselves from these risks [13]. The OCTAVE method consists of a set of rules and a skilled analysis team. The team is made up of people within the organization and is designed to conduct risk assessment procedures. Collect opinions from the analysis team and participants through questionnaires and surveys [15], [16].

Based on the inputs provided, the analysis is done in a structured and organized manner. There are several pre- and post-assessment activities. The risk assessment process consists of three main steps and eight of these three steps.

The OCTAVE method can be extended to the OCTAVE standard, which is designed to meet the requirements of various situations. For example, a standard set can be applied to large organizations to small organizations. But the method is still the same and can be described as four main phases. The OCTAVE basic model is shown in the Fig. 3 below.

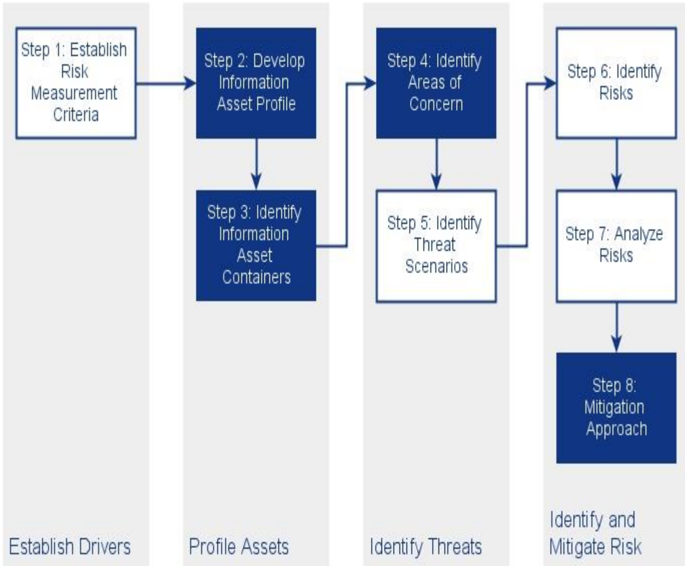


Fig. 3. OCTAVE Basic Model [13]

C. The Facilitated Risk Assessment Process (FRAP)

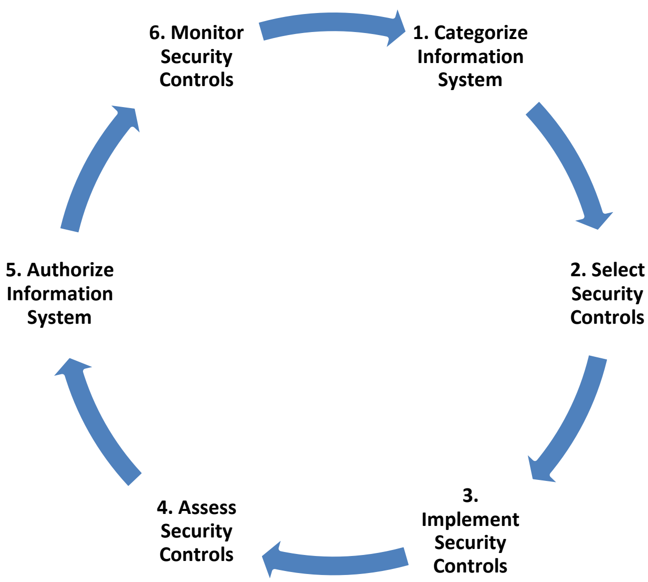


Fig. 4. FRAP Basic Model [18]

The Facilitating Risk Assessment Program (FRAP) was established by Thomas Peltier [17]. Peltier aims to implement risk management techniques in a cost-effective manner to adapt to the rapid development of the business sector. Peltier also emphasizes the involvement of employees in the organization, rather than the advice of external experts. Since the model is designed to prioritize time-cost efficiency, the program includes only the pre-FRAP meeting, the FRAP meeting and the FRAP meeting. In the pre-FRAP meeting phase, the goal is to introduce participants to FRAP and announce the procedures and goals of the meeting. Once the participants reach an agreement, they can hold a FRAP meeting. There are two steps involved during the FRAP meeting. The first step is to browse the logistics, introduce the entire team and briefly repeat what was discussed in the pre-FRAP meeting. The scope statement will then be exposed. In the second step, the FRAP team will review the elements to be reviewed, such as integrity, confidentiality and availability. The team also identifies threats, issues, and any other issues that may pose a vulnerability to the system. Next, the team will recommend controlling these vulnerabilities. After the FRAP meeting, the business manager, project leader and moderator will hold a meeting after the FRAP meeting and complete the action plan. The deliverables for this meeting include a summary of threats and existing controls, as well as a final report. The basic FRAP cycle model is shown in Fig. 4 below.

D. The Central Risk Analysis and Management Method (CRAMM)

The Central Computer and Telecommunications Authority (CCTA) Risk Analysis and Management Method (CRAMM) was developed by the British government in 1985. This tool has been developed and has been commercialized by Insight Consulting [19]. CRAMM is a qualitative tool that provides methods, calculations, and reports for security risk assessment.

The method and tool were developed mainly for application in large-scale organizations, but can be also applied to SMEs [20]. CRAMM can also be used to (a) Justify investment decisions in the security of information systems and networks, based on measurable results and (b) demonstrate the compatibility of the organizations' information systems with the British standard during an auditing process. CRAMM consists of five phases which shown in the Fig. 5.

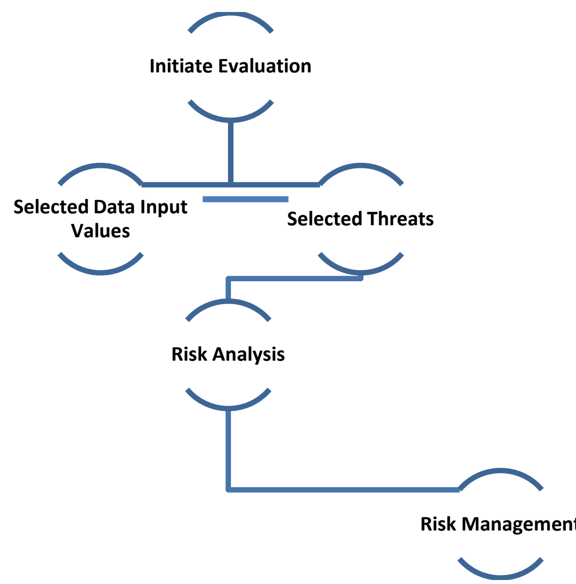


Fig. 5. CRAMM Basic Model [20]

III. Discussion

In general, OCTAVE and CRAMM methods are qualitative methods while FRAP is quantitative. NIST method is a combination of qualitative and quantitative types which is more dynamic and suitable for an organization. This makes the NIST model suitable for quantitative or qualitative research.

NIST risk assessment method is the most well-formed method. Each step has a specific target and enumerates several approaches to facilitate the procedure. Unlike the OCTAVE, CRAMM and the FRAP method, NIST method's collection to the data is not limited to participants' knowledge; it also includes conclusions and discoveries mentioned in other related documentation.

Furthermore, OCTAVE, CRAMM and FRAP merely offer descriptions of each step; while for the NIST method, each step enumerates all the possible approaches to process the data. On the other hand, the OCTAVE and FRAP method are usually applied to the business area while CRAMM specifically for an aviation area. Especially for the FRAP method, the author of the FRAP method, explicitly stated that the FRAP method is not designed to assess the compliance of security requirements.

TABLE I. COMPARISON AMONG SEVERAL RISK ASSESSMENT METHODS

Risk Assessment Methods	References	Types	Approach Phases	Resource Required
NIST	[1], [4], [12], [11], [21], [10], [22], [23]	Qualitative and Quantitative	• System characterization • Threat identification • Vulnerability Identification • Control analysis • Likelihood Determination • Impact analysis • Risk Determination • Control Recommendations • Result Documentation	Non-government organization
OCTAVE	[13], [14], [15], [20], [24]	Qualitative	• Profile threats • Identify infrastructure vulnerability • Develop a security strategy and plan	Internal and non-expert
FRAP	[16], [17], [25], [26]	Quantitative	• Pre-FRAP meeting • FRAP Session • Post-FRAP Process	Internal Manager
CRAMM	[18], [19]	Qualitative	• Asset Identification • Threat and vulnerability assessment • Countermeasure selection and recommendation	Qualified and experienced participant

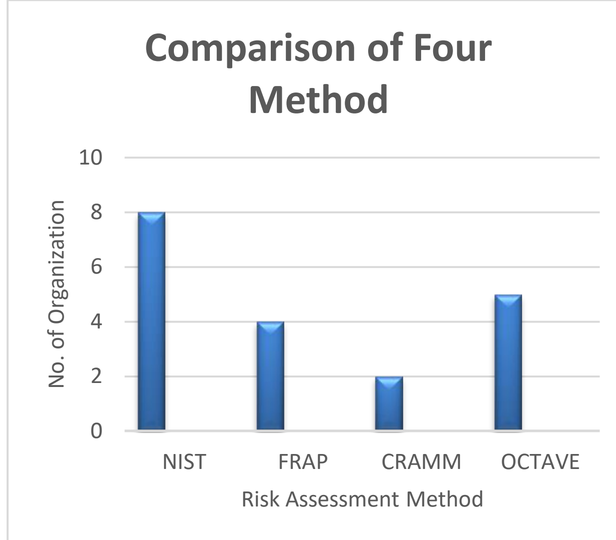


Fig. 6. Graph of four method

Both the FRAP and OCTAVE method is implemented to meet the business need and requires less time and resources. As mentioned earlier, the OCTAVE method has eight steps and needs knowledge from three levels – senior management, operational area management and staff. The FRAP method only has a pre-Frap meeting, FRAP session and post-FRAP discussion, which can be accomplished by the FRAP team in one day. Obviously, the OCTAVE method is more complicated than the FRAP method and requires more people's corporation.

In a word, the OCTAVE method is a workshop-oriented method and requires the participation from a different department. The FRAP method is designed for business analysis instead of a security assessment. The OCTAVE, CRAMM and FRAP is largely dependent on the participants' knowledge. As for the NIST method, the risk assessment process is refined into nine steps. Each step has a clear goal and all the possible approaches to accomplish the goal, which alleviate the bias brought by merely depend on participants' or security evaluator's knowledge.

The differences between all four methods have been simplified in Table I.

Figure 6 shows the comparison of the used method in the industry. Based on the graph it shows that NIST has 8 number of an organization has been used. Compared with FRAP 4 organization, CRAMM 2 organization and OCTAVE 5 organization. Findings of this study indicate that the NIST method more famous and well known used in an organization for risk assessment.

NIST method allows organizations to individually assess threats most relevant to their operations "and to develop a risk-based approach to resource allocation". It enables organizations to express their insider threat management efforts in terms of critical assets (identify); implemented controls and safeguards (protect); manifested threats (detect); formulated incident response strategies (respond); and business continuity plans (recover).

Therefore, the NIST method provides the most complete and scientific approach among all the methods.

IV. Conclusion and Future Works

Several case studies have been made to provides a risk-based detection method for insiders threats. It is not only to understand possible threats, but also help reduce overhead in the unified monitoring process. The results showed that the NIST method is well accepted in many organizations due to the systematic and convincing risk assessment planning. Besides this method is easy operative and practical. The framework can be improved further by assigning users to different classes according to their privileges and assigning different threshold values to each class.

Acknowledgment

The authors also would like to acknowledge Universiti Teknikal Malaysia Melaka. We also would like to thank the funding of this TRGS research grant: TRGS/1/2016/FKP-AMC/01/D00005 for funding this research.

References

[1] Y. Cherdantseva et al., "A review of cyber security risk assessment methods for SCADA systems," Comput. Secur., vol. 56, pp. 1–27, 2016.

[2] Z. Yunos, R. Ahmad, and N. A. Mohd Sabri, "A Qualitative Analysis for Evaluating a Cyber Terrorism Framework in Malaysia," Inf. Secur. J., 2015.

[3] L. Xiangyu, L. Qiuyang, and S. Chandel, "Social Engineering and Insider Threats," in 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2017.

[4] V. Morgagni, N. Nostro, A. Ceccarelli, and F. Brancati, "Insider Threat Assessment : a Model-Based Methodology," pp. 3–12, 2014.

[5] Z. Lai, Y. Shen, and G. Zhang, "A security risk assessment method of website based on threat analysis combined with AHP and entropy weight," in Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, 2017.

[6] Y. Y. Haimes, Risk Modeling, Assessment, and Management, Third Edition. 2008.

[7] E. Zio, "The future of risk assessment," Reliab. Eng. Syst. Saf., vol. 177, no. March, pp. 176–190, 2018.

[8] D. W. Hubbard, The Failure of Risk Management: Why It's Broken and How to Fix It. 2009.

[9] R. K. Abercrombie, F. T. Sheldon, K. R. Hauser, M. W. Lantz, and A. Mili, "Risk assessment methodology based on the NISTIR 7628 guidelines," in Proceedings of the Annual Hawaii International Conference on System Sciences, 2013.

[10] National Institute of Standards and Technology, "NIST SP 800-53A, Assessing Security and Privacy Controls in Federal Information Systems and Organizations," NIST Spec. Publ., 2014.

[11] J. Kouns and D. Minoli, Information Technology Risk Management in Enterprise Environments: A Review of Industry Practices and a Practical Guide to Risk Management Teams. 2010.

[12] National Institute of Standards and Technology, "BEST PRACTICES IN CYBER SUPPLY CHAIN RISK MANAGEMENT," Nist, 2016.

[13] R. A. Caralli, J. F. Stevens, L. R. Young, and W. R. Wilson, Introducing OCTAVE Allegro : Improving the Information Security Risk Assessment Process. 2007.

[14] M. T. Jufri, M. Hendayun, and T. Suharto, "Risk-assessment based academic information System security policy using octave Allegro and ISO 27002," in Proceedings of the 2nd International Conference on Informatics and Computing, ICIC 2017, 2018.

[15] D. C. Felegeanu et al., "A combined method for the analysis and assessment of risks and industrial safety," Environ. Eng. Manag. J., 2016.

[16] M. Masky, S. S. Young, and T. Y. Choe, "A novel risk identification framework for cloud computing security," in 2015 IEEE 2nd International Conference on InformationScience and Security, ICISS 2015, 2016.

[17] T. Peltier, "Information security risk analysis," Philos. Trans. A. Math. Phys. Eng. Sci., 2005.

[18] T. R. Peltier, "Implementing an information security awareness program," Inf. Syst. Secur., 2005.

[19] Z. Yazar, "A Qualitative Risk Analysis and Management Tool - CRAMM," 2002.

[20] T. Yang, E. D. Berger, S. F. Kaplan, and J. E. B. Moss, "CRAMM : Virtual Memory Support for Garbage-Collected Applications," in Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation - OSDI'06, 2006.

[21] A. Syalim, Y. Hori, and K. Sakurai, "Comparison of risk analysis methods: Mehari, magerit, NIST800-30 and microsoft's security management guide," Proc. - Int. Conf. Availability, Reliab. Secur. ARES 2009, pp. 726–731, 2009.

[22] V. Jovanovic and J. K. Harris, "Systems and software assurance - A model Cyber Security course," in 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2016 - Proceedings, 2016.

[23] National Institute of Standards and Technology, "NIST SP 800-37 Revision 1 Guide for Applying the Risk Management Framework to Federal Information Systems. A Security Life Cycle Approach," p. 102, 2010.

[24] A. Sarkheyli and N. Binti Ithnin, "Improving the current Risk analysis techniques by study of their process and using the human body's Immune System," 2010 5th Int. Symp. Telecommun. IST 2010, pp. 651–656, 2010.

[25] M. Tseng, C. Byrne, K. A. Evers, and M. B. Daly, "Dietary intake and breast density in high-risk women: a cross-sectional study.," Breast Cancer Res., vol. 9, no. 5, pp. 12–15, 2007.

[26] A. Shameli-Sendi, R. Aghababaei-Barzegar, and M. Cheriet, "Taxonomy of information security risk assessment (ISRA)," Comput. Secur., vol. 57, pp. 14–30, 2016.

BAAC: Bangor Arabic Annotated Corpus

Ibrahim S Alkhazi
College of Computers & Information Technology
Tabuk University, Tabuk, Saudi Arabia

William J. Teahan
School of Computer Science Bangor University
United Kingdom

***Abstract*—This paper describes the creation of the new Bangor Arabic Annotated Corpus (BAAC) which is a Modern Standard Arabic (MSA) corpus that comprises 50K words manually annotated by parts-of-speech. For evaluating the quality of the corpus, the Kappa coefficient and a direct percent agreement for each tag were calculated for the new corpus and a Kappa value of 0.956 was obtained, with an average observed agreement of 94.25%. The corpus was used to evaluate the widely used Madamira Arabic part-of-speech tagger and to further investigate compression models for text compressed using part-of-speech tags. Also, a new annotation tool was developed and employed for the annotation process of BAAC.**

***Keywords*—Component; arabic language; corpus; annotated corpora; analysis results**

I. Background and Motivation

The Arabic language "العربية" is acknowledged to be one of the most largely used languages, with 330 million people using the language as their first language, as shown in Table 1, plus 1.4 billion more using it as a secondary language [1]. The majority of the speakers are located across twenty-two nations, primarily in the Middle East, North Africa and Asia, and the United Nations considers the Arabic language as one of its five official languages. The Arabic language is part of the Semitic languages that includes Tigrinya, Amharic, Hebrew, etc., and shares almost the same structure as those languages. It has 28 letters, two genders – feminine and masculine, as well as singular, dual and plural forms. The Arabic language has a right-to-left writing system with the basic grammatical structure that consists of verb-subject-object and other structures, such as VOS, VO and SVO [2]–[4].

TABLE I. The Most Universally Used Languages

Rank	**Language**	**Users (millions)**
1	Mandarin	1051
2	English	508
3	Hindi	497
4	Spanish	392
5	Arabic	330
6	Russian	277
7	Bengali	211
8	Portuguese	191
9	Malay	159
10	French	129

The non-colloquial written text for the Arabic language can be divided into two types: Classical Arabic and Modern Standard Arabic [5]–[8]. The Classical Arabic (CA) epoch, as shown in Figure 1, is usually measured from the sixth century which is the start of Arabic literature. It is the language of the Holy Quran, the 1,400-year-old primary religious book of Islam with 77,430 words [9] and other ancient Islamic books from that era, such as the Hadith books [10]. With the beginning of journalism and the spread of literacy in the eighteenth century came Modern Standard Arabic or MSA. MSA is the language of current printed Arabic media and most Arabic publications.

Most Arabic natural language processing (NLP) tasks perform better for MSA [11]. One example of those tasks is parts-of-speech tagging (POS) of the Arabic language as reported in [10], [12], [13], where the performance of the taggers is best when tagging MSA text. The reason for the variation in performance between MSA and CA is that most Arabic language NLP systems were trained using MSA text [14], [15]. More effort is currently being made, such as the creation of manually annotated CA corpora [16] and the evaluation of different Arabic POS taggers on CA text by Alosaimy and Atwell [12], to fill this gap in research.

The term corpus can be defined as a computerised set of genuine texts or discourses provided by language speakers and saved in a machine-readable form [17]–[20]. Xiao [21] argues that a corpus is not a randomly collected collection of texts nor an archive, but a file that manifests four essential aspects: a corpus is a set of (1) machine-readable (2) genuine texts (that includes transcripts of spoken data) that are (3) tested to be (4) representative of a specific or a group of languages.

يُبكون من تبكي السماواتِ يومَه ومن قدِ بكته الأرض فالناس أكمد
وهل عدلت يوما رزية هالك رزيـــة يـــوم مـــات فيـــه محمد

Fig. 1. A Classical Arabic Poem.

Corpora play a significant factor in the development, improvement and evaluation of many NLP applications such as machine translation [22], [23], part-of-speech tagging [24] and text-classification [14], [23]. The design of any corpus depends on its intended applications [25]. Some corpora are for general use and can be utilised in many applications, and others may serve a specific purpose, such as building dictionaries or examining the language of a specific author or duration of time [10].

There are several kinds of annotations which could be applied to corpora, and each annotation is usually designed to handle a certain aspect of the language [26]. One type of corpora annotation is the structural annotation of the corpus by attaching descriptive information about the text, like mark-ups that specify the boundaries of the sentence, section and chapter, or a header file that names the author of the text or adds information about participants, such as the age and gender. Another type of annotation is the morphological annotation, where information about the text, like the stems or root based in a language like Arabic, is added to the corpora. This research applies the most common type of corpora annotation, which is POS tagging of the text [26], where a tag, such as a noun, verb or particle is combined with each term in the corpus, and the number of tags used in the annotation varies from a few to 400 tags or more [27].

Based on the type of text and creation purposes, the corpus can be categorised into six categories: Raw Text Corpora, Annotated Corpora, Lexicon Corpora, Annotated Corpora and Miscellaneous Corpora. Examples of corpora for the Arabic language are provided below.

1) Raw Text Corpora can be Divided into:

A. ***Monolingual corpora***, such as the BACC [28], Ajdir Corpora [29], the King Saud University corpus of Classical Arabic [30], Alwatan [31], Tashkeela [32] and the Al Khaleej Corpus [33]. The monolingual corpora consist of a raw text written in a single language.

B. ***Multilingual corpora***, also known as comparable corpora or parallel corpora, are corpora that are written in two or more languages. Multilingual corpora, such as the UN corpus [34] which is the most important and widely known free corpus [23], Corpus A [22], the Hadith Standard Corpus [35], [36] and MEEDAN Translation Memory [37], are widely used in NLP fields such as machine translation [22], [23].

C. ***Dialectal Corpora***, where the corpus is written in a specific language dialect, such as the Bangor Twitter Arabic Corpus for the Egyptian, Gulf, Iraqi, Maghrebi and Levantine Arabic dialects [38]. Such corpora are used in fields such as text-classification [14].

D. ***Web-based corpora***, such as the KACST Arabic Corpus [39], the Leeds Arabic Internet Corpus [40] and the International Corpus of Arabic [41], where the corpora are only accessible online by an inquiry interface and the corpora cannot be downloaded.

2) The second type is Lexicon corpora, that can be divided into:

A. ***Lexical Databases***, such as the BAMA 1.0 English-Arabic Lexicon [42] and the Arabic-English Learner's Dictionary [43].

B. ***Words Lists*** such as the Word Count of Modern Standard Arabic [43] and the Arabic Wordlist for Spellchecking [44], [45].

These types of corpora act like a vocabulary or a list of words and can be employed by linguists to study many aspects of a language or combined with the lexicons of systems, like spell checking applications, to improve their performance [23].

*3) **Miscellaneous Corpora**, such as Speech Corpora [46], Handwriting Recognition Corpora [47], are beneficial for a number of NLP correlated tasks such as plagiarism detection [48], speech recognition systems [46] and question answering [49].*

*4) **Annotated corpora** are essential for the development of many NLP systems, such as part-of-speech tagging [24], text parsing [50]. Annotated corpora are divided into:*

A. ***Named Entities Corpora*** such as JRC-Names [51] and ANERCorp [52]. Most corpora of this type include the names of persons with the company or organisation name and the locations.

B. ***Error-Annotated Corpora***, such as the KACST Error corpus [53], is a beneficial resource for systems such as spelling correction and machine translation corrected output [54].

C. ***Miscellaneous Annotated Corpora***, such as the OntoNotes corpus [55] and the Arabic Wikipedia Dependency Corpus [56] which are semantically annotated corpora [55].

D. ***Part-of-Speech (POS)*** tagged corpora are an important resource for the training and development of POS systems [24]. Some of the existent resources will be presented in detail in the existing resources section below.

POS annotated corpora are essential for the development of many NLP systems, such as part-of-speech tagging [24], statistical modelling [57] and tag-based compression which provides more effective compression for Arabic text than word or character-based compression methods [13]. The lack of such resources limits some researchers from progressing further in their efforts. The limited availability of some existing annotated corpora and the cost of acquiring others are one of the main reasons that contribute to resource scarcity. Several efforts have been made to overcome the lack of resources [12], [16], [20].

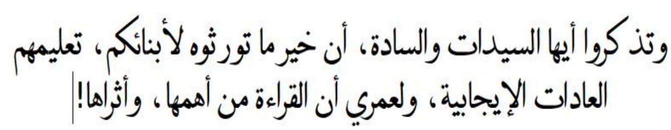

وتذكروا أيها السيدات والسادة، أن خير ما تورثوه لأبنائكم، تعليمهم
العادات الإيجابية، ولعمري أن القراءة من أهمها، وأثراها!

Fig. 2. A Social News from Press Sb-corpus [28] in MSA text.

There exist some annotated corpora for the Arabic language that cannot be utilised by many researchers, such as the tag-based text compression research applied by Alkhazi, Alghamdi and Teahan [13] due to availability, and cost issues, such as the Arabic Treebank corpus [58]. Other resources are designed to be used for particular research or annotated using a distinctive tagset produced for an explicit purpose. The Qur'anic Arabic Dependency Treebank is one example where the text is written in CA text and the corpus uses a tagset which is designed to tag CA text using traditional Arabic grammar [16], [22]. This need for annotated corpora, which are necessary for the development of many NLP systems, provided the motivation to create a manually annotated corpus for the Arabic language.

This research produces a manually annotated POS tagged corpus that is written in MSA. The tagset used in the new corpus was suggested by Alkhazi, Alghamdi and Teahan [13]; further details about the tagset will be discussed in the annotation tagset section (section III-B), and the annotation process follows the annotations guidelines prescribed by Maamour [59] .

II. Existing Resources

In 2001, the Linguistic Data Consortium (LDC) published the first versions of the Penn Arabic Treebank (ATB) [58]. This resource is widely used in many Arabic NLP applications such as the training of POS taggers, like the Madamira Arabic POS tagger [60] and the Stanford Arabic POS tagger [3]. The corpus consists of three parts with a total of 1 million annotated words. The first part v2.0 was a newswire text written in Modern Standard Arabic and consisted of 166K terms acquired from the Agence France Presse corpus. The second part was obtained from the Al-Hayat corpus which was distributed by Ummah Arabic News Text and consists of 144K [58]. The last part of the ATB corpus, part 3 v1.0, as shown Figure 3, is a newswire text obtained from the An-Nahar corpus and consists of about 350K morphologically annotated words. For non-members of the LDC, the cost of acquiring any part of the ATB corpus exceeds several thousand US dollars which prevents access to researchers with a limited budget [57], [58].

Khoja [61]–[63] has published a 50,000 terms manually annotated POS tagged corpus written in MSA text. According to the author, the corpus is divided into two parts; the first part is a newspaper text consisting of 1,700 terms that are manually tagged using a tagset that differentiates between the three moods of the verb and case structures of the noun [64]. The second part of the corpus is tagged using a simple tagset that includes only the following POS tags: noun, verb, particle, punctuation or number [62]. However, access to this resource was not provided.

Another annotated corpus was published by Mohit [56]. The AQMAR Arabic Wikipedia Dependency Tree Corpus is a manually annotated corpus that contains 1262 sentences collected from ten Arabic Wikipedia articles and the 36K terms of the corpus are manually annotated using the Brat annotation tool [56]. The ten articles were annotated for named entities beforehand [65]–[67] and cover topics such as Linux, Internet, Islamic Civilisation, Football, etc. The tagset used in this corpus contains a small number of tags and therefore cannot be used for the research concerning tag-based text compression.

The Columbia Arabic Treebank (CATiB) [27] is another manually annotated Treebank corpus that consists of newswire feeds, from the year 2004 to 2007 and written in MSA. The corpus was initially tokenized and then POS tagged by the MADA&TOKAN toolkit [15], [27]. The TrEd annotation interface [68] was utilised in the annotation process. The number of tags used by CATiB is relatively small as it consists only of six POS tags, NOM, PROP, VRB, VRB-PASS and PRT, where each tag comprises a group of subtags, for example, the tag "NOM" can be used to tag nouns, adverbs, pronouns and adjectives.

III. BAAC: The Bangor Arabic Annotated Corpus

The goal of this annotated corpus is to contribute by filling the gap created by the scarcity of freely available Arabic resources, manually annotated POS tagged corpora in particular, which is caused by the lack of availability and cost issues. Another goal is to provide a new resource required by many kinds of research, such as the ongoing tag-based text compression research conducted by Alkhazi, Alghamdi and Teahan [13], where the only annotation required at this stage is POS tags. The tagset used to annotate the new corpus is the same as used by the Madamira Arabic tagger, for reasons that will be discussed in the annotation tagset section (section B). Since the Madamira Arabic POS tagger is trained by the Arabic Treebank corpus [13], [14], and that corpus is written in MSA, the newly annotated corpus must also be written in MSA.

A. *The Data Source.*

The data source for the new corpus is the Press sub-corpus from the BACC corpus [28]. The BACC corpus was created originally to test the performance of various text compression algorithms on different text files. The results of the text classification performed by Alkhazi and Teahan [14] reveal that the Press sub-corpus is 99% written in MSA, as shown in Figure 2. According to the authors, the sub-corpus is a newswire text consisting of 51K terms, gathered from various news websites between 2010 and 2012 and covers many topics such as political and technology news.

B. *The Annotation Tagset.*

```
<Annotation id="202" type="word">
  <Feature name="lookup-word">AlvAnwyp</Feature>
  <Feature name="comment">ADJ_SHOULD_BE_NOUN</Feature>
  <Feature name="selection">Annotation206</Feature>
</Annotation>
```

Fig. 3. A sample POS tag from the ATB Part 3 v 1.0.

The tagset used in the BAAC corpus is the same as used by the Madamira tagger [60], which was used initially by the MADA tagger [15]. The tagset is the subset of the English tagset which was presented with the English Penn Treebank and consists of 32 tags and was initially proposed by Diab, Hacioglu and Jurafsky [69]. The experiments conducted by Alkhazi, Alghamdi and Teahan [13] have concluded that the quality of tag-based compression varies from one tagset to another. The different tagsets, some of which are shown in Table 3, were used to compress MSA text using POS tags, and tag-based compression using the Madamira tagset outperforms other tagsets such as Stanford [70] and Farasa [71]. Since one of the main goals of creating a gold-standard POS annotated text is to investigate the effect of manual annotation on the tag-based text compression, as described below in the experiments, therefore, the Madamira tagset, which outperformed other tagsets and consists of only 32 tags that are shown in Table 2, is used to annotate the BAAC POS tag and to create the ground-truth data which will be used later for training and evaluation purposes.

TABLE II. THE AGREEMENTS, DISAGREEMENTS AND BSERVED AGREEMENT

Tag	Agreements	Disagreements	Observed Agreement %
noun	23570	529	97.80
verb	5714	44	99.24
prep	5574	10	99.82
adj	4632	1235	78.95
noun_prop	2272	520	81.38
conj_sub	1534	17	98.90
conj	1148	79	93.56
pron_rel	992	37	96.40
pron_dem	767	11	98.59
noun_quant	574	1	99.83
part_neg	498	2	99.60
pron	367	6	98.39
adv	166	195	45.98
adj_comp	265	15	94.64
noun_num	252	7	97.30
part_verb	221	0	100.00
verb_pseudo	203	0	100.00
adj_num	156	26	85.71
adv_interrog	25	111	18.38
adv_rel	83	3	96.51
abbrev	60	2	96.77
part_restrict	59	16	78.67
part	25	27	48.08
pron_interrog	19	30	38.78
part_focus	14	9	60.87
part_interrog	22	0	100.00
part_fut	12	0	100.00
part_voc	10	0	100.00
part_det	8	2	80.00
interj	2	0	100.00
Total	**49244**	**2934**	**94.38%**

C. Automatic POS Tagging.

Madamira [60] was utilised to automatically tag the corpus by POS. The manual annotation process of the BAAC corpus followed annotation guidelines proposed by Maamouri [72] for annotating POS tags. All the previous corrections that are made to a tag are shown to the annotators during the process of annotation, as illustrated in section III-E, and the Madamira tagset used to annotate this corpus applies the criteria proposed by the author.

D. The Annotation Tool.

Most existing tools, such as TrEd tool [68], [73] which was used in the annotation of The Prague Dependency Treebank, are developed to annotate Treebank types of corpora, such as dependency trees corpora, that contain other information about the term, such as the gloss or a comment from an annotator, as shown in Figure 3. As mentioned earlier, the first stage of the BAAC annotation process will only add the POS tags to the corpus. Other linguistic information, such as the structural annotation, will be adapted in future work, therefore, the tool which will be used to manually annotate this corpus will only annotate POS tags. During the preparation for the annotation process, many constraints arose and defined four requirements that had to be met by the annotation tool. First, as the annotators are native Arabic speakers, a well-detailed Arabic translation of the tagset was provided with examples during the annotation process. Second, the software used for the annotation had to comply with the hardware and software requirements of the computer used to perform the annotation. Thirdly, the annotation tool, as shown in Figure 4, had to be executed on different operating systems, therefore, the tool was designed to be portable. Finally, online backing up procedures with the ID of the annotators was done to ensure the safety of the data.

TABLE III. DIFFERENT ARABIC TAGSETS

Term	Madamira Tag	Stanford Tag	Farasa Tag
الادارة	noun	DTNN	NOUN-FS
ترحب	noun_prop	VBP	E/ES/SV
بالتزام	verb	NN	NOUN-MS
الامين	noun	DTNN	NOUN-MS
العام	noun	DTJJ	ADJ-MS
بزيادة	adj	NN	NOUN-FS
عنصر	noun	NN	NOUN-MS
الميزانية	noun	DTNN	NOUN-FS
العادية	noun	DTJJ	ADJ-FS
لمكتب	noun	NN	NOUN-MS
الامم	noun	DTNN	NOUN-MP
المتحدة	noun	DTJJ	ADJ-MP

The previous requirements were met by developing a new annotation tool. First, a detailed Arabic translation of the tagset, which was obtained from Alrabiah [10] and then examined by Arabic specialists, was coded in the annotation tool as shown in Figure 4. The annotation tool also offers examples of the tag if required by the annotator. To comply with the hardware requirements and reduce memory dependency, the tool loads only one sentence to be modified at a time. To follow the Maamouri [72] annotation guidelines, the tool also displays the history of annotation by showing two types of modifications, the original tag assigned by the Madamira tagger and any tag chosen by previous annotators, if they exist. A current status of the annotation process is also displayed to the annotator, such as the number of annotated tags in the current session and the number of modified tags in the total document. The Java programming language was used to develop the annotation tool, and therefore, the tool can be executed on different operating systems. The tool also provided online backing up procedures each time the annotator modified a tag to eliminate any data loss.

E. Data Preparation.

After using Madamira [60] to automatically POS tag the corpus, a copy of the corpus was given to each annotator. Each copy was split into batches of documents that have 10-20 sentences and the ID of the annotator was coded with each batch to be used later in the evaluation section. The two annotators, who are native Arabic speakers and postgraduate students in Arabic Studies, started working to manually annotate the corpus on a full-time basis in two stages.

In the first stage of the annotation process, the annotators were required to work on-site to resolve any issues with the annotation tool and the annotation of the corpus was completed using the facilities provided by Tabuk Public Library. When the annotation process was finished, the two versions were evaluated and the Inter Annotator Agreement was calculated using two metrics, as will be discussed below in the BAAC evaluation section. The differences between the two versions were examined and adjusted off-site by a third annotator, who is a native Arabic speaker and PhD candidate student in Arabic Studies, to produce a final version of the corpus. The total time needed to annotate the corpus was two months – three weeks for the first stage and the rest for the final stage.

IV. BAAC Evaluation

The quality of the annotated corpus affects the quality of the NLP application that utilises it. For instance, Reidsma and Carletta [74] has illustrated that the errors produced by machine learning tools are the same errors made by the annotators of the corpus that was used for training those tools. Two metrics were used to evaluate the quality of the BAAC, the Kappa coefficient [75] to calculate the inter-annotator agreement (IAA) among the two annotators and a direct percent agreement for each tag [76]. Using the data in Table 4, the obtained Kappa value is 0.956, which is recognised as perfect according to Landis and Koch [77]. The total observed agreement from Table 2, which displays the number of agreements and disagreements of different tags between the two annotators in a reverse frequency order, is 94.25%. Taking the number of tag occurrences into consideration, Table 2 shows that the tag verb or 'فعل' has the highest agreement between the annotators with 99.24% agreement. It also shows that the annotators agreed only 25 times out of 136 (18%) on the tag 'adv_interrog' or 'حال'. Also, the annotators agreed only on (45.98%) on the tag 'adv', and (38.78%) on the tag 'pron_interrog'. The reasons for such variation between the annotators were:

TABLE IV. The Bacc Agreement Table

	abbrev	adj	adj_comp	adj_num	adv	adv_interrog	adv_rel	conj	conj_sub	interj	noun	noun_num	noun_prop	noun_quant	part	part_det	part_focus	part_fut	part_interrog	part_neg	part_restrict	part_verb	part_voc	prep	pron	pron_dem	pron_interrog	pron_rel	verb	verb_pseudo	Total
abbrev	59								1															2							62
adj		4363	1	1							247		22											1					4		4639
adj_comp		7	260	1							4		2																6		280
adj_num				142								12	2																		156
adv		92		12	108		1	6	67		63		3	1										8							361
adv_interrog		106		6		9		1	3		7																		4		136
adv_rel		8					74																				1				83
conj								1148																							1148
conj_sub								52	1455																			32			1539
interj										2																					2
noun	1	1151		4	42		4	5	18		22762	2	98		1								1					4	41		24134
noun_num		3		15							5	235	1																		259
noun_prop		166	4	1	1			1	1		450	3	2121											11					36		2795
noun_qua		1									2			57																	576

nt														3																	
part					8			2	1		1		1		23						16										52
part_det		1									1					8															10
part_focus								9									14														23
part_fut																		12													12
part_interrog																			22												22
part_neg		1						1												498											500
part_restrict																					58			1							59
part_verb																						221									221
part_voc											1												9								10
prep		18						1			6		1								1			5556	1				2		5586
pron		4									1				1										366				1		373
pron_dem		1			7			1					1													767			1		778
pron_interrog						16	7		1															2			18	5			49
pron_rel													1											3				988			992
verb		18							4		20		19																5663		5724
verb_pseudo																														203	203
Total	60	5940	265	182	166	25	86	1227	1551	2	23570	252	2272	574	25	8	14	12	22	498	75	221	10	5584	367	767	19	1029	5758	203	

- The different understanding of the tag and, in some cases, its subset of tags by the annotators. For example, Table 4 shows that the two annotators disagreed concerning the tag 'noun' and the tag 'adj' in many instances. The different understanding of the tag 'adv_interrog' and the tag 'adj' has also caused a noticeable number of disagreements between the two annotators.
- Human error in the annotation process contributed to some of the errors in the annotated corpus. This was confirmed by random samples taken to be re-annotated by the same annotator.

Fig. 4. The Annotation tool.

TABLE V. THE TEN MOST FREQUENT TAGS BY THE FIRST ANNOTATOR

Tag	Frequency	%
noun	24099	47.52
verb	5714	11.27
prep	5574	10.99
adj	4632	9.13
noun_prop	2792	5.51
conj_sub	1534	3.02
conj	1148	2.26
pron_rel	992	1.96
pron_dem	778	1.53
noun_quant	575	1.13

The previous reasons were taken into consideration, and all the disagreements were highlighted, which was then given to the third annotator who went through all the disagreements and modified them based on his judgment. Finally, a final version of the corpus, which contains the agreements from the first two annotators and the agreements of the third one, was produced and used for further applications, as illustrated in the experiments section.

V. CORPUS STATISTICS

As stated, the text of the BAAC corpus was obtained from the sub-corpus Press of the BACC. The first annotator made 3150 changes to the originally tagged corpus and the second made 2959 modifications. Table 5 and Table 8 list the first ten most frequent tags for the annotators. The most frequent tag is 'noun' representing 47.52% for the first annotator and 46.48% for the second. The least used tag is 'noun_quant' being 1.13% of the tags for both annotators. A noticeable difference between the two annotators is the use of the tag 'adj' which represents 11.57% for the first annotator and occurring 1235 more times for the second annotator (9.13%).

Table 6 shows the ten most frequently used terms in the BAAC. The first and second most frequent words in the BAAC are 'في' which is a 'prep', that translates as 'in', and 'من', which is also a 'prep', that translates as 'from' representing 2.83% and 2.65% of the text respectively. The table also shows that the most commonly used bigram is 'من خلال', which translates as 'through' occurring 37 times in the corpus. Since the Press sub-corpus, which is the source of the BAAC, was gathered between 2010 and 2012 from several Arabic news websites, the most commonly used trigrams in the BAAC are 'في ميدان التحرير' which translates as 'In Tahrir Square', and 'الأعلى للقوات المسلحة' which translates as 'Higher Council of the Armed Forces', which were mentioned 12 times, and both trigrams relate to the events that happened in Egypt during the same period.

Figure 5 plots using log scales the ranked tag, bi-tag and tri-tag sequences versus their frequencies in the BAAC. There are 32 unique tags used in the annotated corpus, as mentioned earlier. The corpus also has 433 unique bi-tags where the sequence 'noun noun' dominates most of the bi-tags sequences. Finally, there are 2,113 distinct tri-tags used in the BAAC. The figure shows a Zipf's Law-like behaviour which mirrors the behaviour of a similar plot for the English language [78]. More details about the BAAC n-tag sequences are found in Table 7 and will be discussed below.

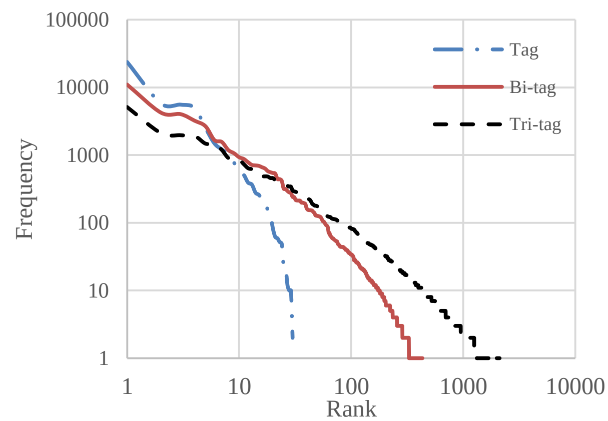


Fig. 5. Rank versus Tag, Bi-tag and Tri-tag Frequencies for the BAAC.

TABLE VI. WORD N-GRAM STATISTICS FROM THE BAAC

Rank	Word	Freq	%	Bigram	Freq	%	Trigram	Freq	%
1	في	1437	2.83	من خلال	37	0.07	في ميدان التحرير	12	0.02
2	من	1345	2.65	إلى أن	37	0.07	الأعلى للقوات المسلحة	12	0.02
3	و	735	1.45	الولايات المتحدة	34	0.07	المجلس الأعلى للقوات	11	0.02
4	أن	698	1.38	ميدان التحرير	30	0.06	القانون رقم لسنة	10	0.02
5	على	615	1.21	في مصر	28	0.05	غفر الله له	9	0.02
6	إلى	401	0.79	عدد من	28	0.05	قال أبو عبدالله	8	0.02
7	التي	352	0.69	من قبل	26	0.05	عبدالله غفر الله	8	0.02
8	عن	351	0.69	ثورة يناير	26	0.05	اللجنة الوطنية للاستقدام	8	0.02
9	أو	275	0.54	بعد أن	26	0.05	الكسب غير المشروع	8	0.02
10	لا	245	0.48	أن يكون	25	0.05	أبو عبدالله غفر	8	0.02

TABLE VII. MOST FREQUENT TAG, BI-TAG AND TRI-TAG SEQUENCES FROM THE BAAC

Rank	Tag	Freq	%	Bi-tag	Freq	%	Tri-tag	Freq	%
1	noun	23782	46.9	noun noun	11035	21.8	noun noun noun	5133	10.1
2	verb	5801	11.4	prep noun	4255	8.39	noun prep noun	2121	4.18
3	prep	5574	11	noun adj	4037	7.96	prep noun noun	1970	3.88
4	adj	4995	9.85	verb noun	3229	6.37	noun noun adj	1918	3.78
5	noun_prop	2532	4.99	noun prep	2679	5.28	noun adj noun	1482	2.92
6	conj_sub	1501	2.96	adj noun	1676	3.31	verb noun noun	1467	2.89
7	conj	1212	2.39	noun verb	1566	3.09	noun noun prep	1195	2.36
8	pron_rel	1025	2.02	verb prep	1190	2.35	noun verb noun	909	1.79
9	pron_dem	774	1.53	noun noun_prop	1066	2.1	verb prep noun	886	1.75
10	noun_quant	573	1.13	noun_prop noun_prop	932	1.84	adj noun noun	858	1.69

TABLE VIII. THE TEN MOST FREQUENT TAGS BY THE SECOND ANNOTATOR

Tag	Frequency	%
noun	23570	46.48
adj	5867	11.57
verb	5758	11.35
prep	5584	11.01
noun_prop	2272	4.48
conj_sub	1551	3.06
conj	1227	2.42
pron_rel	1029	2.03
pron_dem	767	1.51
noun_quant	574	1.13

Table 7 illustrates the ten most frequently used tag, bi-tag and tri-tag sequences in the BAAC. The tag 'noun' was utilised 23,782 times (46.9%) followed by the tag 'verb' that appeared in 11.44% of the text. The sequence of two nouns, the bi-tag 'noun noun', appeared in 11,035 occasions (21.76%), followed by the bi-tag 'prep noun' which was used 4,255 times in the BAAC. The sequence of three nouns came 5,133 times in the text, which represents 10.12% of the text, followed by the tri-tag 'noun prep noun' which came in 4.18% of the BAAC.

TABLE IX. MOST FREQUENT TAG, BI-TAG AND TRI-TAG SEQUENCES OF THE KHALEEJ SUB-CORPUS 'NEWS'

Rank	Tag	Freq	%	Bi-tag	Freq	%	Tri-tag	Freq	%
1	noun	485250	50.2	noun noun	243525	25.2	noun noun noun	122386	0.13
2	adj	120187	12.4	noun adj	91607	9.47	noun noun adj	49187	0.05
3	prep	104158	10.8	prep noun	81537	8.43	prep noun noun	43107	0.04
4	verb	91064	9.41	verb noun	52016	5.38	noun prep noun	39116	0.04
5	noun_prop	51985	5.37	noun prep	48968	5.06	noun adj noun	35544	0.04

To further analyse the n-tag results of the BAAC, Table 9 shows the tag, bi-tag and tri-tag statistics of the News sub-corpus from a different corpus, the Khaleej corpus [31], which also was tagged using Madamira tagger for comparison purposes. The sub-corpus contains 967K terms gathered from news websites. The table shows that both corpora, the News and the BAAC, share the same most frequent tag, bi-tag and tri-tag sequence, where the tag 'noun' in the sub-corpus News represents 50.2% of the text, the bi-tag 'noun noun' was used 243,525 times (25.2%) and the tri-tag 'noun noun noun' appeared in 0.13% of the text. These results confirm that the tag statistics are comparable between the different corpora.

TABLE X. TAG-BASED COMPRESSION RESULTS

Annotator	File size	Compressed size (bytes)	Compression ratio (bpc)
1	824,151	111,009	1.0776
2	819,482	110,954	1.0832
Original File	818,508	110,874	1.0837

VI. EXPERIMENTS AND FUTURE WORK

We have utilised the BAAC corpus in two applications, to evaluate the performance of the Madamira tagger, and to further investigate the tag-based text compression models as applied in by Alkhazi and Teahan [13]. Using the BAAC corpus, the Madamira tagger achieved an accuracy of 93.1%. To evaluate the effect of manual annotation on the tag-based text compression, the two versions of the BAAC were compressed using tag-based text compression models. The results of the compression were then compared to the compressed results of the original Madamira auto-tagged corpus. Table 10 illustrates the compression size (in bytes) and ratio (in bits per charactar) of all three files, and the results confirm that (1) manual annotation of the text reduces the quality of tag-based compression, as mentioned by Teahan and Alkhazi [13], [78]–[82], and (2) compressing the text using other text compression algorithms outperforms the tag-based text compression when compressing small text files, such as the BAAC corpus, as mentioned by Alkhazi and Teahan [13].

Further investigation is required to study the effect of using POS tagging systems, such as the OpenNLP project [83], trained using the BAAC on the tag-based text compression. Future work will add more annotated MSA text and will expand to cover CA text. More linguistic information, such as the structural annotation, will also be added to the BAAC to increase the possible NLP applications of the corpus.

VII. CONCLUSION

A new corpus, BAAC, was presented in this paper. It is an MSA corpus that contains 50K words manually annotated by part-of-speech tags. The annotated corpus used the same tagset utilised by the Madamira tagger and followed annotation guidelines proposed by Maamouri for annotating the POS tags. Also, a new annotation tool was developed and employed for the annotation process of BAAC which obtained a Kappa value of 0.956, and an average observed agreement of 94.25%. The BAAC was used to evaluate the Madamira tagger and to study the effect of the manual annotation on the performance of the tag-based Arabic text compression.

ACKNOWLEDGEMENT

Many thanks to our sponsor, Tabuk Public Library, for providing the facilities required by this research, and to its head, Mr Khalaf Alonazy, for his support. I would also like to show my appreciations to Ms Khawla S Alkhazi for the time and effort she spent contributing to this research.

REFERENCES

[1] A. Soudi, A. Farghaly, G. Neumann, and R. Zbib, Challenges for Arabic machine translation, vol. 9. John Benjamins Publishing, 2012.

[2] M. A. Alghamdi, I. S. Alkhazi, and W. J. Teahan, “Arabic OCR Evaluation Tool,” in Computer Science and Information Technology (CSIT), 2016 7th International Conference on, 2016, pp. 1–6.

[3] S. Green and C. Manning, “Better Arabic parsing: Baselines, evaluations, and analysis,” COLING ’10 Proc. 23rd Int. Conf. Comput. Linguist., no. August, pp. 394–402, 2010.

[4] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. S. Khorsheed, and A. Al-Rajeh, “Automatic Arabic text classification,” in Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data, 2008.

[5] P. Damien, N. Wakim, and M. Egea, “Phoneme-viseme mapping for Modern, Classical Arabic language,” in ACTEA’09. International Conference on Advances in Computational Tools for Engineering Applications, 2009., 2009, pp. 547–552.

[6] M. M. Najeeb, A. A. Abdelkader, and M. B. Al-Zghoul, “Arabic natural language processing laboratory serving Islamic sciences,” Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 3, 2014.

[7] K. C. Ryding, A reference grammar of modern standard Arabic. Cambridge university press, 2005.

[8] M. A. Alghamdi and W. J. Teahan, “A New Thinning Algorithm for Arabic Script,” Int. J. Comput. Sci. Inf. Secur., vol. 15, no. 1, p. 204, 2017.

[9] K. Dukes and N. Habash, “Morphological Annotation of Quranic Arabic.,” in LREC, 2010.

[10] M. S. Alrabiah, “Building A Distributional Semantic Model for Traditional Arabic and Investigating its Novel Applications to The Holy Quran,” Ph.D. thesis, King Saud University, 2014.

[11] K. Dukes, “Statistical parsing by machine learning from a Classical Arabic treebank,” Ph.D. thesis, University of Leeds, 2013.

[12] A. Alosaimy and E. Atwell, “Tagging Classical Arabic Text using Available Morphological Analysers and Part of Speech Taggers,” J. Lang. Technol. Comput. Linguist., 2017.

[13] I. S. Alkhazi, M. A. Alghamdi, and W. J. Teahan, “Tag based models for Arabic Text Compression,” in 2017 Intelligent Systems Conference (IntelliSys), 2017, pp. 697–705.

[14] I. S. Alkhazi and W. J. Teahan, “Classifying and Segmenting Classical and Modern Standard Arabic using Minimum Cross-Entropy,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 421–430, 2017.

[15] N. Habash, O. Rambow, and R. Roth, “MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization,” in Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt, 2009, pp. 102–109.

[16] K. Dukes and T. Buckwalter, “A dependency treebank of the Quran using traditional Arabic grammar,” in Informatics and Systems (INFOS), 2010 The 7th International Conference on, 2010, pp. 1–7.

[17] R. R. Jablonkai and N. Čebron, “Corpora as Tools for Self-Driven Learning: A Corpus-Based ESP Course,” in Student-Driven Learning Strategies for the 21st Century Classroom, IGI Global, 2017, pp. 274–298.

[18] R. Mitkov, The Oxford handbook of computational linguistics. Oxford University Press, 2005.

[19] M. Wynne, “Archiving, distribution and preservation,” Dev. Linguist. corpora A Guid. to good Pract., pp. 71–78, 2005.

[20] R. A. Abumalloh, H. M. Al-Sarhan, and W. Abu-Ulbeh, “Building Arabic Corpus Applied to Part-of-Speech Tagging,” Indian J. Sci. Technol., vol. 9, no. 46, 2016.

[21] R. Z. Xiao, “Theory-driven corpus research: using corpora to inform aspect theory,” Lüdeling, A., Kytö, M. Corpus Linguist. An Int. Handb., vol. 2, pp. 987–1008, 2008.

[22] S. Alkahtani, “Building and verifying parallel corpora between Arabic and English,” Ph.D. thesis, Bangor University, 2015.

[23] W. Zaghouani, “Critical survey of the freely available Arabic corpora,” in In LREC’14 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT), 2017, pp. 1–8.

[24] F. Al Shamsi and A. Guessoum, “A hidden Markov model-based POS tagger for Arabic,” in Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France, 2006, pp. 31–42.

[25] S. Atkins, J. Clear, and N. Ostler, “Corpus design criteria,” Lit. Linguist. Comput., vol. 7, no. 1, pp. 1–16, 1992.

[26] C. F. Meyer, English corpus linguistics: An introduction. Cambridge University Press, 2002.

[27] N. Habash and R. M. Roth, “Catib: The columbia arabic treebank,” in Proceedings of the ACL-IJCNLP 2009 conference short papers, 2009, pp. 221–224.

[28] K. M. Alhawiti, “Adaptive models of Arabic text,” Ph.D. thesis, Bangor University, 2014.

[29] “Ajdir Corpora.” [Online]. Available: http://aracorpus.e3rab.com/argistestsrv.nmsu.edu/AraCorpus/. [Accessed: 22-Sep-2018].

[30] “King Saud University Corpus of Classical Arabic (KSUCCA) | Maha Al-Rabiah – Blog.” [Online]. Available: https://mahaalrabiah.wordpress.com/2012/07/20/king-saud-university-corpus-of-classical-arabic-ksucca/. [Accessed: 22-Sep-2018].

[31] M. Abbas, K. Smaili, and D. Berkani, “Evaluation of Topic Identification Methods for Arabic Texts and their Combination by using a Corpus Extracted from the Omani Newspaper Alwatan,” Arab Gulf J. Sci. Res., vol. 29, no. 3–4, pp. 183–191, 2011.

[32] T. Zerrouki and A. Balla, “Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems,” Data Br., vol. 11, pp. 147–151, Apr. 2017.

[33] “Khaleej corpus.” [Online]. Available: https://sourceforge.net/projects/arabiccorpus/. [Accessed: 22-Sep-2018].

[34] “UN Corpus(Arabic portion).” [Online]. Available: http://www.sibawayh-nlp.com/?q=node/1158. [Accessed: 23-Sep-2018].

[35] “Islamic books | Leicester | Al Kunuz.” [Online]. Available: https://www.alkunuz.co.uk/. [Accessed: 23-Sep-2018].

[36] A. El Shamsy, “Al-Shāfiʿī’s Written Corpus: A Source-Critical Study,” J. Am. Orient. Soc., vol. 132, no. 2, pp. 199–220, 2012.

[37] “Meedan’s Open Source Arabic/English Translation Memory.” [Online]. Available: https://github.com/anastaw/Meedan-Memory. [Accessed: 23-Sep-2018].

[38] M. Altamimi, O. Alruwaili, and W. J. Teahan, “BTAC: A Twitter Corpus for Arabic Dialect Identification,” in Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018), 2018, pp. 5–10.

[39] A. O. Al-Thubaity, “A 700M+ Arabic corpus: KACST Arabic corpus design and construction,” Lang. Resour. Eval., vol. 49, no. 3, pp. 721–751, 2015.

[40] L. Al-Sulaiti and E. S. Atwell, “The design of a corpus of Contemporary Arabic,” Int. J. Corpus Linguist., vol. 11, no. 2, pp. 135–171, 2006.

[41] S. Alansary and M. Nagi, “The international corpus of Arabic: Compilation, analysis and evaluation,” in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, pp. 8–17.

[42] Linguistic Data Consortium., Buckwalter Arabic morphological analyzer : Version 1.0. Linguistic Data Consortium, 2002.

[43] “Arabic-English Learner’s Dictionary.” [Online]. Available: http://www.perseus.tufts.edu/hopper/opensource/download. [Accessed: 22-Sep-2018].

[44] M. Attia, P. Pecina, A. Toral, L. Tounsi, and J. van Genabith, “A lexical database for modern standard Arabic interoperable with a finite state morphological transducer,” in International Workshop on Systems and Frameworks for Computational Morphology, 2011, pp. 98–118.

[45] “Arabic Wordlist for Spellchecking.” [Online]. Available: https://sourceforge.net/projects/arabic-wordlist/. [Accessed: 29-Sep-2018].

[46] K. Almeman, M. Lee, and A. A. Almiman, “Multi dialect Arabic speech parallel corpora,” in Communications, Signal Processing, and their

Applications (ICCSPA), 2013 1st International Conference on, 2013, pp. 1–6.

[47] S. Al Maadeed, W. Ayouby, A. Hassaine, and J. M. Aljaam, “QUWI: an Arabic and English handwriting dataset for offline writer identification,” in Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, 2012, pp. 746–751.

[48] I. Bensalem, P. Rosso, and S. Chikhi, “A new corpus for the evaluation of arabic intrinsic plagiarism detection,” in International Conference of the Cross-Language Evaluation Forum for European Languages, 2013, pp. 53–58.

[49] Y. Benajiba, P. Rosso, and J. M. G. Soriano, “Adapting the JIRS passage retrieval system to the Arabic language,” in International Conference on Intelligent Text Processing and Computational Linguistics, 2007, pp. 530–541.

[50] D. Chiang, M. Diab, N. Habash, O. Rambow, and S. Shareef, “Parsing Arabic dialects,” in 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.

[51] R. Steinberger, B. Pouliquen, M. Kabadjov, and E. der Goot, “JRC-Names: A freely available, highly multilingual named entity resource,” CoRR, vol. abs/1309.6, 2013.

[52] Y. Benajiba, P. Rosso, and J. M. Benedíruiz, “Anersys: An arabic named entity recognition system based on maximum entropy,” in International Conference on Intelligent Text Processing and Computational Linguistics, 2007, pp. 143–153.

[53] M. I. Alkanhal, M. A. Al-Badrashiny, M. M. Alghamdi, and A. O. Al-Qabbany, “Automatic stochastic arabic spelling correction with emphasis on space insertions and deletions,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 20, no. 7, pp. 2111–2122, 2012.

[54] S. Jeblee, H. Bouamor, W. Zaghouani, and K. Oflazer, “CMUQ@QALB-2014: An SMT-based System for Automatic Arabic Error Correction,” in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, pp. 137–142.

[55] S. Pradhan and L. Ramshaw, “OntoNotes: Large Scale Multi-Layer, Multi-Lingual, Distributed Annotation,” in Handbook of Linguistic Annotation, Springer, 2017, pp. 521–554.

[56] B. Mohit, “Named Entity Recognition,” in Natural Language Processing of Semitic Languages, I. Zitouni, Ed. Springer, USA, 2014.

[57] J. Hajic, O. Smrz, P. Zemánek, J. Šnaidauf, and E. Beška, “Prague Arabic dependency treebank: Development in data and tools,” in Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools, 2004, pp. 110–117.

[58] M. Maamouri, A. Bies, T. Buckwalter, H. Jin, and W. Mekki, “Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis),” LDC2005T20, 2005. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2005T20. [Accessed: 25-Nov-2016].

[59] M. Marcus et al., “The Penn Treebank: annotating predicate argument structure,” in Proceedings of the workshop on Human Language Technology, 1994, pp. 114–119.

[60] A. Pasha et al., “MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic,” Proc. 9th Lang. Resour. Eval. Conf., pp. 1094–1101, 2014.

[61] S. Khoja, R. Garside, and G. Knowles, “A tagset for the morphosyntactic tagging of Arabic,” Proc. Corpus Linguist. Lancaster Univ., vol. 13, 2001.

[62] S. Khoja, “APT: An automatic arabic part-of-speech tagger,” Lancaster University, 2003.

[63] Y. O. M. Elhadj, A. Abdelali, R. Bouziane, and A. H. Ammar, “Revisiting Arabic Part of Speech Tagsets,” Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA, vol. 2014, pp. 793–802, 2015.

[64] J. A. Haywood, H. M. Nahmad, and G. W. Thatcher, A new Arabic grammar of the written language. Lund Humphries London, 1965.

[65] N. Schneider, B. Mohit, K. Oflazer, and N. A. Smith, “Coarse lexical semantic annotation with supersenses: an Arabic case study,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, 2012, pp. 253–258.

[66] Behrang Mohit, “AQMAR Arabic Dependency Corpus.” [Online]. Available: http://www.cs.cmu.edu/~ark/ArabicDeps/. [Accessed: 18-Sep-2018].

[67] Behrang Mohit, “brat rapid annotation tool.” [Online]. Available: http://brat.nlplab.org/. [Accessed: 16-Sep-2018].

[68] “TrEd User’s Manual.” [Online]. Available: https://ufal.mff.cuni.cz/tred/documentation/tred.html. [Accessed: 20-Sep-2018].

[69] M. Diab, K. Hacioglu, and D. Jurafsky, “Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks,” in Proceedings of HLT-NAACL 2004: Short papers, 2004, pp. 149–152.

[70] S. Green, M.-C. de Marneffe, and C. D. Manning, “Parsing Models for Identifying Multiword Expressions,” Comput. Linguist., vol. 39, no. 1, pp. 195–227, Mar. 2013.

[71] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, “Farasa: A fast and furious segmenter for arabic,” in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2016, pp. 11–16.

[72] M. Maamouri, A. Bies, and S. Kulick, “Enhancing the Arabic Treebank: a Collaborative Effort toward New Annotation Guidelines.,” in LREC, 2008, pp. 3–192.

[73] P. Pajas and J. Štěpánek, “Recent advances in a feature-rich framework for treebank annotation,” in Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 2008, pp. 673–680.

[74] D. Reidsma and J. Carletta, “Reliability measurement without limits,” Comput. Linguist., vol. 34, no. 3, pp. 319–326, 2008.

[75] J. Cohen, “A coefficient of agreement for nominal scales,” Educ. Psychol. Meas., vol. 20, no. 1, pp. 37–46, 1960.

[76] M. L. McHugh, “Interrater reliability: the kappa statistic,” Biochem. medica Biochem. medica, vol. 22, no. 3, pp. 276–282, 2012.

[77] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, pp. 159–174, 1977.

[78] W. J. Teahan and J. G. Cleary, “Tag Based Models of English Text.,” in Data Compression Conference, 1998, pp. 43–52.

[79] W. J. Teahan, “Modelling English text,” Ph.D. thesis, Waikato University, 1998.

[80] W. J. Teahan and D. J. Harper, “Using compression-based language models for text categorization,” in Language modeling for information retrieval, Springer, 2003, pp. 141–165.

[81] Z. Chang, “A PPM-based Evaluation Method for Chinese-English Parallel Corpora in Machine Translation,” no. September, pp. 1–106, 2008.

[82] W. J. Teahan, Y. Wen, R. McNab, and I. H. Witten, “A compression-based algorithm for Chinese word segmentation,” Comput. Linguist., vol. 26, no. 3, pp. 375–393, 2000.

[83] T. Morton, J. Kottmann, J. Baldridge, and G. Bierner, “Opennlp: A java-based nlp toolkit,” 2005.

Implicit Thinking Knowledge Injection Framework for Agile Requirements Engineering

Kaiss Elghariani[1], Nazri Kama[2], Nurulhuda Firdaus Mohd Azmi[3], Nur Azaliah Abu bakar[4]
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia

***Abstract*—Agile has become commonly used as a software development methodology and its success depends on face-to-face communication of software developers and the faster software product delivery. Implicit thinking knowledge has considered as a very significant for organization self-learning. The main goal of paying attention to managing the implicit thinking knowledge is to retrieve valuable information of how the software is developed. However, requirements documentation is a challenging task for Agile software engineers. The current Agile requirements documentation does not incorporate the implicit thinking knowledge with the values it intends to achieve in the software project. This research addresses this issue and introduce a framework assists to inject the implicit thinking knowledge in Agile requirements engineering. An experiment used a survey questionnaire and case study of real project implemented for the framework evaluation. The results show that the framework enables software engineers to share and document their implicit thinking knowledge during Agile requirements documentation.**

Keywords—Software development methodology; agile methodology; requirements engineering; requirements documentation; and implicit thinking documentation

I. Introduction

The implicit thinking knowledge has become very significant issue for researchers. The number of researches have been increased since that the implicit thinking knowledge considered as a primary key of success through the ability of team members to create and share their implicit thinking knowledge in software product. This consideration has attracted researchers in investigating the ways in which implicit thinking knowledge can be successfully captured, identified, categorized, shared, and documented. The topic of documenting the implicit thinking knowledge has arisen to address this need. Thus, a few knowledge frameworks designed to define practices related to knowledge. Consequently, in Agile methodology, software requirements are defined and documented in the form of user story cards. Besides, as one of Agile features that a clear discussing about user requirements is conducting by having a regular meeting [14], for example, in Scrum approach every day meeting is conducted to have to check the requirements implementation for maintaining the iterations. Traditional software development methodologies are focus more on documenting all software development phases while Agile methodology does not provide comprehensive documentation. Knowledge is easily and explicitly captured in traditional software development methodologies while the Agile methodology deals with implicit thinking knowledge. The implicit thinking knowledge is kept in the software developers' minds [17]. The most critical part of capturing the implicit thinking knowledge is to retrieve the implicit thinking knowledge from their minds, so as to incorporate the right knowledge at the right requirement when needed and to encourage innovation.

Agile methodology has few inherent practices that assist sharing experiences software developers' knowledge during developing the software, for instance, some of Agile practices helps the process to overcome the challenges of capturing implicit thinking knowledge such as pair programming, face-to-face communication and [12]. The frequent interaction among Agile software developers provides an environment that supports the implicit thinking knowledge sharing and cooperative knowledge detection. In addition, implicit thinking knowledge might be feast more efficiently by documenting the face-to-face communication than and stored in databases to be retrieved by knowledge workers [3].

The mentioned practices help to manage knowledge. Also, other approaches for capturing implicit thinking knowledge requires more consideration. This paper proposes a framework injects the implicit thinking knowledge of software engineers in Agile requirements documentation (IITKARD).

A quantitative method such as survey questionnaire has been used in a case study of a real software project for a purpose of the framework evaluation. During the experiment, software requirements used in a real software project converted into user Agile user story cards as a form of Agile requirements engineering. In order to improve the efficiency and usability of findings, a survey questionnaire has been conducted to collect data from focus group of Agile software engineering experts. Therefore the results of the survey have to evaluate the efficiency and the usability of the proposed framework and its usefulness in Agile software methodology.

The proposed implicit thinking knowledge injection provides a systematic process for Agile software engineering during the requirements engineering phase to inject their implicit thinking knowledge in Agile requirements documentation. The framework will provide a systematic process to successfully address the issue of neglecting of implicit thinking knowledge in Agile requirements engineering. This paper divided into five sections, firstly section 2, which discusses the implicit thinking knowledge and Agile requirements engineering, followed by section 3, which describes the proposed IITKARD framework. Then section 4 discusses the results of the evaluation experiment and finally the conclusion.

II. Implicit Thinking Knowledge and Agile Requirements Engineering

The implicit thinking knowledge management includes a set of processes that control the creation, interaction, and sharing of knowledge [14]. It is a few steps used during developing the software to extract and share developers' knowledge to assist them understand how the software is built. The retrieved knowledge needs to be organized so that they are clear and searchable [5]. Moreover, managing the implicit thinking knowledge also includes strategies to create an environment of sharing knowledge and supportive tools that assist the process of injection of software developers' thinking knowledge, in turn, to learn from each other.

Implicit thinking knowledge management is not only a technical issue [1]. Since it relies on people expertise, the success of managing the implicit thinking knowledge is highly influenced by human. However, these factors named a non-technical factors, which are related to the characterization of human implicit thinking knowledge, and the promoting the knowledge workers to contribute to share their implicit thinking knowledge in order to build shared knowledge repository [3].

Unfortunately, knowledge workers are unenthusiastic to share their own implicit thinking knowledge with other people [5]. Some workers consider their knowledge as a private professional knowledge and they are not often intending to share it [15]. The workers reluctance of sharing knowledge might also be contributed by other organizational practices. The staff, who are rewarded due to their expertise might lead to that knowledge might be kept in their minds [5]. Moreover, psychological issues might also effect the participation of the knowledge workers [16]. For example, in the organizations meetings some workers try to avoid to share their knowledge and this would adversely affect the meetings process as a result [19]. According to Hissen [1] some members are hardly to speak, due to their low status, shyness or controversial ideas. Therefore, the externalization of implicit thinking knowledge should be part of a defined knowledge and it should not be left an optional. For example, in software engineering field, software engineers have deliberations includes different views of software projects lifecycle reviews should be captured. Thus, more attention need to paid regarding to the importance of the integration between software development methodologies and knowledge management.

Obviously, that implicit and explicit are different in terms due to implicit thinking is hard to document. Sandra et al. [20] highlighted that implicit thinking knowledge documentation is not something for discussion sense [21]. Though, the significance of capturing implicit thinking knowledge during requirements engineering led the researchers to pay more attention on topics related implicit thinking knowledge documentation [18]. However, tacit and implicit knowledge are slightly different [22]. Researches stated that implicit thinking knowledge is not organized knowledge but it can be structured and documented. On the other hand, organizing the implicit thinking knowledge is not easy. Moreover, implicit thinking knowledge categorized as an expression, assumptions, developers thought that, might be translated to principle. According to Correia and Aguiar [8] there is insufficient information of implicit knowledge, while [9] argues that implicit thinking knowledge is "how" to do things, but some issues to be explicitly described.

Additionally, the technical issues contribute in the difficulties of capturing the implicit thinking knowledge, the implicit thinking knowledge is not well structures and there is not any standard format to be followed. Therefore, using traditional data models could hardly capture he implicit thinking knowledge. Few models in the software engineering aimed to codify the implicit thinking knowledge such as DRL [4], QOC [11] and IBIS [13]. Even though, these models support the implicit thinking knowledge capturing, but it is hardly to cover all implicit thinking knowledge. Sometimes the tacit knowledge is differently expressed; it might be a body language expression. Thus, these all variations of implicit thinking knowledge expression is hardly to be accommodated.

Elghariani and Kama [7] stated that Agile requirements engineering practices assist to resolve the traditional requirements engineering challenges and highlighting appropriate capability of teams' cross-functional development [2]. In addition, in software industry, some challenges have been faced during practicing Agile requirements engineering such as less documentation and ignoring non-functional requirements documentation e.g. usability and security [7]. Neglecting of implicit thinking knowledge documentation is considered as a major issue for both Agile and traditional software methodologies [10]. One of Agile requirements engineering practices is creating user story cards, which includes few attributes related to the user software requirement, for example, story card number, story date, story priority, story description, etc. [6]. Therefore, Shim and Lee [25] addressed that ignoring implicit thinking knowledge documentation in Agile methodology cause some major issues as following:

- Time waste of asking the same questions by software developers.
- Software issues might be frequently faced by developers but they forget how were solved.
- Losing information once particular developer left the project.
- Lack of recording developer's communication.
- Software usability issue
- Neglecting of unstructured knowledge contribution.

The major challenges of documenting the implicit thinking knowledge in Agile requirements engineering is how implicit information software engineers can be converted into explicit information, as well as how to convert explicit information from individual software engineer to groups in the organization [21]. The retrieved knowledge needs to be recorder and stored in a repository. Additionally, the process of sharing software engineer's implicit knowledge needs to be followed by all software developers [15]. Also, this includes understanding the purpose of storing and documenting the implicit thinking knowledge [12].

Chau and Maurer [23] stated that there are serious issues to capture team members' implicit thinking. Most of the members are not able to exactly define their thinking. So Ahmed [24] suggested that software developers need to have more face-to-face discussion on software requirements. Another suggestion is using electronic databases will solve the issue, if the face-to face communication is not possible then some alternatives can be used such as e-mail, social media, audio and video conferencing [18].

III. Related Work

Based on the previous studies, few frameworks were proposed to manage the knowledge using traditional software development methodologies. A framework for supporting knowledge management (KMS) recommends two significant considerations, which are, the existence and structure of the knowledge [12].

The proposed framework KMS layered into seven layers known as interface, access, collaborative, application, transport, integration and repositories [22]. In addition, Hahn and Subramani [12] proposed an approach for knowledge management system known as Soft-System Knowledge Management. This approach was proposed to confirm the suitability between the organizational requirements on new product development (NPD) and knowledge management creativities. There are three main components are included in this framework, which are Knowledge sharing methods, Organizational level and Key Enablers.

A Knowledge Management System in Open Source Software Development Environment (KMSOS2oD) framework is particularly designed to support Open Source Systems development life cycle and it includes five core components, which are layers, components, process, knowledge, and Communities of Practice (CoP) [15].

Several tools have been proposed by researchers to support distributed Agile software development, and those tools are especially developed to support sharing applications. The proposed tools adopted the concept of collaboration and communication among team members. Therefore, the tools used are messengers, e-mails and newsgroups. Tools like videos/audios conferencing are also used as a real-time collaboration tools [8]. Moreover, tools providing a Plug-in like JSPWiki and MASE are used to support environment tailored for Agile development teams [17].

Kavitha and Irfan [17] have proposed a framework to capture knowledge retrieved from the Agile team members, which may either be distributed or collocated. The framework helps to increase the organizational learning process to capture and maintain the knowledge on the required information. The framework included a set of integrated tools that support capturing and storing knowledge. The tacit knowledge sharing was through a real-time collaboration tools such as news groups, NetMeeting and e-mails.

Shim and Lee [25] suggested an approach to capture and manage knowledge. This approach was built by using the main characteristics of the social software and expands them to merge the knowledge and its structure then manage them probably by using online tools. The authors used web technologies to extract the implicit thinking knowledge and injected to the knowledge base; the injecting was used with the platform as provider to enable the users to retrieve a documented knowledge related to the certain software practice, which also affect the user interaction.

Based on literature related to knowledge sharing within software development, there are many approaches aimed at this objective. Most of the proposed frameworks aimed for managing software knowledge designed for traditional methodologies [17]. Some approaches are realized practically, while others are simply theoretical such as Rayan and O'Connor [22] that propose a theoretical model for defining, acquiring and sharing tacit knowledge surfaced through social interaction. KMSOS framework is another approach, which is introduced by Lakulu [15], but it is mainly targeting capturing software knowledge within the Open source software process. However, it does not cater for documenting any sort of implicit thinking knowledge. In terms of approaches targeting the Agile software methodology, though [17] and [25] are most related to our work as they both targeting software knowledge management for Agile software methodology. Nonetheless, both are aimed to tackle the knowledge sharing in all phases of the Agile lifecycle. We on the hand, only focusing on the requirement elicitation phase. We believe that the requirements for capturing implicit thinking knowledge vary in different Agile phases. Hence, a generic knowledge-capturing model would not cater for the capturing richness of the requirements engineering phase. Because it is well known that the requirements engineering phase accommodates most of the deliberations made by the Agile team. However, though both [17] and [25], are targeting the capturing of implicit thinking knowledge within the Agile software process, but they ignored the importance of the User Story Cards, which as at the core of implicit thinking knowledge generation. The same approaches ([17], [25]), also lacks the formal representation of the implicit knowledge, they instead adopting a social media tools as repository to store the implicit thinking knowledge.

IV. The Proposed Knowledge Framework

The framework is proposed to be part of reality [1]. It is meant to summarize the conceptual framework of the reality being demonstrated. Essentially, our proposed framework is built to inject the implicit thinking knowledge created while practicing Agile software requirements documentation. The proposed framework facilitates sharing implicit knowledge [5].

The Internet or local network usually connects Agile software team members. The framework supported by a tool that assists in collecting the knowledge. Experiences generated in the form of arguments categorized as issue, assumption, suggestion, question and opinion are entered by the software engineer and stored in the knowledge repository.

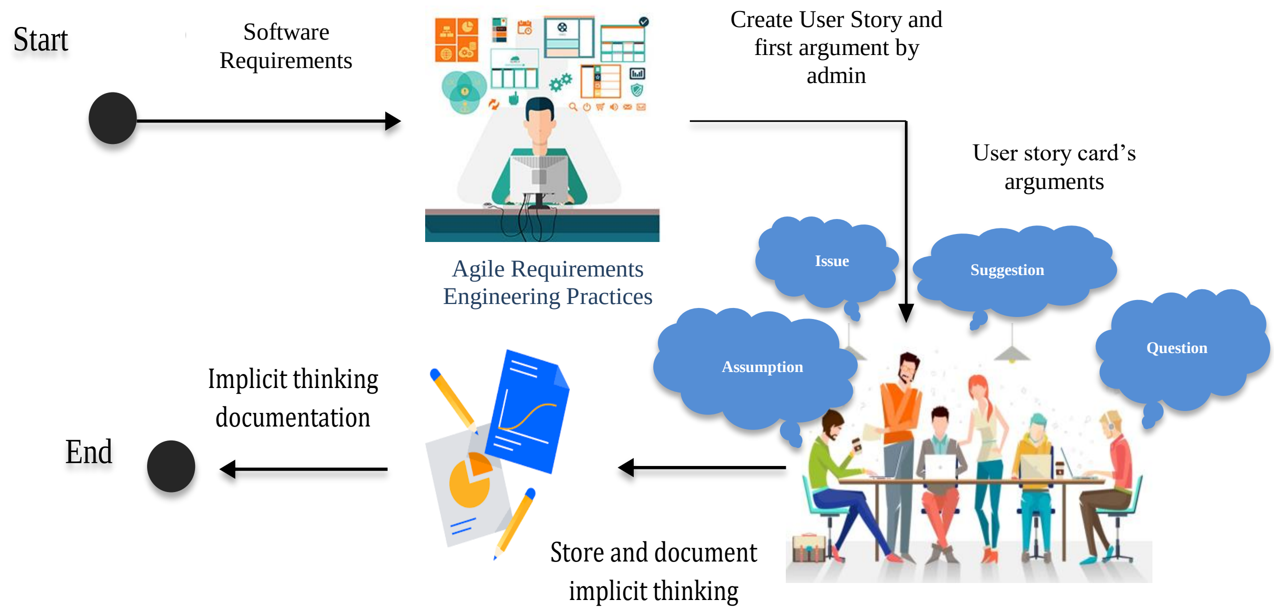


Fig. 1. Conceptual View of IITKARD Framework.

Fig 1 shows a conceptual view of the proposed framework. It shows the stages of the processes of sharing software engineering implicit thinking knowledge in Agile requirements engineering. Based on the proposed IITKAD framework, the knowledge generation loop starts from creating the user story card, which will be the root of the experience knowledge.

The IITKARD framework helps to share informal knowledge. The implicit thinking knowledge generated through the IITKARD framework helps to maintain the up-datedness of the captured knowledge. Moreover, this framework provides Agile software engineers with a supportive tool that helps in injecting the implicit thinking knowledge. The user story card created by the admin (team leader) and adding the first argument of the retrieved knowledge related to requirements engineering practices are entered via an interface by the admin (team leader) and stored in the knowledge database. The implicit knowledge extracted during developing the software requirements is accessible to all team members and easy to find. Agile team members will be able to retrieve the information related to certain user story card from database as structured information. Efficient query mechanisms also provided, which can be helpful to retrieve the required data.

The proposed IIKARD framework, which has four main activities, which are: (1) cerate the user stories; (2) set team leader first arguments; (3) inject the implicit thinking knowledge of software engineers; (4) document the implicit thinking knowledge. Therefore, to evaluate the framework efficiency, a tool was developed to systematize the processes of developing the framework as shown in Fig 2. The tool of the proposed framework has four main steps, which are:

Step 1 Create User Story: The first practice of Agile requirements engineering is to create the user story of each requirement and the proposed framework tool requires the team leader to key in the user story information for example: Story Title, Task Engineer and describe the story (requirement).

Step 2 Set Team Leader first Argument of the user story: The second step of the proposed framework tool is setting the first argument by the team leader (Admin).

Step 3 Inject the implicit thinking knowledge of software engineers: The Third step of the proposed framework tool is to let the team members to key in arguments such as Issues, Assumptions, Suggestions, Questions and Opinions.

Step 4: Document the implicit thinking knowledge: The final step of the proposed framework tool is to document the arguments of each user story by displaying the arguments in sort of storyline. It contains the title and its arguments member's photo, name, text and icon of the argument type whether it is an issue, assumption or suggestion. Hence, this, IITKARD framework was successfully designed and developed to achieve this research objective in order to provide a framework to inject the implicit thinking of software engineers in Agile RE.

Therefore, IITKARD framework controls the deliberation among team members and supports their arguments by detecting the type of the arguments and simplifies the communication about certain issue. The tool which implements IITKARD framework steps shall be applied and helpful for areal data from a real software project adopting Agile methodology as a software development methodology.

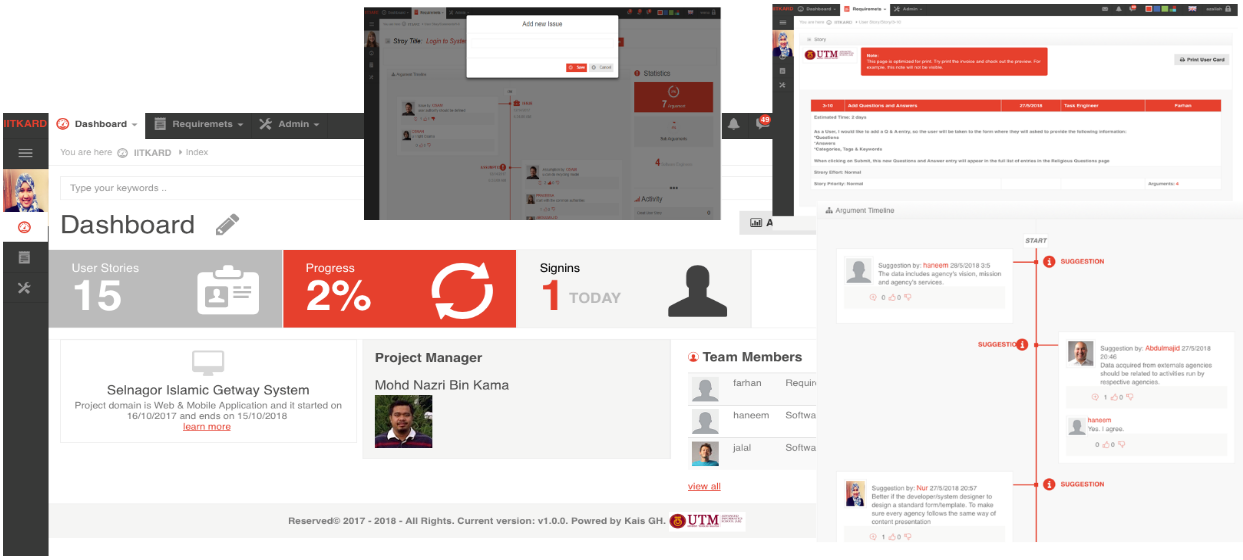


Fig. 2. An Implementation of the Proposed Framework.

V. Results

After performing, the first experiment, evaluation survey questionnaire has been distributed to focus group of experts. The 13 completed questionnaires have been received from 10 Agile software engineering experts. The experts came from different organizations with the experience of Agile software methodology and requirements engineering. The participants started performing the experiment by using traditional Agile requirements documentation. However, the experiment shows that the documentation does not include any implicit thinking knowledge in the user story cards. The experiment shows that the overall results obtained a high level of experts' satisfaction in terms of IITKARD efficiency evaluation. In general, the overall results shows that the participants performed as an experts' in Agile requirements engineering and improves that IITKARD is more efficient in Agile methodology since Agile lean on a strong communication among the team members.

The traditional requirements documentation used for the experiment does not include any implicit thinking knowledge shared by software engineers during developing software system which been adopted as a real project data for this experiment. The experts' participants in the experiment shows their satisfaction on the efficiency of IITKARD due to the improvement of ITKARD framework that can assist software engineers who are newly involved in the software project to understand how the requirement was developed. Also, the experiment proved that the IITKARD improves the usability of injecting the implicit thinking knowledge in Agile requirements documentation. It is noted that the improvement rate of the IITKARD usability is high in Agile methodology.

The proposed models of knowledge's management in both traditional and Agile methodology [8][17][18][19][25], use a social media tools to support sharing and recording the knowledge, but it is well known that social media knowledge is not structured enough, and any searching through this knowledge would not be efficient in terms of retrievability. IITKARD on the other hand, proposed a custom knowledge model, which strikes a balance between ease of use and retrievability, as shown through the tool developed to demonstrate the process of injecting the implicit thinking knowledge among the team members. This is achieved by managing the requirements (User Story Cards) and the implicit knowledge. Therefore, the experts' perceptions of the outcome show that the overall results of the experiment are at satisfactory level in terms of the usability of the framework.

VI. Conclusion

Software engineering is a domain consists a very knowledge thorough and realized in implicit form in many software companies. Knowledge management systems were designed more for structured knowledge while implicit thinking knowledge is ignored. The implicit thinking knowledge is usually held in professionals' minds that might be lost at any time. Therefore, there is significant need to exploit implicit thinking knowledge as well. This paper proposes a framework that injects implicit thinking knowledge in Agile requirements engineering. The framework integrates both explicit knowledge in the form of user story cards and implicit thinking knowledge in the form of arguments that compose the context of the creation criteria of the knowledge captured. In order to inject the implicit thinking knowledge in Agile requirements engineering the captured knowledge documented in the form of user story arguments.

VII. Limitation and Future Work

Even though the research contributions, a few limitations in this research are considered. The designed IITKARD framework has suggested that the shared implicit thinking knowledge should be modified and deleted. The injecting process suggestion was made to the software requirements engineering practices. There are two elements exist in the software requirements engineering process that are dependent to one and another. The requirements engineering elements are Requirements analysis, Requirements Documentation. This research study applied only for requirements engineering as the input and output of IITKARD framework and this is one of the limitations of this study. Therefore, it is important to apply

IITKARD framework to all Agile methodology phases since all the shared knowledge will cover all the phases, which could affect the overall software development process.

Based on the limitations described above, the following areas may constitute possible future work:

- Extending the research study by covering a bigger and more complex Agile software projects as samples of the case selections. Big and complex Agile software projects may involve a longer software requirements implementation and more team members who can interact with proposed IITKARD framework.

- The software requirements engineering elements consist of analysis, documentation in each of the software requirement. These two main Agile requirements engineering elements are correspond to one and another. For example, if one requirement is deleted or changed in the software engineer should consider the implicit thinking knowledge related to that particular requirement to be modified or deleted as well. In this study, the proposed IITKARD framework concentrated only on the Agile requirements documentation. The suggested future work could extend this study by implementing the IITKARD framework all Agile methodology phases.

- Since this study has used academic context to run the experimentation, a future work can replicate the experiment design in the industry context to examine the practitioners' perceptions in using the IITKARD framework. In addition, it can be extended whereby the evaluation does not only measure the practitioners' perceptions but it can also measure the software engineering experts satisfaction of the end product that is developed using the framework of injecting the implicit thinking knowledge in Agile requirements engineering.

ACKNOWLEDGMENT

The author would like to express the acknowledgment to Razak Faculty of Technology & Informatics for the support given in conducting this research study.

REFERENCES

[1] Abdulmajid Hissen, "Capturing Software-Engineering Tacit Knowledge", Departement of Computer Science, Sebha University, 2008 2nd Eroupean Computing Conference.

[2] Aneesa Rida Asghar, Shahid Nazir Bhatti, Atika Tabassum and S Asim Ali Shah, "The Impact of Analytical Assessment of Requirements Prioritization Models: An Empirical Study". International Journal of Advanced Computer Science and Applications, Vol. 8, No. 2, 2017

[3] Bjarnason, E., Wnuk, K., & Regnell, B. (2011b). Requirements are slipping through the gaps—A case study on causes & effects of communication gaps in large-scale software development. In 2011 IEEE 19th international requirements engineering conference (pp. 37–46).

[4] Claudia Müller, Julian Bahrs, Norbert Gronau, "Considering the Knowledge Factor in Agile Software Development", Journal of Universal Knowledge Management, vol. 0, no. 2 (2005), pp.128-147

[5] David Walsh Palmieri, "Knowledge Management through Pair Programming", Thesis

[6] Deki Satria, Dana Indra Sensuse and Handrie Noprisson, 2017 "A systematic literature review of the improved Agile software development", in 2017 International Conference on Information Technology Systems and Innovation (ICITSI). Bandung, Indonesia, PP 94 – 99

[7] Elghariani, K., and Kama, 2016. Review on Agile requirements engineering challenges,in 3rd International Conference on Computer and Information Siences, Kuala Lumpur, Malaysia, pp. 507-512.

[8] Filipe Figueiredo Correia, Ademar Aguiar, ". Software Knowledge Capture and Acquisition: Tool Support for Agile Settings", IEEE Computer, 2009

[9] Meira Levy, Orit Hazzan, "Knowledge Management in Practice:The Case of Agile Software Development", IEEE software, 2009, pp. 60-65

[10] Ghani I., Azham Z., and Jeong SR. "Integrating Software Security into Agile-Scrum Method". KSII Transactions on Internet an Information Systems (TIIS) 8 (2), 646-663. 2014.

[11] Gerardo Canfora, Aniello Cimitile, Corrado Aaron Visaggio, "Lessons learned about distributed pair programming: what are the knowledge needs to address?" IEEE Software, 2003

[12] Hahn, J., & Subramani, M. ''A Framework of Knowledge Management System: Issues and Challenges For Theory and Practice'', Proceedings of the twenty first international conference on Information systems, ACM, Brisbane, 2000

[13] Harald Holz and Frank Maurer, "Knowledge Management Support for Distributed Agile Software Processes", 2003

[14] Inayat et al., (2014). A systematic literature review on Agile requirements engineering practices and challenges. Computers in Human Behavior. (2014).

[15] Modi Lakulu, "A Framework of Collaborative Knowledge Management System in Open Source Software Development Environment" Journal of computer and Information Science, Vol.3, No.1, Feb 2010, pp.81-90

[16] Mustafa Ally, Fiona Darroch, and Mark Toleman, "A Framework for Understanding the Factors Influencing Pair Programming Success", XP 2005, Sheffield, UK, pp.18–23

[17] R. K. Kavitha and M. S. Irfan Ahmed, 2011. "A Knowledge Management Framework for Agile Software Developme nt Teams", in 2011 International Conference on Process Automation, Control and Computing. Coimbatore, India, PP 1-5

[18] Saeed S. and Alsamdi I., 2013. "A software development process for open source and open competition projects", International Journal of Business Information Systems, 12(1), 110-122

[19] Sandra L., Buitrun Francisco J., Pino, Brenda L., Flores-Rios, Jorge E., Ibarra-Esquer, Marua Angulica Astorga-Vargas, 2017. "A Model for Enhancing Tacit Knowledge Flow in Non-functional Requirements Elicitatio", in 2017 5th International Conference in Software Engineering Research and Innovation (CONISOFT). Mérida, Mexico

[20] Shankar, R., Acharia, S., & Baveja, A. ''Soft-system Knowledge Management Framework for New Product Development'', Journal of Knowledge Management, 2009, 13(1) , pp.135-153.

[21] Sufian Fannoun ; John Kerins, 2017. "Evaluating current practice and proposing a system to enhance knowledge assets within a small software development unit", 2018 4th International Conference on Information

[22] Ryan, S. and O'Connor, R., Acquiring and Sharing Tacit Knowledge in Software Development Teams: An Empirical Study, Information and Software Technology, Vol. 55, No. 9, pp. 1614 - 1624, 2013.

[23] Thomas Chau, Frank Maurer, "Integrated Process Support and Lightweight Knowledge Sharing for Agile Software Organizations", 2005

[24] Usman Ahmed, 2018. "A review on knowledge management in requirements engineering", in 2018 International Conference on Engineering and Emerging Technologies (ICEET), Lahore, Pakistan, PP 1-5.

[25] Woogon Shim and Seok-Won Lee, 2017. "An Agile Approach for Managing Requirements to Improve Learning and Adaptability", in 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW). Lisbon, Portuga, PP 435-438

Usability Testing for Crop and Farmer Activity Information System

Halim Budi Santoso[1], Argo Wibowo[4]
Department of Information System
Duta Wacana Christian University
Yogyakarta, Indonesia

Rosa Delima[2], Emylia Intan Listyaningsih[3]
Department of Informatics
Duta Wacana Christian University
Yogyakarta, Indonesia

***Abstract*—Information System usability level depends on acceptance and system convenience to be run by users. One of the methods to measure usability level is by conducting usability testing. This article elaborates usability testing for Crop and Farmer Activity Information System. This system is one of the agriculture information systems that is developed to record system activities for each farm field. This system is considered as one of the important role of Information and Communication Technology (ICT) for agriculture. This system has been developing since 2017 and needs to be assessed and tested. To assess the system, usability testing was conducted by taking sampling from two regions in Central Java, Temanggung and Gombong. The respondents are system administrators, farmers, and general users with each of respondents has different criteria. There are 58 respondents participated in this research: 49 farmers, 3 system administrators, and 6 general users. Usability testing was carried out by giving respondents several test tasks based on the system. Each respondent had different kind of test task in accordance with the system functionality for each user. The result of the test found that system administrator user interface assessment value gained average percentage of 69%, while the farmers gained 76% and general users gained 79%. From the test, it also bring some recommendations for system refinement. Those recommendations were taken from user inputs and user test results. The recommendations have been made to bring better system environment.**

Keywords—Usability testing; crop and activity information system; improvement recommendation; precision farming; information technology for agriculture

I. Introduction

Agriculture is one of the biggest sectors in Indonesia. Indonesia is popular with diversity of crop: rice, corn, sugar, and soybeans [1]. Historically, agriculture contributed to Indonesia's economic growth and decrease unemployment. It also helped government to decrease poverty rate [1]. Agriculture sector has low productivity while the number of people engaged in this sector is very high [2]. This problem becomes one of the challenges for the government. To increase the farmer's productivity, there are some important things should be handled [1]: 1) Government should focus on farmer incomes; 2) Government should increase the productivity through research and extension system; 3) Government is able to provide funding by giving farmer credit from state budget; 4) Government is able to help farmer by giving access to irrigation and building infrastructure for irrigation; 5) Government should protect agriculture sector and farmer from imported diseases and increase agriculture product standards.

Information and Communication Technology (ICT) is one of tools that can help government to increase farmer's productivity. Initiatives to develop rural information and communication technology (ICT) bring an opportunity to disseminate information to rural communities. Evolution of smartphone user also helps ICT penetration to rural area. Implementation of ICT for agriculture encourages an innovation in ICT for agriculture. Implementation of ICT in agriculture is able to help in three main processes [3]: (1) Land selection and calendar definition by giving information systems including Decision Support System, Management Information System, or Geographic Information System; (2) ICT enabled learning and knowledge exchange by helping in calendar definition, land preparation, access in credit for farmer, water management, and input management; (3) Networking and e-Commerce to help agriculture product marketing.

Comment [PP1]: paper

One of the systems that is enabled ICT role for agriculture is by providing information system for crop and farmer activity. Researchers has been developing crop and farmer activity in order to collect and issue crop data collection, planting calendar definition, farmer activity, and agriculture product data collection [4]. Crop and Farmer activity Information System is also able to provide report graphically and summarize data. The system is also able to predict harvest time for farmers in some specific calendar [4].

Crop and Farmer Activity Information System has been developing since 2017. This system needs to be evaluated and tested to measure system satisfaction level and give improvement recommendation. Therefore, usability testing and evaluation should be conducted to achieve system efficiency, effectiveness, and satisfaction. The objective of this research is to evaluate system interface using usability testing. From the evaluation, some improvement recommendation is conducted to bring better user interface and the system is able to be used easily.

This paper is discussed the study background and what has been done in the previous research. Following the first part, researchers discuss about usability testing and some researches were conducted related with usability testing. Research methodology is the next part and followed by discussion and analysis. The last part is the conclusion and recommendation for future works.

II. Literature Review

A. Crop and Farmer Activity Information System

Crop and Farmer Activity Information System is a system that gives information to the user related to crop specification, planting calendar, and prediction of harvest time [4]. This system is developed in Indonesia Language since the system is targeted to users (farmer, farmer group representative, and academician). Crop data collection provides detail information about crop and specific characteristics. Figure 1 shows crop information page.

Figure 1 shows Crop Information Page which provides detail information about the crop. Some detail information like root characteristic, trunk characteristic, leave characteristic, fruit and seed characteristic. This information is available for farmer, farmer group representative, and academician [4]

Fig. 1. Crop Information Page [4].

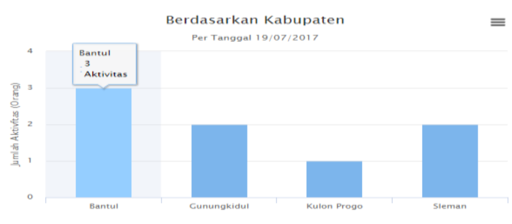


Fig. 2. Number of Land Preparation Activity in Yogyakarta Special Region [4].

Fig. 3. Farmer Planting Activity Based on Crop and Region [4].

Another function available in this system is farmer activity. This feature provides a system for farmer to entry specified information related with what activity they do. This data bring information about the number of farmer activities (figure 2, figure 3) and yields (figure 4, figure 5) based on district and crop.

Figure 2 and 3 shows the farmer activity reports and summary. This report provides information about the number of farmers or users which conducting specific activity (figure 2). On the other hand, figure 3 shows information about the number of crop planted by farmers. This two reports can be drilled down based on region or province.

Fig 4 and 5 brings information of the amount of yield for specified date. This information are generated from the data entered by farmers. These four graphics are able to cascade down to get more detail information.

No	Nama Tanaman	Jumlah Panen (Kuintal)
1	Padi	140
2	Bawang Merah	15
3	Bawang Putih	
4	anggrek	
5	Jagung	30
6	Gandum	
7	Bayam	
8	Kentang	
9	Kedelai	
10	Ketela Pohon	20
11	Salak	50
12	Cabai Rawit	10
	Total	265

Fig. 4. Information related the Number of Yields (in quintal) [4].

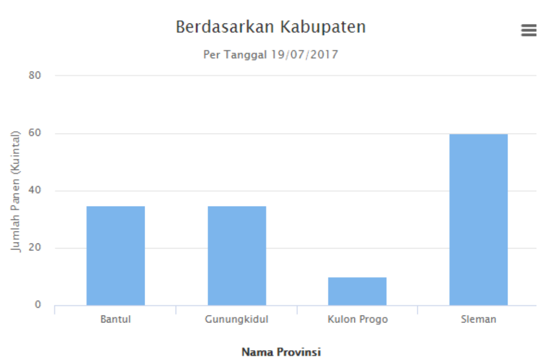


Fig. 5. Yield Information based on Region [4].

B. Usability Evaluation and Testing

Usability has been defined as a measurement of system effectiveness, efficiency, and satisfaction [5]. Usability is also an assessment to measure quality level and human point of view about the systems [6] [7] . Mack and Nielsen [8] categorized the evaluation method into four classification: (1) Usability evaluation through evaluation software; (2) Usability testing is assessed by evaluating the user interface with real user; (3) Usability measures by using models and formulas; (4) Usability is measured based on rules of thumb and the general skill, knowledge, and experience of the evaluators. The usability testing is an assessment method to measure the degree of interactive system is easy and pleasant to use with a view of identifying usability problems and / or a collection of usability measures / metrics [9]. Thus, usability is a media to gain good quality of web, by measuring reliability, functionality, usability, efficiency, maintainability, and portability [7].

Usability has 5 quality components [10]: (1) Learnability: How easy the users to finish task given and how fast user understand functionalities in the system; (2) Efficiency: How fast the user can perform the tasks given after learn the system; (3) Memorability: How easily user can reestablish proficiency after a period of time do not use it; (4) Errors: How many times users do error and how easily the users can recover from the errors? (5) Satisfaction: How positive the users against the system user interface. Usability is necessary for system survival. People will choose the system that is easily to be used. People will not use the system that difficult to be used [10]. Thus, usability testing is done to test those 5 quality components.

Usability is concerned with quantitative and qualitative analysis, which is measured timed-based and traffic-based data [9]. Usability level is gained by conducting testing for respondents by distributing the questionnaire. Each question will be tested validity and readability that the data obtained from the questionnaire can be used as a solid basis for research data [11].

Usability testing has been conducting for some research topics. Lestari [12] conducted usability testing to test web application for small and medium enterprise in Bandung, West Java, Indonesia. Usability is measured using Post-study System Usability Questionnaire (PSSUQ). There are four reviews: overall satisfaction score (OVERALL), system usability (SYSUSE), information quality (INFOQUAL), and user interface quality (INTERQUAL). From the study, it is found that 56% of the respondents are satisfied with the system (4.22 for OVERALL points). 64% of respondents agree that web application helps them to promote their products (4.43 for SYSUSE points). Small and medium enterprise site does not have complete information for product catalogue. It makes information quality score only 3.71 with 36% respondents agree. On the other hand, 60% respondents agree that the system has good quality user interface.

Usability testing was also done to evaluate government websites [13]. This paper is to study whether the government websites fulfills the Nielsen usability components. To find the result, 30 users were given some tasks and post-test questionnaires. Those tasks are related with finding some information in the e-Government sites and post-test questionnaires included Nielsen's usability components. From the tasks and post-test questionnaires, it is found that most of users (27 respondents) agreed that the e-Government websites is quick to perform. 28 of 30 respondents also agree that the e-Government sites are easily to be remembered and they are satisfied with the websites. With the new user interface, e-Government sites quicker to perform, easier to understand, and more satisfied. The researchers already conducted the third test and it is found that the result from the third test increased comparing with the other two tests.

Usability evaluation was also done to test the e-Learning system in one of the public universities in Kenya [14]. This university has implemented Moodle e-Learning system. This research is to find what factors that are affected the usability of e-Learning system. The components that were tested in this evaluation: learnability, user-friendliness, technological infrastructure, usability policy, culture, and gender. From the study, it is found that the learnability brings significant affect to usability of e-Learning system. It is suggested that to enhance e-Learning system in the university, the system should be easily to be learnt. User friendly factor also affects the user ability. E-Learning should be user friendly to be usable in the university [14].

Usability testing is also to test one of the biggest community site in Indonesia, kaskus.co.id. There are some problems found in this research, such as the process of the posting, the advertisement on the page which makes user inconvenience, and difficulties to organize picture in this community [15]. Some influential factors of website, such as simplicity, user-friendliness, comfort, navigability, link visibility, high and readable color contrast, and right to the point information got medium point, which range from 0.2 to 0.6. The finding is also got moderate level of usability for kaskus, the community site in Indonesia [15]

III. Research Methodology

Starting this research, researchers did literature study and exploring crop and farming activity system. This initial study was conducted to observe research methodology and explore system functionalities. After finishing initial study, researchers define the respondents. There are 58 respondents from two regions in central java, Gombong and Temanggung, involved in this research. Those respondents are categorized into 3 kinds of users with each user has specific ability and requirements: (1) Farmer: Farmer respondents should be able to operate computer and / or smartphone, minimum age is 20, and graduate minimum from junior high schools; (2) System Administrator. System administrator should be able to operate computer and understand the basic computer operation, and graduated from bachelor degree; (3) General User. General user graduated at least from junior high school, minimum age is 20, has the ability to operate computer and / or smartphone. General user can be government as a representative from department of agriculture, academician, or civil society.

After defining the respondents, the next step in this research is defining tasks with these following criteria: (1) Task description should be available in each task; (2) The tested page should appear when user do the task on the tested page; (3) There should be explanation about the task step in each task given; (4) Success task is given when respondent is able to finish the task; (5) Maximum time (second) is time limitation for respondents to finish each task. Table 1 below describes task for farmers:

Table I, II, and III shows the detail tasks for specific users. To test this website usability, there are some indicators: (1) Task Success. This indicator measures user effectiveness in order to finish the task; (2) Time on task. This indicator measures how much time respondent needs to finish a task; (3) Error. This indicator evaluates respondent performance to use the system; (4) Efficiency. This indicator measures the system efficiency to finish the tasks by count the number of click.

Researchers also prepared test scenario for respondents, as follows: (1) Usability testing is carried out one by one respondent of each required criteria; (2) Respondent will be provided with a laptop with the system displayed on the laptop; (3) Usability testing is conducted at the flexible places; (4) Researchers explain each step of testing step; (5) Respondent will be given a document about system description and task; (6) Researchers record processing time for a respondent to finish a task, count errors that happen during doing the test; count the number of click to finish a task.

Researchers will analyze the data with metric usability – task success, metric usability – time on task, error data analysis, and efficiency data analysis. The next step after analyzing the data is making some improvement recommendation. Improvement recommendation is given based on the usability test result.

TABLE I. TESTED TASKS FOR FARMER RESPONDENT

No	Description	Things to Do	Success Criteria	Maximum Time (in second)	Minimum Clicks
1	Login to Crop and Farmer Activity Information System Username: us2 Password : 123	a. Search and find link to login to the system b. Entry some textboxes to login page	Respondent is able to login to the system. The system will give output WELCOME ….. on the main menu	31.32	4
2	Sort farmer activity based on activity date descending	Search and choose textbox to do data sorting	System will sort the data and show the sorted data	97.12	5
3	Change the description of a crop. Use garlic for an example	Search EDIT button and update data	After success to change the description, system will show pop up message “Data is successful to be updated”.	91.24	4
4	Add a data in menu crop. Spesies : Corn Harvest Year : 2018 Harvest period : 4 Land area : 30 m2 Total weight : 20 kuintal Product : corn seed	a. Search link Add Data b. Entry some fields in the menu add data c. Save the new data which is entered in the previous step	After succeceed to find link Add Data, respondent should entry some data in the textboxes and save it. There will be a pop up messages	219.6	13
5	See detail information of onion plant data.	Find and click link Show	System will show detail information of onion plant data	98.76	5
6	See the family of plant morphology	a. Find plant morphology b. Find the family data of specific data	System will show family of given plant data	74.6	3
7	Print all data of planting calendar	a. Find link to print data and download the data b. Search data in the download folder and show the data	System will show the downloaded data	56.2	5
8	See the end of the corn season on the land map menu	Find the end month of the spesific plant season	System will show the end month of the season	93.28	3
9	Show the harvest graph of rice in 2017	a. Find the link of Harvest graph b. Fill the information on the textboxes		36.76	5
10	See the summary data of the harvest graph	Find the Summary Links		14.76	1
11	Print to excel format the harvest graph which is got from task 10.	a. Find the link “Print” b. Show the data on the excel format	System will show the graph, download the data, and show on the excel format	21.24	3

TABLE II. TESTED TASKS FOR SYSTEM ADMINISTRATOR RESPONDENT

No	Description	Things to Do	Success Criteria	Maximum Time (in second)	Minimum Clicks
1	Login to Crop and Farmer Activity Information System Username : us1 Password : 123	a. Find and click Login link b. Fill some textboxes from login page	Respondent is able to login to the system and system shows WELCOME message	13.7	4
2	Add corn detail morphology ID : Corn Name : Corn Morphology Division : Angiosperm Subdivision : Zea Ordo : Poales Familiae : Poaceae	a. Find and click Link : Add Data from the system b. Fill some textboxes with the data c. Save the new data	Respondent will get messages Data is succeed to be entered and system will show the data	82.9	10
3	Edit data of Corn Morphology	a. Find and click link Edit b. Update some spesific data of morphology lists c. Save the changes	System will show the pop up messages to indicate that the data is successfully to be updated.	49.8	5
4	Print the plant data and show the downloaded data	a. Find and click Print link b. Find the downloaded data	System will show the data	29.1	6
5	Find and see the corn planting calendar in Bantul region	Search and click link SHOW	System will show the complete corn planting calendar data	31.6	4
6	Print the data of land map and show the data	a. Search and click link PRINT b. Search and print data	System will download data on the excel format	21.15	5
7	On the farmer activity menu, find the data persiapan1. Delete that data.	Search and click link DELETE	System will delete the data permanently	34.65	5
8	On the farmer activity menu, find the land preparation data : 12 and update the period become period : 8	a. Find the search bar b. Click update c. Do some changes in data d. Submit	System will show the pop up messages " Data is successed to be updated"	44.5	7

TABLE III. TESTED TASKS FOR GENERAL USER RESPONDENT

No	Description	Things to Do	Success Criteria	Maximum Time (in second)	Minimum Clicks
1	See detail information of onion plant	a. Search plant menu b. Search onion plant data c. Press button "Show"	System will show the detail information of onion plant	23.8	4
2	Print all plant morphology data	a. Search plant morphology menu b. Search and press button "Print"	System will download data into xls format	22.64	3
3	Show the printed data which has been done in task 2	a. Search the downloaded data b. Show the downloaded data	The data will be opened in spreadsheet application	26.04	2
4	Find the detail information of Corn on the planting calendar menu	a. Find the planting calendar menu b. Find button "Show"	System will show corn detail data and information	19.52	4
5	Find end month of corn season with latitude and longitude coordinat in the land map menu	a. Search land mapping menu b. Search corn data c. Find button "+" to see end month of corn season	System will show end month of corn season data with longitude and latitude coordinate	18.2	3

No	Description	Things to Do	Success Criteria	Maximum Time (in second)	Minimum Clicks
6	Look at the graphic detail of plantation data.	a. Search graphic plant menu b. Choose and click plantation graphic	System will open the plant data and show the graphic	31.76	3
7	Look at the potato detail information	a. Search potato data b. Search and click button "Show"	System will open potato data	20.96	1
8	Download the type of soil graphic on the PNG format	a. Search the type of soil data b. Download and Open the graphic	System will download the graphic and open it	29.64	4
9	Open the PNG graphic	Search and open the downloaded data	Monitor will show the graphic	9.24	1
10	Search the cassava data on the farmer activity	a. Find the search bar b. Fill the search bar c. Click and press button "Search"	System will show the data that contains word: cassava	47.12	4
11	Search summary data from regional activity graphic. Year : 2016 Plant : Onion Activity : Land Preparation	a. Search regional activity graphic b. Fill the textbox c. Click and press button "Submit" d. Find and search button "Summary"	System will show summary garlic data	60.48	7
12	Find the number of activities on plant activity graphic with the following data : Province : Special Region of Yogyakarta Year : 2015 Activity : Land Preparation	a. Find and search plant activity graphic b. Fill some textboxes c. Click "Submit" button d. Direct mouse to the graphic	System will show the number of land preparation activity	66.32	8
13	Filter the harvest crop based on name. Name : rice	a. Find the harvest crop menu b. Fill some textbox filter c. Find and press filter button	System will show the data that contains word rice	100.4	6
14	See and show the ammount of rice harvest graphic in Yogyakarta Special Region	a. Find the menu b. Fill the textbox c. Press submit button d. Direct the cursor to the new data		44..84	6
15	See the ammount of rice harvest in Yogyakarta Special Region from 2012 – 2016	a. Find the menu of harvest data b. Fill some textboxes c. Press submit button	System will show the ammount of harvest graphic based on data entered	99.4	12
16	See the rice plant data and the amount of harvest crop in 2017	a. Find the graphic menu on detail information of harvest b. Fill some data and click submit button c. Find and click Summary data	System will show summary and the number of rice plant harvest in Special Region of Yogyakarta on 2017	48.4	1

IV. Analysis and Discussion

A. Descriptive Statistics

There are 58 respondents participated in this research: 49 farmers, 3 system administrators, and 6 general users. Based on age, those respondents can be categorized as follows:

Table IV shows the number of respondent based on age category. Almost 50% of respondent's age is 20 to 30 years old. There are 9 respondents are 41 to 50 years old as a farmer. There are also 3 general user respondents which are aged more than 51 years old.

TABLE IV. Respondent Profile Based on Age

No	Age	Number of Respondent
1	< 19 years old	0
2	20 – 30 years old	24
3	31 – 40 years old	22
4	41 – 50 years old	9
5	> 51 years old	3

TABLE V. RESPONDENT PROFILE BASED ON EDUCATION BACKGROUND

No	Education	Number of Respondent
1	Junior High School	30
2	Senior High School	25
3	Undergraduate Degree	2
4	Graduate Degree	1

Table V shows the number of respondent based on education background. There are 30 respondents which are junior high school graduates. Those are farmer respondents. There are also 19 farmer respondents who graduate from senior high school. System administrator respondents graduate from senior high school. The rest are general user respondents and system administrator. There is a general user respondent who already completed master degree.

B. Testing Result

This research involved 58 respondents to finish several tasks to measure task success, time on task, error, and efficiency (number of click). Time on task is a measurement for user effectiveness to finish a task. While task success is a measurement of how many respondents are able to finish the task. Error is to calculate how many mistakes respondents did. And the last is efficiency as a measurement of how many efforts respondents did in a system. In this research, effectiveness is measured by counting the number of click respondent did during finishing a task.

1) System administrator respondents

The first step to conduct usability testing is measuring maximum time to do the tasks. Maximum time is got by taking 4 random samples and double average times to finish tasks.

Table VI shows the maximum time limit to do tasks. Each task has different maximum time. If the respondents are not able to finish task until the maximum time, respondents will be considered and assumed as an error. After completing the task, there are 3 respondents who are not able to finish task 4 on time, 2 respondents who are not able to finish task 5 on time, and 1 respondents who are not able to finish task 6 on time.

Table VII shows the success rate of system administrator task. Those task is filled with 1 if the respondent is succeed to finish the task and 0 if the respondent is failed to finisih the task. It is seen that there is no respondent who is able to finish the task 4 and only 1 respondent who is able to finish task 6. Respondents are not able to finish task 4 since there is no symbol on the page menu or active label if the page is active. Thus, respondents faced difficulties while doing the task.

Contrary to the success rate calculation, error rate is to measure the respondents' error while finishing tasks. On task 4, respondents were not able to show crop menu. Respondents show crop / plant morphology data. Respondents are also not able to show the data which is given from the task.

The last measurement is efficiency (number of click). Efficiency is to measure how much efforts respondents do in a system. The number of click is compared with minimum click to finish task. The minimum click to finish the task is 46. Table VIII below shows efficiency of system administrator by counting number of click.

Table VIII shows the system administrator efficiency. The number of click is slightly above the minimum number of click. The respondents explore the system functionality well. It makes respondents are able to click the tasks efficiently.

TABLE VI. BENCHMARK MAXIMUM TIME FOR SYSTEM ADMINISTRATOR

Task	User 1	User 2	User 3	User 4	Avg	Max Time (Avg * 2)
1	7.3	8	5.3	6.8	6.85	13.7
2	33.8	49.3	38.9	43.8	41.45	82.9
3	18.8	25.5	22.5	32.8	24.9	49.8
4	10.7	17.8	13.4	16.3	14.55	29.1
5	16.6	11.3	17	18.3	15.8	31.6
6	12.6	9.1	10.3	10.3	10.575	21.15
7	14.2	13.9	22.7	18.5	17.325	34.65
8	21.9	18.1	23.5	25.6	22.275	44.5

TABLE VII. SYSTEM ADMINISTRATOR TASK SUCCESS

User	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Success
1	1	1	1	0	1	0	1	1	75%
2	1	1	1	0	1	0	1	1	75%
3	1	1	1	0	1	1	1	1	88%
Succes	100%	100%	100%	0%	100%	33%	100%	100%	79%

TABLE VIII. SYSTEM ADMINISTRATOR EFFICIENCY

User	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Sum
1	4	12	9	6	7	4	6	9	57
2	4	12	5	4	5	4	8	6	48
3	4	10	5	4	5	8	5	7	48

TABLE IX. BENCHMARK MAXIMUM TIME FOR FARMERS

Task	**User 1**	**User 2**	**User 3**	**User 4**	**User 5**	**Avg**	**Max Time (Avg * 2)**
1	14.7	15.6	11.3	14.8	21.9	15.66	31.32
2	15.9	43.9	17.3	69.6	96.1	48.56	97.12
3	21.2	67.8	30.4	39.8	68.9	45.62	91.24
4	59	124.5	68.5	132.3	164.7	109.8	219.6
5	24.9	61.3	30.8	40.8	89.1	49.38	98.76
6	16.9	30.4	29.9	43.8	65.5	37.3	74.6
7	18.8	27.7	15.6	34.6	43.8	28.1	56.2
8	59.4	10.9	22.4	65.8	74.7	46.64	93.28
9	12.4	12.8	10.6	23.4	32.7	18.38	36.76
10	3.9	7.3	5.5	8.4	11.8	7.38	14.76
11	9.6	7.1	6.5	12.5	17.4	10.62	21.24

2) Farmer respondents

There are 49 farmer respondents are involved in this research. Among those 49 respondents, there are 13 data is invalid because the respondents got help from the others to finish the task. Those invalid data are not analyzed. 5 respondents are chosen to benchmark maksimum time.

Table IX shows the maximum time limit to do several tasks. Based on those benchmark, it is shown the number of successful respondents based on time limitation.

Table X shows the successful respondents to finish the task based on the time limitations. Some tasks has good success rate. On the contrary, less than 50% respondents is not able to finish task 2, 5, 6, and 8. From the task 6 test, respondents are difficult to find edit button. Some of respondents think that '+' button is edit button. On task 9, 10, and 11 respondents should scroll down the page to find the summary button. From this task test, it is suggested to refine the system user interface to ease user to access the menu.

TABLE X. NUMBER OF RESPONDENTS WHICH NEEDS TIME ABOVE AND BELOW MAXIMUM TIME LIMIT AND AVERAGE TIME TO FINISIH TASK

Task	**Number of Respondents Below Max Time**	**Number of Respondents Above Max Time**	**Average Time Used**
1	30	6	6.197
2	11	25	43.519
3	21	15	28.533
4	12	24	97.025
5	6	30	48.427
6	17	19	19.036
7	18	18	21.261
8	9	27	28.172
9	26	10	7.458
10	32	4	1.411
11	32	4	2.25

TABLE XI. SUCCESSFUL RATE OF EACH TASK

Task	**Success Rate**
1	100%
2	89%
3	94%
4	92%
5	94%
6	94%
7	75%
8	78%
9	97%
10	94%
11	94%

Task Success is calculated with binary success. If the success rate percentage is bigger than 78%, the task can be completed properly. Table XI shows the success rate of each task given to the respondents.

Table XI shows the success rate of each task. When doing the task 7, some respondents did some mistakes by choosing the wrong data to be filtered. Respondents also print and display the wrong planting calendar. Another variable to be analyzed is farmer respondent's efficiency. Average click to finish all tasks is 59 while the predetermined value for farmer respondents to finish the task is 51. It means that there is difference on both two values. It is found during the test that some respondents faced difficulties to find some buttons on its page.

3) General user respondents

Similarly with Farmer respondents and system administrators, researchers carried out analysis for benchmark maximum time limit. It is determined by calculating average time of 5 sample respondents for each task.

Table XII shows the maximum time for general user respondents. There are 6 respondents for general user. After finishing the test, the data is analyzed to get the number of successful respondents and unsuccessful respondents. Table XIII below shows the number of successful and unsuccessful respondents for each task.

TABLE XII. BENCHMARK MAXIMUM TIME FOR GENERAL USERS

Task	User 1	User 2	User 3	User 4	User 5	Avg	Max Time (Avg * 2)
1	12.4	13	8.6	10.3	15.2	11.9	23.8
2	8.7	10.1	10.3	12.7	14.8	11.32	22.64
3	2.3	6.6	19.3	13	23.9	13.02	26.04
4	11.9	6.6	8.8	7.9	13.6	9.76	19.52
5	6.8	7.6	11.9	6.9	12.3	9.1	18.2
6	17.7	17.4	4	16.6	23.7	15.88	31.76
7	6.8	5.8	18.4	8.9	12.5	10.48	20.96
8	18.9	14.7	11.9	12.2	16.4	14.82	29.64
9	3.7	5.2	4	4.2	6	4.62	9.24
10	20.1	25.8	18.4	23.9	29.6	23.56	47.12
11	17.1	37.1	10.8	40.3	45.9	30.24	60.48
12	25.2	23.1	17.3	30.8	69.4	33.16	66.32
13	40.2	43.2	39.9	50.8	76.9	50.2	100.4
14	26.9	15	13.6	17	39.8	22.46	44.84
15	40.8	39.5	48	44.3	75.9	49.7	99.4
16	27.8	21.1	15	24.4	32.7	24.2	48.4

TABLE XIII. NUMBER OF RESPONDENTS DID TASK BELOW AND ABOVE AVERAGE TIME AND AVERAGE TIME TO FINISIH TASK

Task	Number of Respondents Below Average Time	Number of Respondents Above Average Time	Average Time Used
1	3	3	12.5
2	5	1	16.18
3	6	0	9.016
4	2	4	16.1
5	1	5	22.4
6	2	4	15.75
7	6	0	9.25
8	4	2	18.24
9	6	0	5.53
10	2	4	32.15
11	3	3	38.7
12	6	0	27.85
13	5	1	51.38
14	4	2	23.85
15	5	1	53.16
16	3	3	19.73

Table XIII shows the number of successful and unsuccessful respondents finishing the task before maximum time. There are some tasks with the number of unsuccessful respondents are so high (more than 3 respondents). While doing the test task 1, 4, and 5, respondents were not able to find the display button and made respondents are not able to finish the tasks on time. Respondents found similar problem with display or detail button on task 6. Another case found on this test is respondents found difficulty to use search bar. Respondents are also hard to find “Summary” button since this button is located on the bottom of page. Respondents should scroll down the page. There were some mistakes that the respondents did: (1) respondents are not able to visit summary page; (2) respondents are failed to filter based on specific keywords; (3) respondents are wrong to choose the appropriate menu.

TABLE XIV. SUCCESSFUL RATE OF EACH TASK

Task	Success Rate
1	100%
2	100%
3	100%
4	100%
5	100%
6	100%
7	100%
8	100%
9	100%
10	66.67%
11	50%
12	100%
13	83.33%
14	66.67%
15	83.33%
16	66.67%

TABLE XV. SUCCESSFUL RATE OF EACH TASK

Respondents	Average			
	Time Per Task	Task Success	Clicks	Efficiency
System Administrator	52%	79%	95%	50%
General Users	66%	84%	70%	83%
Farmers	66%	91%	69%	91%

Table XIV shows success rate of each task. Task 1-9, there is no respondent made errors during finishing the task. Otherwise, there is 33.33% respondents who did some errors finishing the task 10, 14, and 16. There is 16.67% respondents who are not succeed to finish the task 13 and 15. While 50% respondents are failed finishing the task 11.

Minimum number of click to finish tasks is 69 and the average of respondents click finishing the task is 101. This minimum number of click cannot be achieved since respondents are difficult to find the summary button. The location of the button is not able to be seen clearly.

C. Combining Metrics based on Percentage

Table XV shows the metric based on percentage. This technique is to combine different scale of value, convert it into percentage, and calculate the average of each parameters (time per task, task success, number of clicks, and efficiency). Based on table XV, average time to do task for system administrator is 52% from maximum time. The usage of option button cannot be understood easily for the respondents. System Administrator respondents only 79% are succeed to finish tasks because some respondents did the activities on the wrong page. The number of clicks is 95%. It means that users click the correct button. On the contrary, the accuracy is only 50% due to system administrator respondents are not able to finish some tasks.

General user respondents conducted and finished the task well. The average time to do task is 66% from the maximum time with the success rate around 84%. There are some failed activities which users are not able to finish it because of difficulties to translate English button, for example submit button. While the number of click is 70% and the efficiency is 83%. The average number of efficiency for general users is higher than system administrator.

Farmer respondents finished tasks with the average time 66%. Some respondents are failed to finish the task because of the usage of the wrong symbol. Thus, the respondents are not able to understand the meaning. User interface of the system does not provide good information to the respondents. On the contrary, the number of succeed respondents to finish the tasks is 91% with efficiency around 91%. It means that most of the farmer respondents are able to finish the task but some of them need additional time to finish the task.

D. Improvement Recommendation

Improvement recommendation is given according to the task test result, feedback from the respondents, and the analysis of the data from the task test result. Table XVI shows the improvement recommendation for the systems.

Table XVI shows the improvement recommendation for each user. Recommendations are categorized into 3 users. Some recommendations are the same, for instance in each user, there is a recommendation to display current active username on page, additional icon is required for each button, different colour between header and footer.

TABLE XVI. IMPROVEMENT RECOMMENDATION

Respondents	**Recommendations**
System Administrator	1. Menu and navigation should use different colour 2. Active menu should be highlighted 3. System should display that the system is an agriculture system 4. Information about how to care of plant should be available 5. Additional picture for each plant is required 6. There should be changes in menu naming 7. User interface makes some distractions 8. Searching is not addressing what is being sought 9. Icon is required on each button 10. Navigation should be changed to accordion type that will made easy to be scrolled. 11. Header and footer should use different colour 12. Active username should be visible
General Respondents	1. Icon (+) can be replaced with icon (▼) 2. Active menu should be highlighted 3. Accordion menu type should be used in order to ease user to scroll 4. Active username should be visible 5. Sort function will be added with icon and background 6. Header and footer should use different colour
Farmers	1. Icon is required on each button 2. Indonesian language should be used in each button 3. Button edit, delete, and summary should be visible to the user 4. Accordion menu type should be used in order to ase user to scroll 5. Active username should be visible 6. Summary button will be placed closer to the content in order to ease user finding the button 7. Distance between column in each table will be narrowed down. User does not need to click additional button

E. Improvement Result

Improvement recommendation has been generated. Based on those recommendations, researchers try to fulfill the improvement needs. The improvements are arranged based on the recommendations for all 3 category user.

1) Highlighting the active menu

Figure 6 shows the changes before and after the revision. On the left, it is shown that active menu was not highlighted. On the right, it is shown that active menu has been highlighted.

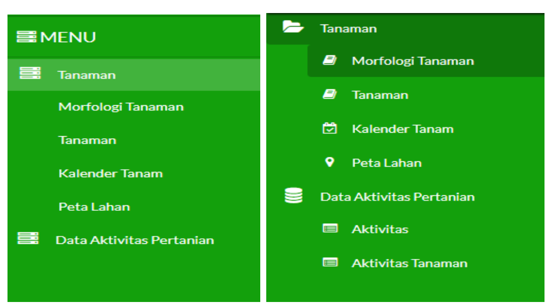
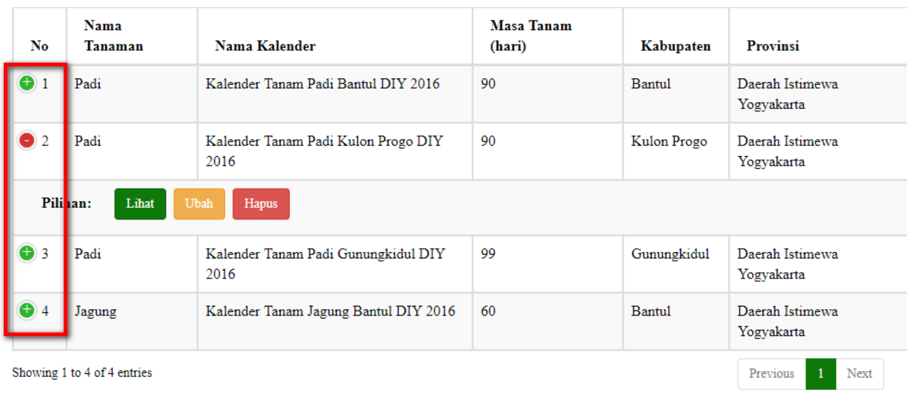


Fig. 6. Menu before the Revision (Left) and after the Revision (Right).

2) Changes in Icon (+) to Icon (▼)

Figure 7 shows the icon (+) before the changes. That icon has been replaced with icon (▼) which is displayed in figure 8.

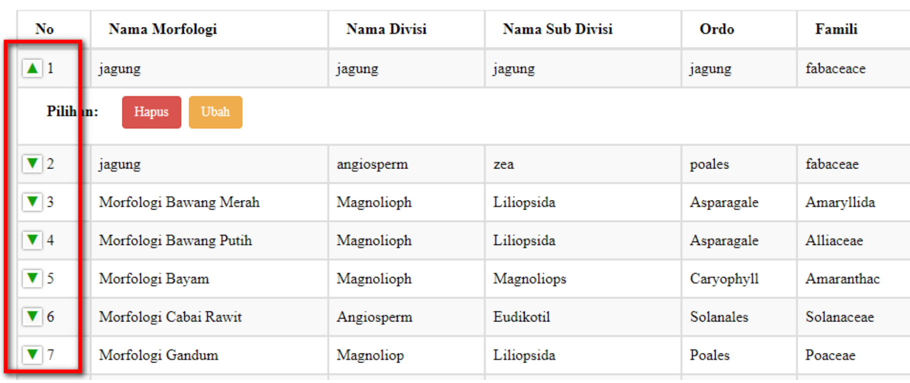
The changes also include removal of show button. Show button has been removed (figure 8). User is only able to delete and edit the record.



No	Nama Tanaman	Nama Kalender	Masa Tanam (hari)	Kabupaten	Provinsi
1	Padi	Kalender Tanam Padi Bantul DIY 2016	90	Bantul	Daerah Istimewa Yogyakarta
2	Padi	Kalender Tanam Padi Kulon Progo DIY 2016	90	Kulon Progo	Daerah Istimewa Yogyakarta
Pilihan: Lihat Ubah Hapus					
3	Padi	Kalender Tanam Padi Gunungkidul DIY 2016	99	Gunungkidul	Daerah Istimewa Yogyakarta
4	Jagung	Kalender Tanam Jagung Bantul DIY 2016	60	Bantul	Daerah Istimewa Yogyakarta

Showing 1 to 4 of 4 entries Previous 1 Next

Fig. 7. Icon (+) before Recommendation Changes.



No	Nama Morfologi	Nama Divisi	Nama Sub Divisi	Ordo	Famili
1	jagung	jagung	jagung	jagung	fabaceace
Pilihan: Hapus Ubah					
2	jagung	angiosperm	zea	poales	fabaceae
3	Morfologi Bawang Merah	Magnolioph	Liliopsida	Asparagale	Amaryllida
4	Morfologi Bawang Putih	Magnolioph	Liliopsida	Asparagale	Alliaceae
5	Morfologi Bayam	Magnolioph	Magnoliops	Caryophyll	Amaranthac
6	Morfologi Cabai Rawit	Angiosperm	Eudikotil	Solanales	Solanaceae
7	Morfologi Gandum	Magnoliop	Liliopsida	Poales	Poaceae

Fig. 8. Improvement Result of Icon and Display Button.

3) Moving the searching text box to make it visible to user

Some respondents had problem while they should search specific data. They found difficulties to search and the menu is required to be moved.

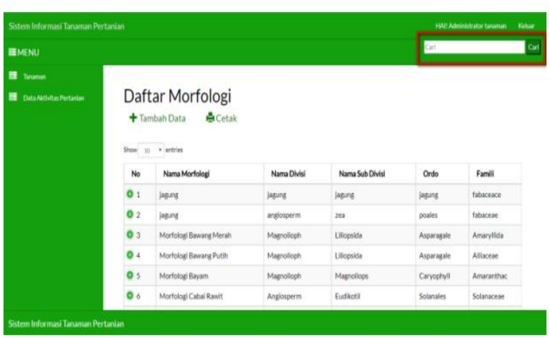


Fig. 9. Search Menu Location before Improvement Changes.

Figure 9 and 10 shows the location of the searching menu. Searching menu before the improvement changes is located on the left top on each page. After the improvement changes, searching menu is located little bit down. This movement is based on the test recommendation. While doing the task test, respondents needed extra time to find the searching menu.

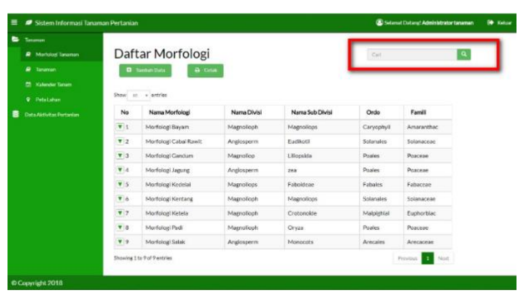


Fig. 10. Search Menu Location after Improvement Changes.

4) Changes on add, print, sorting menu, and searching menu

Farmer respondents found difficulties while adding, printing, sorting, and searching data. Therefore, additional icon is needed to ease user. Figure 11 shows the menu before changes and figure 12 shows the menu after changes.

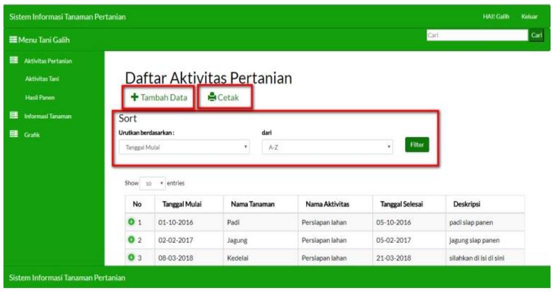


Fig. 11. Add, Print Button, Sorting Menu, and Searching Menu before Changes.

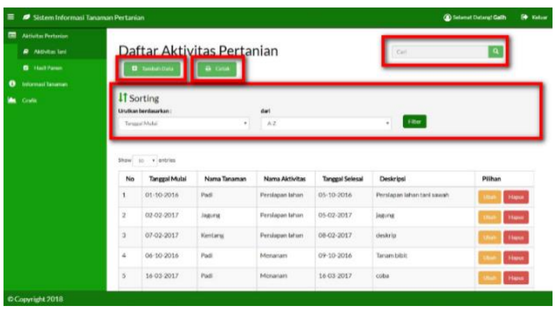


Fig. 12. Add, Print Button, Sorting Menu, and Searching Menu after Changes.

Figure 11 shows the add button, print, sorting menu, and searching menu before changes. Button should be changes to ease the user while searching menu should be moved down slightly. Figure 12 shows the improvement changes for those buttons and menus.

V. Conclusions

Based on the research, there are some conclusions as follows:

1) Usability problems for agriculture activity information system are the usage of inappropriate button icon, button position does not locate on the right position, and navigation of menu sidebar does not display current active menu. User also faced difficulties problem with symbol button.

2) Task test from usability analysis for agriculture activity information system: three kinds of respondent (system administrator, farmer, and general user) assess that agriculture activity information system is good with average overall assessment value 69% - 79%.

3) There are some improvement recommendations that will improve system usability. Those recommendations has been followed up by doing several refinements.

References

[1] The World Bank, "Priority Issues for Indonesian Agriculture," The World Bank, 2005.

[2] The Organisation for Economic Co-operation and Development, "OECD Economic Surveys Indonesia," The Organisation for Economic Co-operation and Development, 2015.

[3] Deloitte, "e-Transform Africa: Agriculture Sector Study - Sector Assessment and Opportunities for ICT," Delloite Project Report, 2012.

[4] R. Delima, F. Galih and A. Wibowo, "Development of Crop and Farmer Activity Information System," *Researchers World*, vol. VIII, no. 4, pp. 180 - 189, October 2017.

[5] T. K. Chiew and S. S. Salim, "Webuse: Website Usability Evaluation Tool," *Malaysian Journal of Computer Science*, vol. 16, no. 1, pp. 47-57, 2003.

[6] N. Bevan, "Measuring Usability as Quality of Use," *Software Quality Journal*, vol. 4, pp. 115-150, 1995.

[7] J. Mvungi and T. Tossy, "Usability Evaluation Method and Principle for the Web," *International Journal of Computer Science and Software Engineering*, vol. 4, no. 7, pp. 165-171, 2015.

[8] R. L. Mack and J. Nielsen, Usability Inspection Methods, New York: John Wiley & Sons, 1994.

[9] N. F. M., EL-firjani, E. K. Elberkawi and A. M. Maatuk, "A Method for Website Usability Evaluation: A Comparative Analysis," *International Journal of Web & Semantic Technology*, vol. 8, no. 3, pp. 1-11, 2017.

[10] J. Nielsen, "Usability 101 - Introduction to Usability," 4 January 2012. [Online]. Available: https://www.nngroup.com/articles/usability-101-introduction-to-usability/. [Accessed 16 September 2018].

[11] W. E. Y. Retnani, Y. P. Prayogi, B. Prasetyo and A. N. M. M. A. R, "Usability Testing to Evaluate the Library's Academic," in *International Conference on Computer Applications and Information Processing Technology*, Bali, 2017.

[12] S. Lestari, "Analisis Usability Web (Studi Kasus Website UMKM Binaan BPPKU KADIN Kota Bandung)," *Jurnal Ilmiah Teknologi Informasi Terapan*, vol. 1, no. 1, pp. 46-51, 2014.

[13] M. Ashraf, F. S. Cheema, T. Saba and A. Mateen, "Usability of Government Websites," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 8, pp. 163 - 167, 2017.

[14] N. K. Kiget, G. Wanyembi and A. I. Peters, "Evaluating Usability of e-Learning Systems in Universities," *International Journal of Advanced Computer Science and Application*, vol. 5, no. 8, pp. 97-102, 2014.

[15] D. P. Sari, N. Susanto and Z. L. Z. Polii, "Analysis of the Community Website Usability Using Koohang's Model Based on Task Performance Comparasion Result," in *International Conference of Industrial, Mechanical, Electrical, Chemical Engineering*, Surakarta, 2016.

A Novel Architecture for 5G Ultra Dense Heterogeneous Cellular Network

Sabeen Tahir
Information Technology Department, Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah 21589, Saudi Arabia

***Abstract*—The mounting use of wireless devices and wide range applications in an ultra-dense heterogeneous wireless network has directed to the challenging circumstances that could not be handled till 4G. In order to deal with the critical challenges, the fifth generation (5G) wireless network architecture requires an efficient well-organized wireless network. In this paper, a novel architecture for the 5G ultra-dense heterogeneous cellular network is proposed. In the proposed architecture two main aspects, massive MIMO-OFDMA and IP-based vertical handover are considered. In order to have full network coverage, the whole macro area network is divided into the microcells and each microcell is further divided into the smaller cells. The heterogeneity of different types of base stations (macro area network base station, micro-cell base station, and small cell base station) provides efficient network coverage. By reducing the area of the cell, the frequency efficiency and network coverage are improved. All the base stations are equipped with the massive MIMO-OFDMA antennas and different radio access technologies so a single wireless device can switch from one radio access technology to the other. In order to prevent the link disconnection and save IP address, whenever wireless devices need to perform vertical handover first a new connection is established with new radio access technology. It is important to note the same IP address is used whereas the current connection is disconnected. By utilizing IP-based vertical handover, the new 5G wireless network can become the principal goal of service continuity and minimize the handover processing delay. The simulation results show the improvement in the network performance.**

***Keywords*—Massive MIMO-OFDMA; 5G; IP-based interoperability; heterogeneous**

I. Introduction

The inventive and efficient use of Information and Communication Technologies (ICT) is becoming more attractive and significant to advance the economy of the world. A wireless communication technology plays a role of the most critical element in the global ICT and it is a fast-growing trend in the world. The main requirements of the 5G system will have to include the simultaneous full-duplex transmission, use of massive multi-input - output (massive MIMO). In addition, millimeter wave communication, cloud-based radio access network and energy efficient technologies due to the energy starving multimedia applications [1]. Due to the promising trends of the frequency virtualization and energy efficient communication, the 5G network's structural design will be more complex than 4G network design [2, 3]. Finally, minimize delays in the network to improve the overall performance [4]. All the current mobile wireless networks are going to adapt the all-IP (Internet Protocol) principle, which means that all the data and signaling will be transmitted via IP (a unifying technology) on the internet layer [5, 6]. Recent years have seen incredible an enlargement in mobile data traffic and this trend is anticipated to continue in the future. To handle this enlargement, dense Transmission Points (TPs) e.g. Ultra-Dense Networks (UDNs) need to be deployed. The 5G cellular networks have to support the ubiquitous mobile broadband wireless access and advance the handling of extremely dense and heterogeneous networks. Fig. 1 shows an example of six typical UDN scenarios [7]. An extremely dense heterogeneous wireless network requires efficient resource management of applications and formation of efficient approach offering energy efficiency, scalability, and flexibility [8].

The upcoming 5G wireless networks will highly base to the highest degree of density and heterogeneity. Therefore, networks need to have an effective density and heterogeneity controlling techniques to handle the increasing amount of data traffic and wireless devices scattered in a skillful manner. The data traffic load of wireless network increases than the limit of future wireless networks. It is observed that an analogous connection exists between interference and scattered networks, where the whole throughput is not in line through the maximum scattered networks. Hence, the network scalability is a serious issue for the future 5G wireless network [9-11]. The energy efficiency can be monitored from the base stations and wireless mobile devices. From the base station perspective, it is worth controlling solutions that decrease energy consumption by connecting wireless devices into as few as the possible or frequent reconstruction of the power cycling method. Therefore, it requires an efficient technique that increases energy efficiency by getting an enhanced trade-off between the expected data traffic and existing energy. From the other perspective of wireless devices, the energy-efficient techniques maximize the energy efficiency.

Fig. 1. Typical UDN Scenarios [7].

The future 5G extremely dense wireless network needs to improve the flexibility with an efficient control of energy consumption. Security is another perspective that conflicts with the scalability and energy efficiency. The wireless connections mostly interrupted due to unwanted interference or security attacks. These interruptions lead to some harmful effects with communication delay. It is an extremely challenging job to tackle scalability, flexibility, and energy competency at the same time as these requirements are in conflict with each other. The improvement in one can lead to the degradation of the other requirements. Suppose security is increased, but with the greater processing overhead, reduced scalability and energy efficiency. If the energy efficiency improves by switching off various network equipment, then the network scalability may be affected [12-14]. The proposed 5G network system will have the ability to control many base stations positioned dynamically in a heterogeneous manner. A macrocell overlaid with the microcells and small cells to carry the enhanced spectral efficiency and coverage within a network. The key assumptions of the proposed technique are 1) all the base stations are equipped with the massive MIMO-OFDMA antennas, 2) a user will have the ability to switch between different RATs without changing their IP address from a single wireless mobile device 3) the wireless devices can perform vertical handover efficiently without disconnecting their links.

This paper targets to develop a new operational design for the 5G network, Section 2 describes the related work and highlights the problems in current wireless techniques. In Section 3, the proposed 5G dense heterogeneous wireless network is explained. Simulation results are given in Section 4. Challenging in the 5G are discussed in Section 5. Finally, in Section 6 the future challenges and conclusions are drawn.

II. Related Work

The use of extra frequency band will direct to an increasingly mobile frequency use in 5G. Advanced wireless techniques will facilitate drastically efficient frequency by using synchronized multipoint and small cell etc. For the sharing of frequency, a mobile architecture will allow the increase of multiplexing, the density of base stations and network heterogeneity. The previous generations 3G (UMTS and HSPA) and 4G (LTE and LTE-Advanced) focused on creating new radio transmission techniques. The 3G has the ability to handle complex data and support high data rates. The 4G network is expected to present comprehensive, secure internet access, gaming and streamed multimedia services. Now the research directs to the 5G, in order to highlight the user-oriented challenges 5G network must focus on the network densification into an ultra-dense environment. It should enhance the features of mobile devices, including autointegration and self -management capabilities. Furthermore, high consistent connections put inflexible latency and reliability requirements on the structural design.

Multi-Input Multi-Output (MIMO) is a wireless radio antenna technique that utilizes multiple transmitters and receivers to broadcast more data at the same time. A base station of massive MIMO technique is equipped with a large number of antennas serving many wireless devices. It allows multiple antennas to transmit and receive multiple spatial streams and combine data streams receiving from different paths in order to increase the receiver signal capturing the power and capacity of a channel. MIMO played an imperative role in 3G and 4G wireless network [15, 16]. Some other current techniques have designed the highly energy efficient massive MIMO [17]. The energy efficiency of a massive MIMO system depends on the circuit energy consumption. The best broadcast power increases with the number of antennas, so a broadcast power is a significant design parameter for high energy efficiency massive MIMO system [18]. Antennas are generally located on the surface of the carrier to get the wanted electromagnetic energy [19]. The MIMO technique has the ability to increase the throughput even under the conditions of signal fading, interference, and multipath [16, 20]. Millimeter wave communication is a potential 5G cellular network technique, which is expected to allow hundreds of MHz bandwidth [21]. The millimeter wave technology has some limitations such as not helpful for all kinds of wireless applications and millimeter wave technology is its limited range. According to the physics law “the shorter the wavelength, the shorter the communication range for a given power”. By the side of logical power in many cases, the limitation contains the range less than 10 meters. The free space loss in dB can be calculated as below:

$$L = 92.4 + 90\log(f) + 20\log(R) \quad (1)$$

Where R is considered as the Line-of-Sight (LoS) distance between transmitter and receiver antennas in kilometers, and f is the frequency in gigahertz. For instance, the loss at 10 m at 60 GHz is:

$$L = 92.4 + 35.6 - 40 = 88dB \quad (2)$$

This loss can be overcome with good receiver intensity and high antenna's gain. The other limitation of the millimeter wave technique is atmospheres that can absorb it and limit its range. The signal attenuation is very high in the air due to the rain, fog, and humidity [22]. METIS has identified the 5G mobile communication scenarios. These situations reflect the predicted challenges like high-data-rate, accessibility, mobility, a huge amount of devices, low latency and consistency [23]. The objective of this project is only to combine the heterogeneous wireless technologies to optimize the range of radio devices. However, problems with this approach are, it does not provide the solution how to stable 5G requirements, which are contradicting each other. METIS's structural design mostly focused on hardware communication techniques. Although, the band utilization is mentioned it does not state anything regarding what type of technique will be in employment to reach an efficient and safe spectrum allocation [8].

A management architecture "Wireless Software-based architecture for Extremely Dense networks" (WiSEED) for the 5G system is proposed that handles three operational services routing, device mobility, and frequency usage. The aim of WiSEED is to handle the contradictory requirements such as scalability, flexibility, and energy efficiency in order to give a high quality and ubiquitous services for mobile broadband internet access in a dense environment. WiSEED focuses on the perspective of software networking, it occupies technologies and operational services based on the perception to support the envisage architecture. The recourse demand for all the advanced applications will be increased for the 5G network. The 5G network should have to advance the wireless transmissions and service rates like 4G. Therefore, 5G networks are supposed to permit transmissions between anybody (person-to-person), anything (person-to-machine, machine-to-machine), wherever the devices are and whatever electronic/wireless devices need [24, 27].

Internet Sockets are the endpoints for data communication streams and these are used to recognize all the communication links between nodes. Each web socket is a unique combination of local IP address, proper transport communication port, target IP address, target communication port, and type of transport protocol. In a heterogeneous wireless network, when wireless mobile devices change their Radio Access Technology (RAT) or perform vertical handover then by default IP addresses of respective devices changes. When the devices connect to other RAT, the devices consume more resources and create a delay. The change of RAT means that the change of IP address, the change of any parameter in a socket means that disconnects the current socket and connect to the new socket (new communication link). This is the main problem with the current internet connections. To overcome this critical issue in a 5G network, IP addresses of the source and destination devices must be fixed and unchanged in case of any interoperability. By this way, it will ensure the handover transparency and preserve the proper layout data packets.

In order to accommodate the high demand for network data traffic, a new 5G network will feature by the frequency reuse, an extreme number of wireless devices, and multiple base stations, etc. Therefore, the system requires more attention to enhance interference management techniques. There are different types of interferences that can badly affect the 5G wireless network [25]. Other-Cell Interference, the devices which are at the edge of the cell usually suffer two types of interference. The first is self-cell interference from the other devices in the same cell and the second is other-cell-interference from the devices of other cells. Cross-tier Interference: in a multi-tier cellular communication network, the interference from one tier to another is indicated as cross-tier interference, it degrades the system performance. Intersystem Interference: the interferences from the signal transmitted by the other system. To control these types of interferences there are some interference coordination techniques like beamforming, power control, user scheduling, and advanced receiver techniques. In a 5G network, carrier aggregation and use of the multi-RAT will be directed to an enhanced mobile spectrum usage. Different radio technologies will empower higher spectrum efficiency by using small cells and massive MIMO. The frequency allocation and mobile infrastructure both can improve the statistical multiplexing and enhance the density of the base station per user. These techniques require a shared backhaul network for the mobile user that further increase the network heterogeneity.

To enhance the current wireless technologies, this paper presents a novel 5G network architecture. Where a whole macro network area is divided into the microcells which are further sub-divided into the smaller cells in order to achieve the full network coverage. The required base stations are equipped with massive MIMO-OFDMA antennas. To prevent the disconnections of vertical handover and delay due to the interoperability, this technique proposed IP-based vertical handover. Where the devices will establish a new RAT link before disconnecting the current link.

III. A Novel Architecture for 5G Ultra Dense Heterogeneous Cellular Network

The proposed 5G network architecture considered the dense heterogeneous wireless network with the deployment of wireless base stations having massive MIMO with OFDMA antennas. The whole Macro Area Network (MAN) into the MicroCells (MCs), each MicroCell has covered by the powerful Micro Base Station (MC_BS). There is one powerful Macro Area Network Base Station (MAN_BS). All the MC_BS are connected with the MAN-BS, which is connected to the core network. The MC is further divided into the Small Cells (SCs) in order to achieve the full network coverage. Each SC has its Base Station (SC_BS), SC is used to upgrade the performance of MAN by offloading the data traffic generated in the network. The degree of integration determines the network performance. The proposed architecture is IP based because it prevents the vertical handover disconnections between mobile devices in case the cells or RAT is changed. The proposed architecture allows switching between different

RATs. The general layout of the proposed 5G ultra-dense heterogeneous network is shown in Fig 2.

The whole network architecture represents a multi-tier backhaul communication system of the ultra-dense network. The whole ultra-dense area is covered by the small cells as the small cells are designed to fill up the uncovered gaps in microcellular areas. These are physically smaller and covered limited areas. The most important question is that how much densification an SC-BS can cover? The whole micro area has divided into multiple SC where each cell has its own base station. As the microcell is an ultra-dense area, therefore in order to provide full coverage, all the wireless devices are connected to its nearest SC-BS. The base stations are equipped with Massive MIMO antennas therefore, a single base station has an ability to connect maximum users. The P_{loss} factor is used as a function of the probability of loss of the distance between transmitter and receiver (wireless mobile device- SC-BS). The probability loss P_{loss} is derived as:

$$P_{loss} = \min(\theta/d, 1) * \left(1 - e^{-\frac{d}{2}*\theta}\right) + e^{-\frac{d}{2}*\theta} \quad (3)$$

Where *d* is the distance between the mobile device and it's serving base station. According to the model, within the specified *Ɵm* theta meter devices connects to the base station. Within the specified limit the P_{loss} is 1. The devices positioned up to the *Ɵm* from the SC_BS has guaranteed a strong loss, therefore, the devices connect to the other SC_BS, which is in the proximity.

In order to establish a backhaul data communication link, SC_BS selects the neighboring MC_BS. An SC_BS has to satisfy the following two conditions:

1) The distance between transmitter and receiver is less than or equal to the radius rd of SC_BS,

2) The distance between SC_BS and MC_BS is less than the distance between the transmitter and MC_BS.

3) When the distance between MC_BS and SC_BS is less than rd then SC_BS establish a link. If conditions are satisfied as a result, the SC-BS efficiently allocates the available spectrum to the wireless devices. In addition, it overcomes the path loss due to the small distance and performs intense frequency reuse.

The ultra-dense Backhaul Network Capacity *B_NC* is estimated by the following formula [26]:

$$B\_NC = \frac{ST(n) \times Tr}{N(n)} \quad (4)$$

Where n indicates the number of SC_BSs, *ST(n)* is the simultaneous transmissions in MC, *Tr* is the transmission rate, and *N(n)* is the average hop number of backhaul data traffic in MC.

The MAN_BS, MC_BS, and SC_BS are equipped with massive MIMO-OFDMA antennas. The control system of base stations has to manage a special module known as frequency module. The frequency module of MC_BS keeps the record of all connected SC-BSs and the frequency module of an SC-BS keeps the record of all the RATs of all connected mobile devices. SC-BSs update their MC_BS and all the MC_BSs of a MAN update their serving MAN_BS. The MAN-BS maintains the database of the whole network and updates frequently. As all the base stations are equipped with massive MIMO antennas, therefore, MC_BS can connect maximum SC_BSs similarly the MAN_BS also can connect and control multiple MC_BSs. In order to control the critical situation, the MC_BS also has the ability to provide direct links to mobile devices. Finally, the MAN_BS connects with the core network.

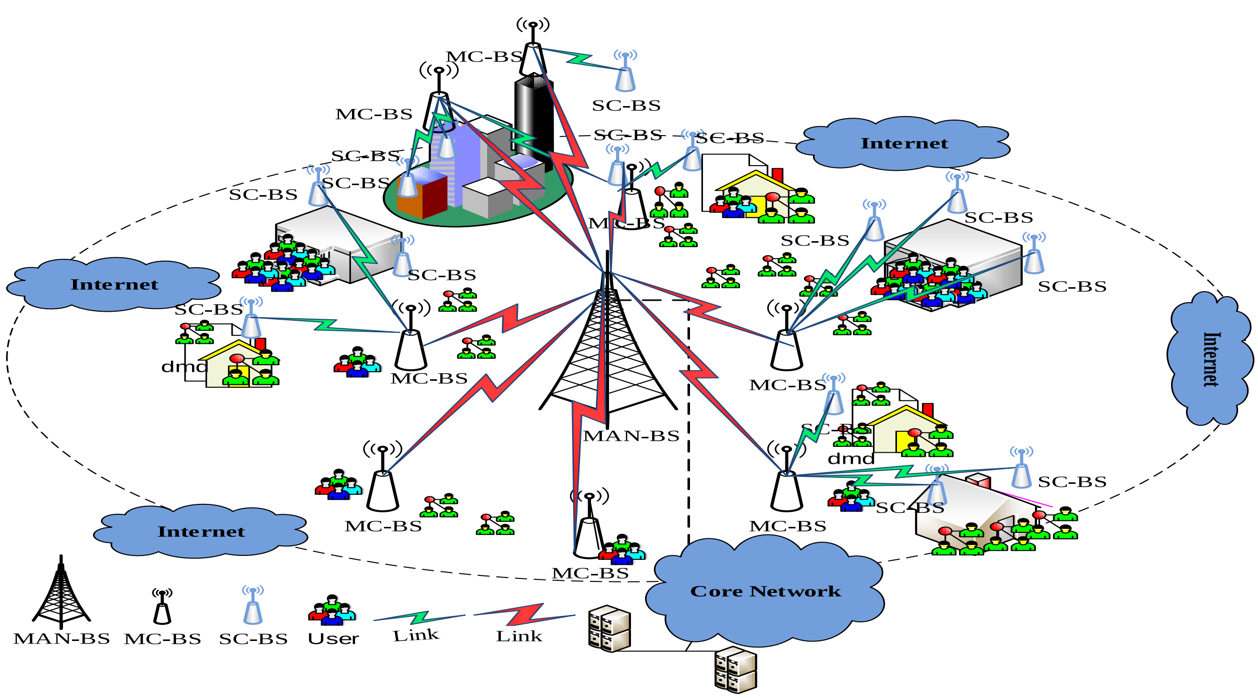


Fig. 2. General Layout of the Proposed 5G ultra-Dense Heterogeneous Network

A. Massive Multi-Input Multi-Output with Orthogonal Frequency Division Multiple Access (Massive MIMO-OFDMA) Technology

In the proposed architecture, the BSs are equipped with Massive MIMO with OFDMA antennas as shown in Fig 3. The antennas of MAN_BS, MC_BS and SC_BS are using spatial multiplexing with Orthogonal Frequency Division Multiple Access (OFDMA) modulations. The spatial multiplexing transmits multiple streams of data on the same downlink resource block. These streams have the ability to link one or more wireless devices. All receiver antennas can accept the streams from all the transmit antennas. Spatial Multiplexing Matrix CMM is given in equation 5.

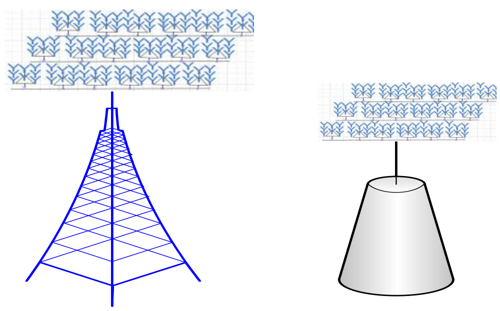


Fig. 3. Base Stations Having Massive MIMO Antennas.

$$CMM = \begin{array}{c} \xrightarrow{Tn} \\ \begin{matrix} C11 & C12 & \dots & C1n \\ C21 & C22 & \dots & C2n \\ \vdots & \vdots & & \vdots \\ Cn1 & Cn2 & \dots & Cnn \end{matrix} \end{array} \downarrow Rn \quad (5)$$

Where *Tn* and *Rn* are used to denote the numbers of transmitting antennas and numbers of receiving antennas. The coefficients C_{ij} of transmitting antenna *i* and receive antenna *j* shows all the possible routes between the receiver and transmitter.

The spatial multiplexing technique formulates the receiver more complex, therefore it is combined with OFDMA, where the problems caused by the multipath channel are handled efficiently. In OFDMA, the frequencies are orthogonal to each other; therefore, it fully controls the interference and prevents bandwidth reduction. In order to extend the spatial dimension, the OFDMA uses a frequency dimension. Consider a MIMO OFDMA base station has *N-Tr* transmit and N-Re receiver antennas (Tr-Transmit signal and Re-Receive signal). For incoming data, the *N-Tr* antenna encodes the channel, maps the encoded incoming data bit streams on spatial dimension, and maps the coded data bits. The *N-Re* antenna performs the detection operation for encoded data streams. Fig 4, is the general structure of the MIMO structure. The massive MIMO-OFDMA transmitter is depicted in Fig 5, a single dimensional encoder encodes the incoming data streams. Following by, the encoded streams are mapped onto the three dimensions by using Space Time Frequency Mapping (STFM) that makes the transmitter a complete OFDMA transmitter. The transmitter performs the interleaving, Quadrature Amplitude Modulation (QAM) mapping and cyclic extension process to the incoming data streams. Finally, DAC forwards the signal in the analog domain where it changes the radio frequency and then transmits.

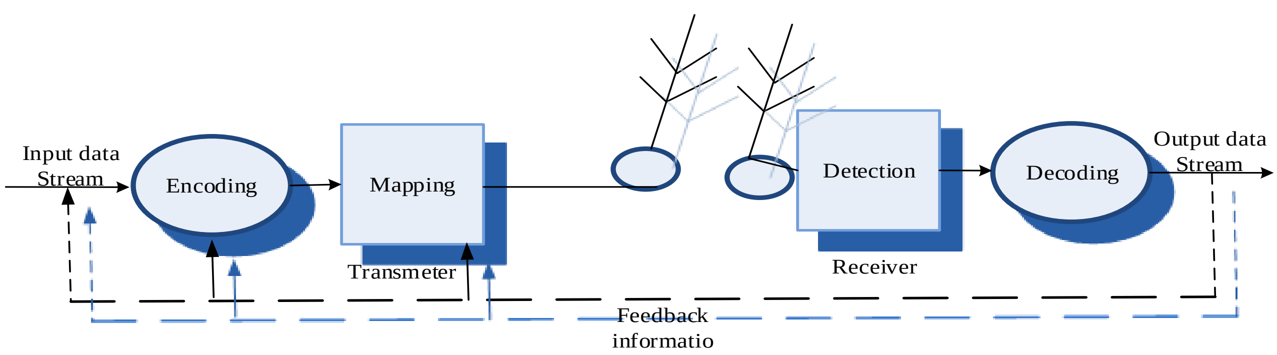


Fig. 4. The General Structure of MIMO Structure.

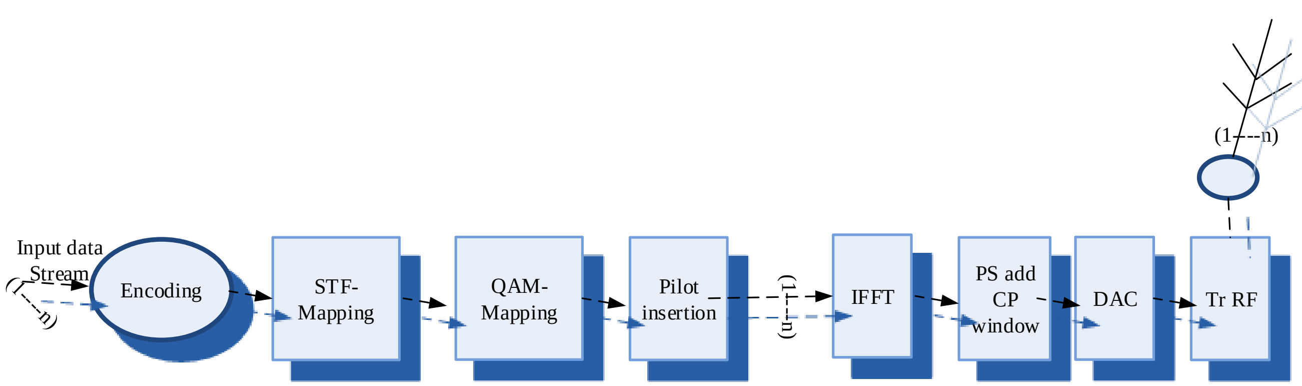


Fig. 5. Massive MIMO-OFDMA Transmitter.

The massive MIMO-OFDMA receiver is used to represent the digital representation of receiving signals N-Re. In a next step, the receiver arranges the frequency and gets the symbol timing according to the sequence. All the receiver branches execute the synchronization process together. For an efficient frequency synchronization of multiple branches, it helps to keep all the branches on one side of a link connected to the same local oscillator in a heterodyne scheme. Once the synchronization process completed, it eliminates the guard time known as Cyclic Extension CE (equal to the last part of the OFDM symbol) that eradicates the channel interference. Subsequently, it performs a Fast Fourier Transform (FFT) for each receiver branch. The Fourier analysis changes a signal from its original domain (space or time) in the frequency domain. In order to recover, data streams the STF detection and decoding is executed. In order to track the phase shift, pilot symbols are included in every MIMO-OFDMA data symbol on the predefined subcarrier. Finally, the data streams are combined and STF de-mapping and decoding are executed on transmitted data streams and get the output data stream.

B. *IP-based Interoperability in Ultra-Dense Heterogeneous Wireless Network*

When a device connects to the new RAT, first it has to disconnect the current link and establish a new link with the new RAT and IP address. This is time-consuming and inflexible way therefore to overcome this deficiency a new handover technique is proposed. The proposed 5G heterogeneous wireless network has the ability to control multiple RATs at the same time. All the available radio technologies are responsible for handling the wireless device mobility and handover while the user of the terminal device makes the final selection among different RATs. Therefore, all the wireless devices must contain an interface to switch between different RATS. The proposed architecture considers the vertical handover, where wireless devices can perform switching between available RATs.

All the base stations have a built-in control system with some special modules, one of them is Frequency Module FM. This FM is responsible for allocating frequencies to all other connected wireless devices and keeping the updated record. All the wireless devices (connected to the network) must have an interface for using different RATs so the devices can easily select the desired RAT and connect. The FM of SC_BS updates the MC_BS about the frequencies of all connected wireless devices and FM of MC_BS further updates the MAN_BS. If a wireless device wants to switch from one RAT to the other than the control system of a base station manages the handover links. The wireless device chooses the new RAT as displayed in its interface and the wireless device transmits an auto message to the base station where the control system establishes the new link with the desired RAT and then disconnect the previous RAT link. In this case, the interoperability between RATs is done on the network protocol layer that is commonly used by all the RATs. As shown in Fig 6, an IP based 5G wireless mobile network. The IP technology is considered to make sure enough control data for the efficient routing of IP packets between wireless devices and servers on the internet.

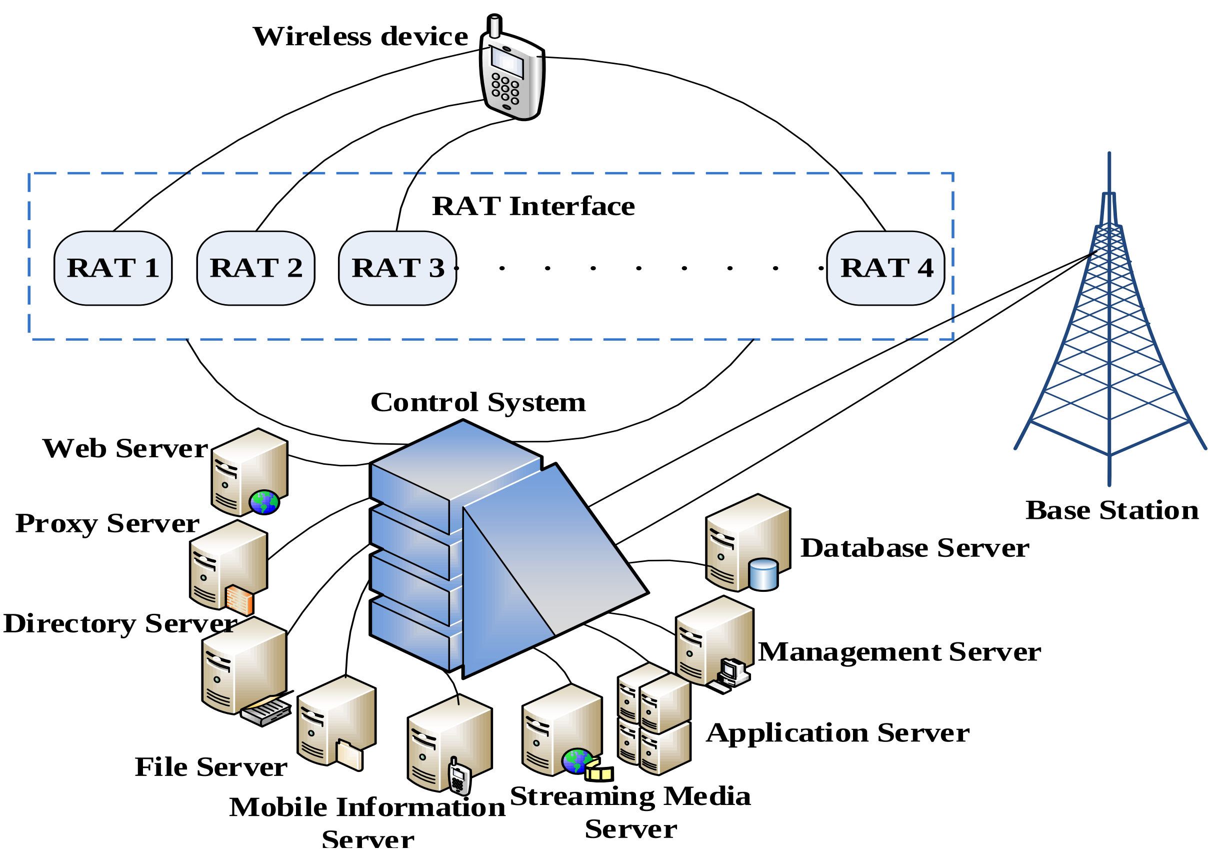


Fig. 6. An IP based 5G wireless Mobile Network.

C. Functionalities of the Wireless Terminal Device

There are three stages (RAT discovery, RAT selection, and RAT handover) of RAT interoperability. The wireless terminal device must be capable of detecting, selecting and activating the RAT links. The vertical handover process goes through the normal procedure within the specified time in a seamless manner. Like base stations, the wireless devices also have some functional modules, the functional modules are shown in Fig 7. The responsibility of the Database Administration Module (DAM) is to keep the record of connected base stations. A list of neighboring base stations, available RATs, RAT's policies, updated status of RATs, periodically monitored Received Signal Strength (RSS) from the base station, device mobility, and its status. The Network Interface Module (NIM) is responsible for identifying the available RATs on the wireless terminal device and performing switching between different RATs. It monitors the status of each RAT and updates the status during handover. The statuses of all RATs are updated by using a scan, idle, active, turn off and sleep modes. The Mobility Administration Module (MAM) is responsible for handling the mobility and connections of wireless terminal devices. In case a wireless device moves out of the radio range of one small cell to the other or even moves out from one microcell to the other or from one macro area network to the other. The MAM efficiently handle the mobility events and immediately updates the status. Finally, the DAM status is updated.

A fully network connected terminal device is considered. When a terminal device needs to switch from its current RAT, it opens the RAT interface and selects the desired RAT. It updates the DAM to change the mode of the selected RAT. After selecting the Target RAT, the DAM calculates the RSS from its SC-base station and send a link request to the base station. The RSS is calculated to make sure either the terminal device is in the radio range or not. Each RAT has a pre-defined threshold (*th*) value for RSS. Whenever a wireless device has to switch its RAT first it calculates the RSS which automatically updates in DAM. If the condition is true, then a terminal device performs the handover. The RSS is calculated by using the following equation.

$$RSS_{Target\ RAT} \geq RSS_{th,Target\ RAT} \qquad (6)$$

Where *RSS* Target *RAT* is the *RSS* that the terminal device receives from the SC_BS and the $RSS_{th,\ \mathrm{Target\ RAT}}$ is the RSS threshold value of the Target *RAT*.

The SC_BS sends an acknowledgment (*Ack*) to the terminal device that shows the connection has been established. After receiving the *Ack* the terminal device connects with T_RAT and changes the mode. Subsequently, it disconnects the previous RAT and updates its current mode. Thus, the terminal device connects to a new RAT by performing a successful IP based vertical handover, therefore it updates its CS_BS by sending an ACK. Finally, an update is sent to the MC_BS. The wireless terminal device IP based vertical handover procedure is explained in Fig 8. Traditionally, wireless devices linked with a base station may decrease the performance as a mobile device moves away from its base station. In the proposed 5G network architecture, the DAM of end devices monitors the states and update the database. Therefore, a device having a weak signal strength automatically connects to the neighboring associated base station.

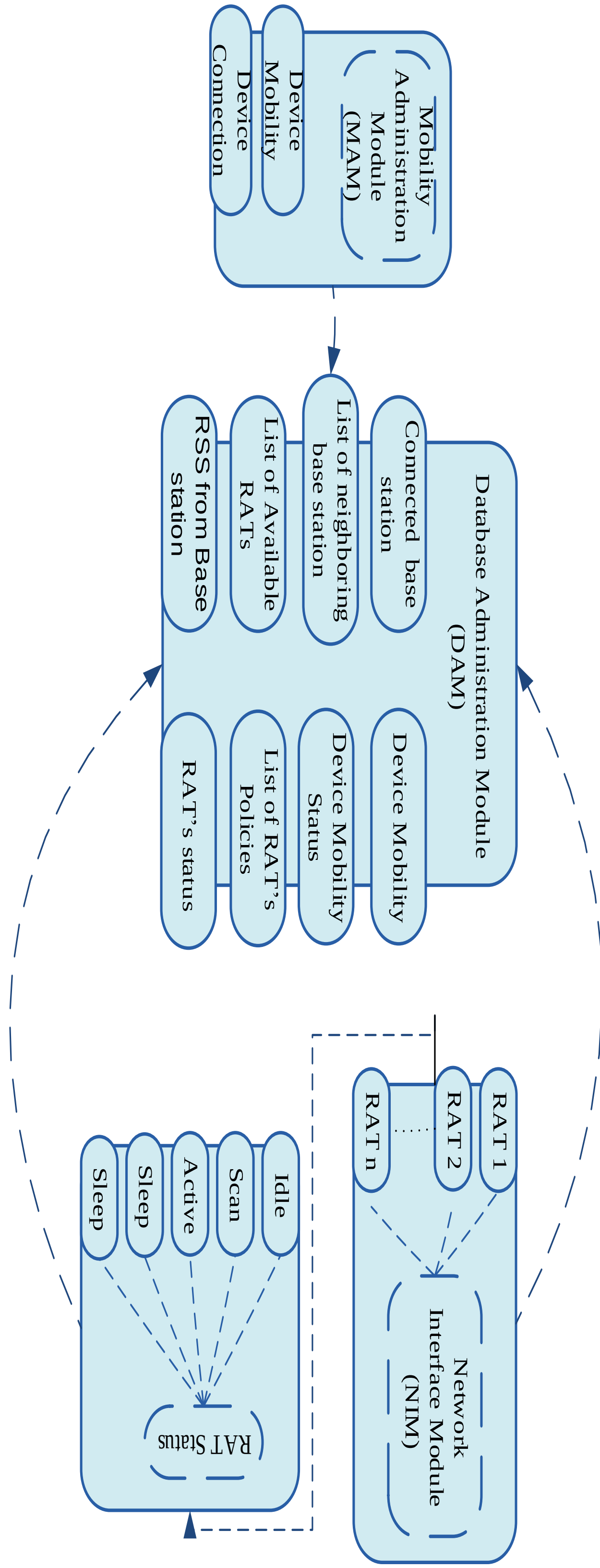


Fig. 7. Functional Modules of wireless Terminal Device.

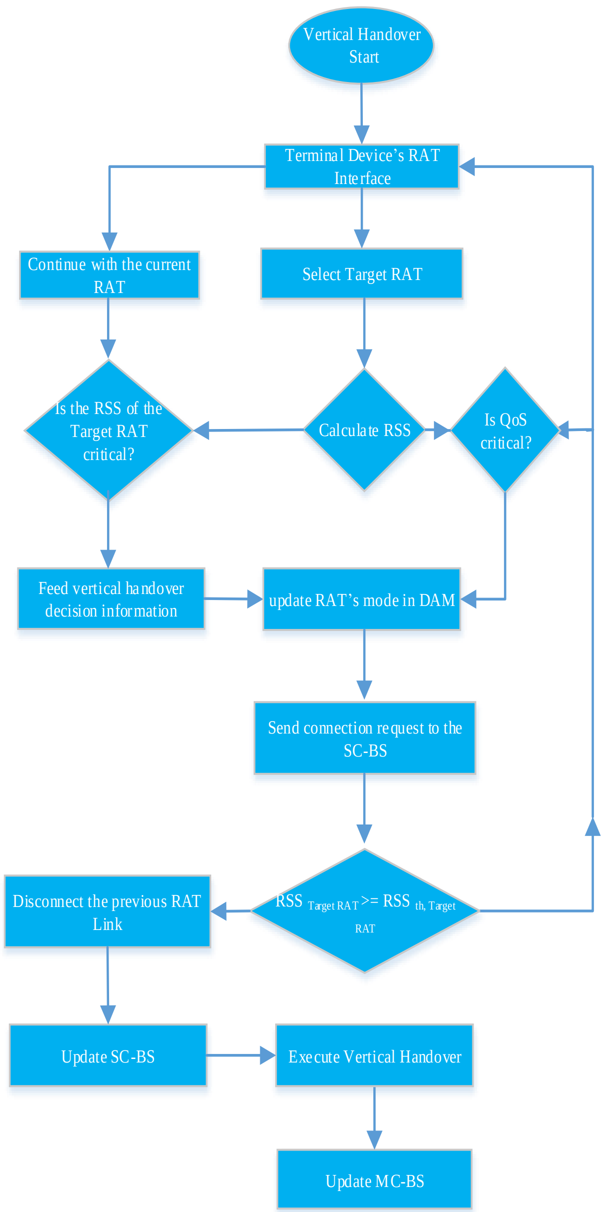


Fig. 8. IP based Vertical Handover of a Terminal Device.

IV. Results and Discussions

This section evaluates the performance of the proposed architecture and assesses how the network coverage and energy efficiency changed with the macro, micro, and small cells' distribution. The cells' deployment is independent and follows the Uniform and Poisson random measure. The performance of the proposed architecture is evaluated in the OPNET simulator. The simulation area is set to 300 × 300 m2 all the base stations are located in the center of the cell whereas the user equipment is randomly located around the base stations. The simulation parameters are listed in table 1. The radius of base stations SC_BS is set as the (100 m, 200 m), MC_BS radius (300 m, 500 m) and MAN_BS is 1 km.

TABLE I. Simulation Parameters

Parameters	Values
Area	300 x 300 m2
SC_BS radius	100 m, 200 m
MC_BS radius	300 m, 500 m
MAN_BS	1 km
Traffic mode	CBR
Per channel bandwidth	100 MHz
Bandwidth 2	20 MHz
Scheduling Algorithm	Proportional fair
Number of MAN-BS	3
Number of MC-BS	9
Number of SC-BS	55

The network energy efficiency is a key constraint that limits the 5G network's densification. The energy consumed by the SC-BS is consist of Embodied Energy (EE) and Operational Energy (OE). The EE is the power utilized by all the process related to the BS implementation. Whereas the OE is the energy utilized by the operations of backhaul within the network lifetime Time To Live (TTL), which is expressed as:

$$OE = BSOP * TTL \tag{7}$$

Where BSOP is the BS operational power and it is considered as a linear function of SC_BS. It is defined as $BSOP = A * P_{Tr} + B$, where P_{Tr} is transmission Power, $A = 7.90\ W$ and $B = 72\ W$. In order to clarify the derivation model, P_{Tr} of a BS is normalized as $P_{Nr} = 1.5$ W and BS backhaul throughput θ is considered as 1 Gb/s. The average throughput A_{Th} is calculated by:

$$P_{Tr} = P_{Nr} \times (A_{Tr} / \theta) \tag{8}$$

The operating power of BS has calculated as below:

$$BS_{op} = A \times P_{Nr} \times (A_{Tr}/\theta) + B \tag{9}$$

The Backhaul Energy Efficiency B_EE is obtained through:

$$B_{EE} = \frac{B\_NC}{n \times (BS\_EC)} \tag{10}$$

Where *B_NC* is Backhaul Network Capacity and *BS_EC* is Backhaul Energy Consumption.

The energy efficiency is analyzed in Fig 9 and 10 with Uniform and Poisson distribution. At the start, the energy efficiency was improving with the increasing number of small cells then it starts decreases after it reaches the threshold value. With the static number of MAN_BSs, the energy efficiency improves with the increase of radius. With the fixed radius, backhaul energy efficiency with respect to the average of SC_BSs' throughput is explained in figures.

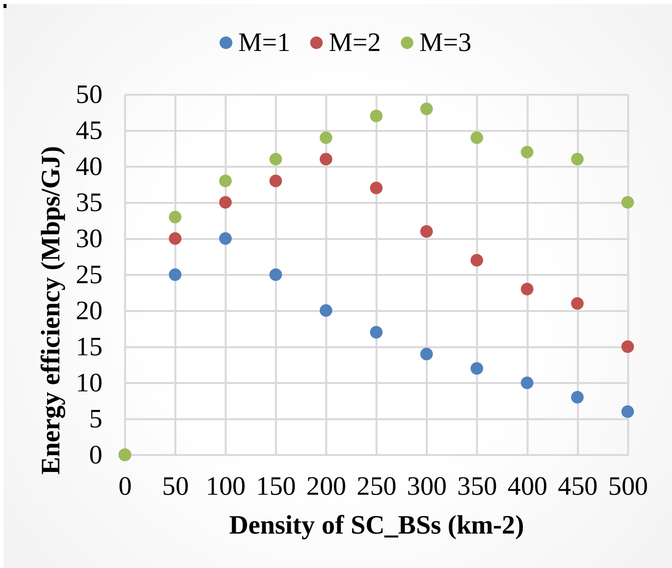


Fig. 9. Energy Efficiency Vs. Density of SC_BSs with Uniform Distribution.

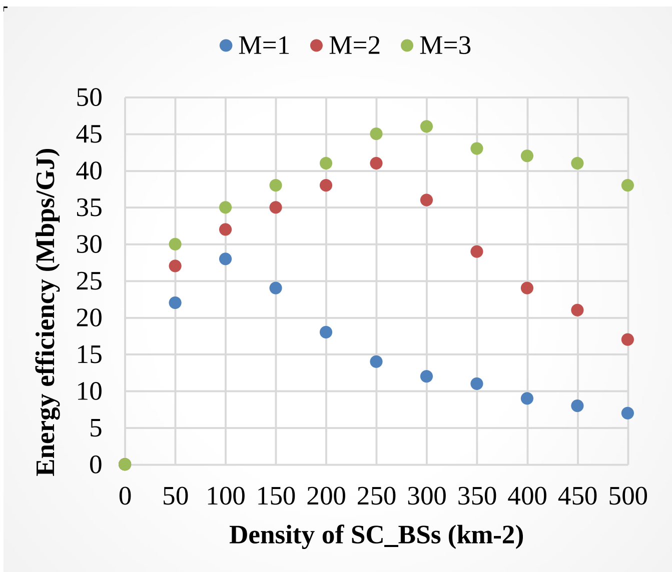


Fig. 10. Energy Efficiency Vs. Density of SC_BSs with a Poisson Distribution.

As a microcell is an ultra-dense area, therefore, to get the full network coverage proposed technique to divide it into the SCs. All the active MTs in a single cell connect to the respective BS where BSs are equipped with the MIMO antennas. With the fixed SC radius, the B_NC with respect to the number of SCs is explained in Fig 11, it was observed that the B_NC increases with the increase of a number of small cells. When B_NC touched its maximum limit, it gradually decreases with the increase of SC. In Fig. 12, the average number of simultaneous transmission is found as 19, 25, and 29 when the small cell radius is set as 300 m, 200 m, and 100 m correspondingly. The backhaul network capacity increases with the increase of simultaneous transmissions. When the average number of simultaneous transmission and the number of SC is fixed the backhaul network capacity start decreasing.

From the simulation results, it has observed that when the average simultaneous transmissions are larger than the predefined threshold value then backhaul capacity get the constant value. This simulation output helps to design the ultra-dense cellular network.

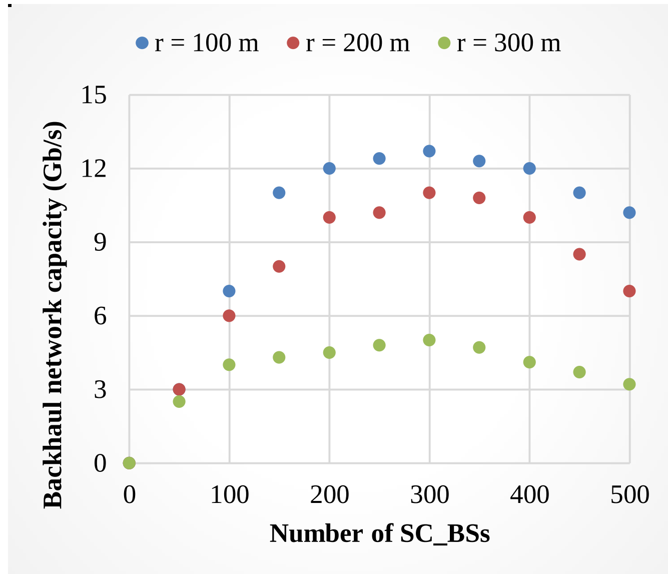


Fig. 11. Number of SC_BSs vs. Backhaul Network Capacity.

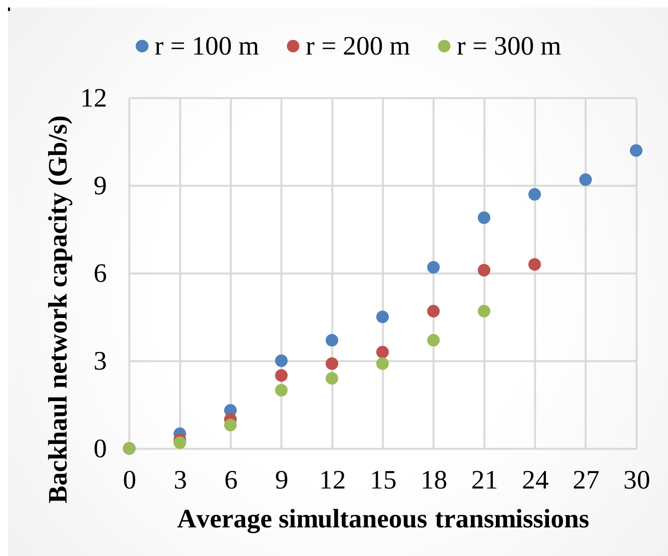


Fig. 12. Average Simultaneous Transmissions Vs. Backhaul Network Capacity.

In the proposed architecture, every BS is equipped with an FM, which has to allocate the frequencies to the other connected MTs. An MT can easily select the desired RAT from its interface. An FM of SC_BS periodically informs the respective MC_BS about the frequencies of all the connected MTs and it further updates the MAN_BS. If a MT needs to switch from its C_RAT first it selects a T_RAT from the available RATs as displayed in the interface. The MT sends an auto message to the BS where the control system manages a new link with the T_RAT and disconnect the C_RAT links. In Fig. 13, the backhaul network capacity with respect to the number of SC_BSs is shown, MC-BS starts increasing with the increase of SC_BSs and with the increase of density of a small cell. When the density of an SC is equal to or greater than the defined threshold value then the capacity reaches a static value. The backhaul network capacity is analyzed with the Uniform and Passion distribution. When the backhaul network capacity with the Uniform distribution of SC_ BSs attains the static values as in Fig 14, the density thresholds of a small number of MAN_BSs (M) = 3, 2 and 1, respectively. When the backhaul network capacity with the Poisson distribution of small cell BSs achieves the static values in Fig.14, the density thresholds of small cell BSs are 0-500 which relates to the number of MAN_BSs (M) = 3, 2 and 1.

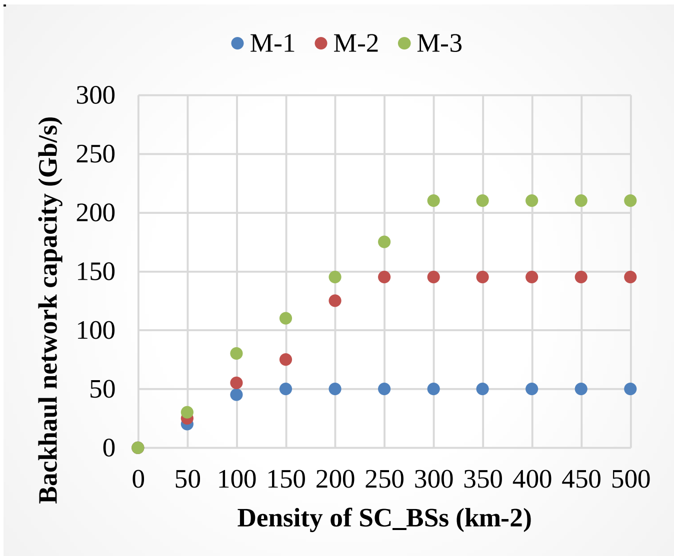


Fig. 13. Backhaul Network Capacity vs. Density of SC_BSs with Uniform Distribution.

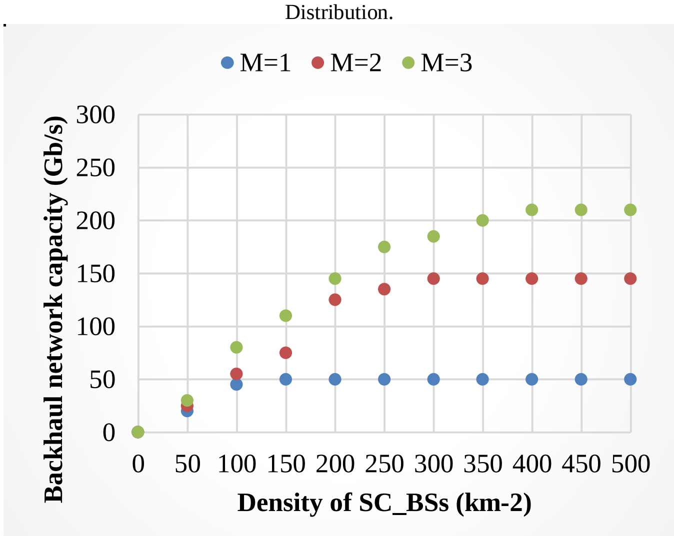


Fig. 14. Backhaul Network Capacity vs. Density of SC_BSs with a Poisson Distribution.

V. Future Challenges

The increasing use of the ultra-dense cellular network is motivated by the massive MIMO-OFDMA and IP based vertical handover, the distributed network architecture is a reasonable solution. In spite of 5G network benefits, the small cell densification opens up new research directions. There are some potential challenges of scheduling, routing relay optimization, interference, and allocation of best massive MIMO antennas etc. Scheduling is considered an efficient technique to utilize the available frequency. On the other side, the proportional fair scheduler can also be used to improve the throughput with a large number of wireless mobile devices. A 5G distributed wireless network needs to carefully select a routing path between the source and destination devices. If base stations are equipped with the complex massive MIMO antennas, then there is highly required reasonable allocation of massive MIMO antennas for achieving efficient transmission.

A 5G wireless network will handle the huge amount of data traffic of the ultra-dense wireless network but interference can be a critical challenge. Therefore, the system needs more attention to implementing the interference avoidance techniques. The use of high frequency, the highest data rate, minimum delay, quick response time, complex antennas, implementation of multiple RATs in a single wireless device pays attention to the development of advanced base stations and wireless devices. The implementation of the 5G network is not limited to the hardware, but it highly requires the implantation of new protocols and algorithms.

VI. Conclusion

Ultra-dense wireless networks are still considered as a complement to the cellular wireless network with an efficient architecture. The massive MIMO antennas facilitate the 5G ultra-dense cellular networks to be scattered in all wireless cellular set of connections. In this paper, a novel 5G wireless network architecture is presented with two main aspects of massive MIMO-OFDMA and IP based vertical handover. The proposed architecture divides the whole macro area network into the microcells and each micro-cell consists of multiple small cells. All the cells are manipulated by their adjacent massive MIMO-OFDMA base stations. The heterogeneity and small-sized cell base stations provide full network coverage. The proposed architecture performs switching between different RATs. All the devices are designed with a multiple RAT interface from where the user can select the desired RAT and perform vertical handover. In order to illuminate the handover delay and improve the service continuity, the vertical handover is IP based. The advantage of using IP based vertical handover is establishing new connections with the same IP address before disconnecting the previous links.

References

[1] A. Zakrzewska, S. Ruepp, and M. Berger., “Towards converged 5G mobile networks - Challenges and current trends,” in Proc. ITU Kaleidoscope Academic Conference. 2014. p. 39-45,

[2] S. Talwar, et al., “Enabling technologies and architectures for 5G wireless,” in IEEE MTT-S International Microwave Symposium (IMS),. 2014. p. 1-4,

[3] ZoraidaFrias and JorgePérez Martínez, 5G networks: Will technology and policy collide?. Telecommunications Policy, Elsevier, 2017.

[4] J. Bernardos, et al., An architecture for software-defined wireless networking. Wireless Commun. IEEE., 2014. 3(21): p. 52-61.

[5] E. Piri et all, 5GTN: A test network for 5G application development and testing, European Conference on Networks and Communications (EuCNC), Athens, p. 313-318.

[6] J. P Romero, et al., “Radio Resource Management Strategies in UMTS”, 2005, John Wiley & Sons, New York, USA.

[7] A. Ejaz et al., "Recent Advances and Challenges in Mobile Big Data", IEEE Communications Magazine 2018. 56(2): p. 102-108.

[8] Vinicius C. M. Borges, et al., "Aspirations, challenges, and open issues for software-based 5G networks in extremely dense and heterogeneous scenarios", EURASIP Journal onWireless Communications and Networking, 2015. 164.

[9] S. T. Bakhsh, et al., Cross-layer-based Adaptive Traffic Control Protocol for Bluetooth Wireless Networks, International Journal of Advanced Computer Science & Applications, p. 102-108, vol 8, June 2017.

[10] S Hong, et al., Applications of Self-interference Cancellation in 5G and Beyond, in Commun. Mag. IEEE 2014. p. 114–121.

[11] WH Chin, Z Fan, and R Haines, "Emerging technologies and research challenges for 5G wireless networks", Wireless Commun. IEEE, 2014. 2(21): p. 106–112.

[12] WH Chin, Z Fan, and R Haines, Emerging technologies and research challenges for 5G wireless networks. Wireless Commun. IEEE., 2014. 2(12): p. 106–112.

[13] C-X Wang, et al., Cellular architecture and key technologies for 5G wireless communication Networks, in Commun. Mag. IEEE.. 2014. p. 122-130.

[14] E Björnson, et al., Multiobjective signal processing optimization: the way to balance conflicting metrics in 5G systems, in IEEE Signal Processing Magazine. 2014. p. 14-23.

[15] Noha Hassan and Xavier Fernando, Massive MIMOWireless Networks: An Overview, Electronics, 2017: p. 29.

[16] Stefano Ruffini, et al., A Novel SDN-Based Architecture to Provide Synchronization as a Service in 5G Scenarios, in IEEE Communications Magazine, 2017.

[17] T Bogale, et al., "User scheduling for massive MIMO OFDMA systems with hybrid analog-digital beamforming", in IEEE ICC. 2015.

[18] W Liu, A Zappone, and C Yang, "Global EE optimization of massive MIMO systems", in IEEE SPAWC. 2015.

[19] Qian Wang, et al., 5G MIMO Conformal Microstrip Antenna Design, Wireless Communications and Mobile Computing, 2017. 2017: p. 11.

[20] M., N.H., et al., Massive MIMO for Fifth Generation (5G): Opportunities and Challenges, in International Conference on Computer and Communication Engineering (ICCCE),. 2016: Kuala Lumpur,

[21] G. Xiaohu, et al., "5G Ultra-Dense Cellular Networks". IEEE Wireless Communications, 2015.

[22] M. Fadi et al., "Energy Efficiency Perspectives of Femtocells in Internet of Things: Recent Advances and Challenges", IEEE Access, 2017, p. 26808-26818.

[23] Osseiran. A., et al., "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," in Communications Magazine, IEEE,. 2014. p. 26-35.

[24] C-X Wang, et al., Cellular architecture and key technologies for 5G. wireless communication Networks. Commun. Mag,, 2014. 52(2): p. 122–130.

[25] Lu. Yang and Wei. Zhang, Interference Coordination for 5G Cellular Networks. SpringerBriefs in Electrical and Computer Engineering, Springer, 2015.

[26] G. Mao, et al., "Towards a Simple Relationship to Estimate the Capacity of Static and Mobile Wireless Networks", IEEE Trans. Wireless Commun. Aug. 2013. 12(8): p. 3883-3895.

[27] I. Muhammad, et al., reconfigurable scatternet formation and maintenance scheme with heterogeneous services for smart Bluetooth devices, Sustainable Cities and Society, vol 36, p. 589-599, 2018.

Audio Augmentation for Traffic Signs: A Case Study of Pakistani Traffic Signs

Abdul Wahab[1], Aurangzeb Khan[2], Ihsan Rabbi[3], Khairullah Khan[4], Nasir Gul[5]
Department of Computer Science
University of Science and Technology Bannu
Bannu, Pakistan

***Abstract*—Augmented Reality (AR) extend the appearance of real-world by adding digital information to the scene using computer graphics and image processing techniques. Various approaches have been used to detect, identify and track objects in real environment depending upon the application, shape of the tracking object and environment type. The marker-based tracking technique is the most commonly used method in augmented reality applications in which fiducial markers are put in the real world for tracking. In this work we proposed a model to detect and identify the traffic signs through marker based technique to improve the usability of marker-based detection in augmented reality applications. We developed an AR application that can detect and recognize the markers designed for Pakistani traffic signs and augment them with voice alert to the driver so that he does prepared for the upcoming hurdle on the road. As identified by literature no work has been performed on augmentation of voice for traffic signs. From experiments the model outperforms than baseline techniques.**

Keywords—Augmented reality; traffic sign detection; traffic sign recognition

I. Introduction

Road safety is becoming increasingly critical day by day world over, particularly in developing countries like Pakistan with the growing rate of ownership of vehicles. In fact, a large number of individuals lose their lives every year due to road accidents and the yearly rate of such individuals seems to be increasing. According to statistics, every year 1.25 million individuals lose their lives in road accidents. A huge number of others sustain wounds, and some experience long lasting disabilities. No nation is saved this toll in lives and suffering that especially influences youngsters. Enormous human potential is crushed, with severe social and financial effects. Road safety is therefore a noteworthy general medical issue around the world[1].

Advanced Intelligent Driver Assistant System (AIDAS) is a critical part of the intelligent transport system which surely warns potential hazards on the road, helps the driver in routing and direction. Traffic signs are landmarks that direct the driver on hazardous road conditions and therefore enhance road safety[2]. Generally, the traffic signs have distinctive color, shape, size and characters or symbols. Most commonly triangular, circular, rectangular and octagonal shapes are used to symbolize warning, prohibitions, recommendations and stop signs respectively[3].

The traffic signs are put beside the road so that they can be seen by the drivers easily. Inability to see traffic signs due to various conditions can bring about lethal car accidents. Currently, traffic risks are increasing due to the increase in number of vehicles. The adoption of an automated system for detecting and recognizing traffic signs thus improves road safety.

In general, global positioning system (GPS), radio frequency transmitter and vision-based techniques are used to improve driver alertness[4]. The first two methods have weaknesses such as failure in GPS connection and installation of RF transmitter at every single traffic sign is very difficult and costly. As a result, the vision based technique is more reliable and inexpensive. From two decades, traffic sign recognition (TSR)has become very attractive research topic among many computer vision communities due to its more practical applications[2].

In this paper we present a model to detect and recognize the traffic signs through marker-based technique. When the car's camera sees the marker, it will generate a sound, corresponding to that traffic sign which will alert the driver.

The rest of the paper is ordered as follows: literature review is described in section-II. Section-III elaborates the proposed model. Section-IV provides the results and discussion of detection and recognition of markers in outdoor environment. Finally, the conclusion is given in section-V.

II. Literature Review

Road signs are defined with a particular shape and color that makes them easy to spot before other common object. In general, three primary approaches are used as a part of traffic sign recognition i.e. color-based, shape-based and neural network based. Due to unique color of traffic signs, the color-based technique has turned out to be exceptionally prevalent.

A. *Color Segmentation*

Color is a predominant visual component and without a doubt represents essential data that the driver must oversee. The most specific colors used for traffic signs are red, blue, yellow, black and white. Color-based detection technique is considered to be the most widely used strategy for initial search space reduction.

The RGB color space is followed by [5]. In light of their investigation they worked just on 3 colors: red, blue and yellow to make 3 binary images containing just those pixels only that belong to these colors. All associated pixels to the same connected area are marked together. Beneath a predefined threshold regions are simply rejected in light of the fact that they are generally not traffic signs or otherwise, the target is such that detection due to its smaller size could not be successfully classified. They accomplished good results for their classifications of selected signs. For example, signs of obligation and prohibition signs.

Numerous researchers believe that the RGB color space is extremely delicate with respect to lighting variations. These strategies are led by HSI or HSL color space. The direct threshold on (RGB) space is rarely used due its sensitivity to variations in lighting. The HSI color models are the most popular as it depends on the human color observation. In addition, it is generally considered invariant to changes in lighting and brightness [6-8]. HSI is used by many researchers such as [9-12]with the reason that the HSI space models is superior to RGB for human vision and permits some variety in the intensity of light.

However, some researchers are not happy with the efficiency of HSI as it does not exhibit any adjustment in color temperature under various climate conditions such sunny, cloudy, foggy or rainy, yet just changes the intensity of light. A group of researchers[13, 14] used CIECAM color space which can foresee color appearance as precisely as a normal observer. Compared with other color models, the data demonstrate that CIECAM surpasses the RGB and HSI color spaces for sunny, cloudy and rainy conditions with a correct fraction of 94%, 90% and 85% respectively. Variations in climate conditions, when there are various artificial lights, the color of traffic signs and lighting sources fluctuate, leading to the situation when the adoption of most techniques in color-based traffic sign segmentation may not work accurately all the time. In fact, till date no generally acknowledged technique is available in color-based segmentation of traffic signs[15].

B. Shape based Detection

Another basic strategy is to exploit shape information in a way got from edge properties. The primary advantage is its strength in connection to the distinctive lighting conditions and the deterioration of the signboard. A few authors favor a colorless approach since they don't consider color segmentation to be absolutely reliable because of their sensitivity to different factors, for example, changing weather situations, reflection of sign board surfaces, time, or extreme camera exposure. On the other hand, shape of the traffic sign is more powerful attribute with respect to the varying lighting conditions and the degradation of signs boards.

A corner detection technique with a specific color was applied by [16]to discover traffic signs. They just used a few points on the boundary of circle rather than every pixel. Subsequently, this technique is susceptible to occlusion issues. The same authors have accomplished the task of traffic sign identification by merging color information and shape information using boundaries[15].

A new approach of shape-based identification is the Haar cascade classifier which needs training a classifier for each traffic sign, prompting a tedious identification process that was tea large amount of the processor time as applied in[6].

Hough transformation and its derivatives dominate the detection phase of traffic signs as used by[17, 18].

The groups of researchers[19, 20]utilized the fast radial symmetry transformation to discover normal signs in the form of a polygon such as triangle, square and octagon. The general polygon detector algorithm is utilized in these attempts to detect rectangular speed signs in the Unified State of America. Barnes and Zelinsky[4] applied radial symmetry detector in their system that uses gradient image to recognize Australian circular speed signs in real time.

C. Classification of Signs

The meaning of the traffic sign lies inside the candidate in the form of symbols. To recognize the meaning of a traffic sign, the classifier generally require either an inside part of the target sign or some specific properties as its inputs, that recognize the candidate sign from other signs. Most general classifiers used as a part of traffic sign detection are neural networks[5, 21] and support vector machine[22, 23]. Previously, some work has merged the color with support vector machines (SVMs) for shape-based detection and in addition SVM-based classification[24].

D. Augmented Reality Applications to Navigation

A tourist application for an Android system called i-Street was introduced by[25], which intends to detect, recognize and read street plates in a video stream and after that measure the relative location to augment them precisely with virtual overlays. Another application developed by [26] augments surrounding geographic information such as street name, virtual roads and current location on the screen as users walk on a street to give them an overview of the surrounding environment. The Smart Multimedia Guide developed by [27] is a location-based application that manipulates various services and technologies. The system can detect the artwork closest to the user, and make these works able to automatically tell their stories using multimedia facilities while users visit.

A device for a finger called EyeRing allows the user to point to an object, take a picture, and listen to comments about what he has just focused on. EyeRing can also act as an assistant for navigation or translation and even help children learn to read [28].

Through the study of relative work done in recent years in traffic sign recognition domain we can say that this is a vast domain where a range of techniques available for classification of signs and still there are a lot of ongoing research. In the literature as identified, no work has been performed on augmentation of voice for traffic signs. To overcome this problem, this research work focuses on the development of an application that will guide driver about road signs through aural instructions.

E. Challenges in Traffic Sign Recognition

The recognition of the road sign at first look seems to be extremely basic and simple, since it is characterized by a few directives, such as color, shape, size and pictogram. However, in reality it is exceptionally difficult job as there are additional factors that influence the task and performance of traffic sign recognition. Automatic detection and recognition of a traffic sign may include same challenges as object identification in natural environments.

In spite of the fact that traffic signs are very much characterized through laws and intended to be seen easily, there are still a lot of difficulties for TSR systems. For example:

a) Traffic signs are similar inside or across the categories

b) Traffic signs may have blurred due to long exposure to sunlight or are grimy so they are not any more in their specified colour

c) The sign board shaft may have twisted and therefore, the sign is no longer symmetrical to the road.

d) Lighting conditions may make colour identification unreliable.

e) Low contrast of environment may make shape identification hard.

f) There may be changing climate conditions (e.g., snowy, sunny, sandy, foggy).

g) In urban regions different objects may look fundamentally the same as traffic signs.

h) The occurrence of other items (e.g., moving autos, plants, walkers, shop signs, and so forth.) can occlude the visibility of a sign.

i) Video or image gained from a moving auto frequently experiences motion blurs. So the detection and recognition of traffic signs from such scenes turn out to be very challenging.

III. The Proposed Model

This research work is mainly focuses is on developing an augmented reality application that can detect and recognizes the markers created for Pakistani traffic signs. When the car's camera see the marker it identify it and then generate a voice command associated with that particular traffic sign to the driver to be alert.

The design of the proposed model is given in fig. 1 while the algorithm of the proposed model is given below:

A. *The Marker Database*

For the recognition of traffic signs through marker, Pakistani traffic signs were collected. A total of 130 traffic sign images were collected for analysis. The images were then converted into black and white for creating markers. The markers were created in Adobe Photo Shop CS4. Images database were classified into four categories i.e. Warning, Mandatory, Prohibitory and Stop as shown in Table 1.

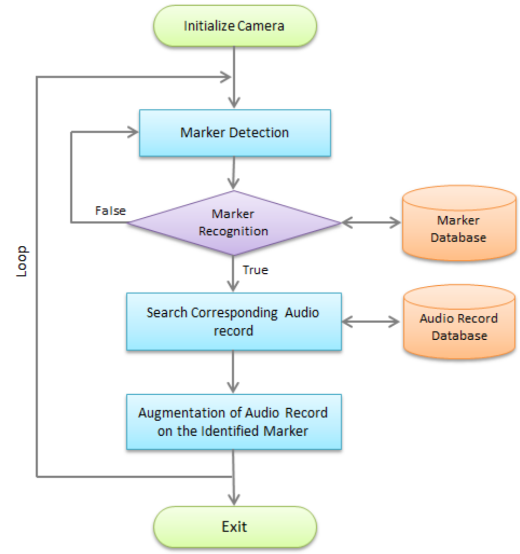
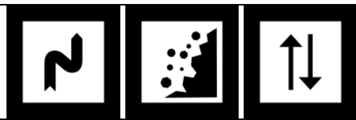
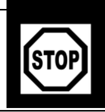


Fig. 1. Design of the Proposed Model.

Algorithm: Audio Augmentation for Traffic Signs

Input: Video frames from dashboard camera of the car
Output: Audio augmented scene of the road on the camera
Initialization(Camera, Variables)
while CameraRecording := True **do**
  Grab next frame from the video
  Search for potential marker in the frame
  if isMarkerFound := True **then**
    Identify the detected marker
    //Audio augmentation
    Search audio record in the audio database
    Augment (play) the corresponding audio on the identified marker
  else
    break
  Repeat Step 4
  end if
  end while
  exit (CameraRecording := False)

TABLE I. SAMPLE MARKER DATABASE

Category	Sample Markers	No. of Markers
Warning		47
Mandatory		40
Prohibitory		42
Stop		01

Marker-based methods utilize pattern files for identifying a specific marker. The pattern file is recorded in stable environment where the lighting and brightness level are constant. The marker tracking failure happens in uncontrolled lighting and brightness as well as varied contrast levels in an environment. To avoid this problem, multiple pattern files are recorded for a single marker of traffic sign. In this technique, single marker is recorded under different lighting conditions according to daytime such as in the morning, noon, afternoon and evening time. This strategy enhances the marker detection performance in conditions where lighting and brightness levels are changing.

B. *Effect of Lighting and Brightness on Marker*

Markers are affected constantly with the change of lighting and brightness conditions of environment which may lead to identification failure or false identification of markers. For the solution of this problem a technique of multiple pattern files at different timing of the day with different lighting and brightness levels of environment is proposed.

When the car's camera detects a marker on the road, these multiple pattern files will be checked for matching with the detected marker. This approach increases the rate of marker identification during different lighting, brightness and contrast levels in environment. Applying this technique may increase the processing time but since our research is only related to the detection and recognition of the markers rather than with continuously tracking the markers, so this processing time does not overload the system.

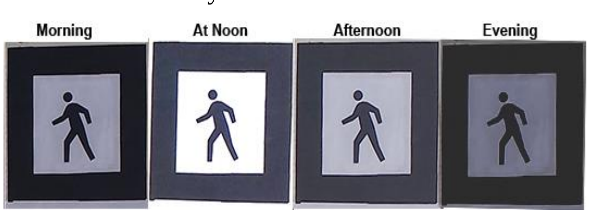


Fig. 2. Multiple Pattern Files for a Single Traffic Sign.

For accurate and fast identification of traffic sign markers, total of 4 pattern files were captured for a single marker at different lighting conditions depending on the day time i.e. morning, noon, afternoon and evening as shown in Figure 2.

C. *Experimental Setup for Checking the System*

Experiments were performed with a car prepared for the evaluation of the system. A webcam of 5.0 Mega pixels is used with a Laptop having Core i5 (quad core) 2.5GHz processor and 4 GB RAM. The webcam was attached with the rearview mirror of the car and connected to the laptop. The laptop was also attached to the speeker of the car through a 3.5mm cable for the voice alert. So that when the camera detect and recognize a marker then it can generate voice command to the driver through the car's speeker loudly. The experimental setup of the car is given in Fig.3.

Fig. 3. Car prepared for Experiments.

For experiment purpose the markers were place at different location on the road side. The experiments were conducted on different kinds of road such as rough and smooth, straight and having turns. A sample marker for STOP sign is place on the road as shown in Fig. 4.

Fig. 4. STOP Sign Marker place in the Road.

IV. SYSTEM EVALUATION AND RESULTS

To prove the effectiveness of the framework, the proposed system was tested on 20 markers. These 20 markers are selected manually from the database such that some of them are very similar to each other to check the accuracy/precision of the system. The resolution of the video is 640×480. Our main focus is only on detection and recognition of traffic signs when they appear in front of camera. We do not consider tracking the signs continuously.

A. *Detection Rate Results*

In this section the quantitative data is presented for the experiments performed on the proposed system. The experiments were conducted at different timing of the day i.e.

at morning, noon, afternoon and evening and with different car speeds i.e. 0-60 km/h, 65-80 km/h and 90-110 km/h. Similarly in the second part the recognition results of the system are given in the similar way.

1) Detection Rate at Morning

The experiment conducted at morning gave good results at all speeds i.e. 100%, 96%, and 86% for speed limits 0-60km/h, 65-80km/h and 90-110km/h respectively as shown in Table II.

TABLE II. DETECTION RATE IN THE MORNING

Speed	No. of Visits (A)	Total no. of Markers × (A)	No. of Detected Markers	No. of Missed Marker	Detection Rate (% age)
0-60 km/h	05	100	100	00	100
65-80 km/h	05	100	96	04	96
90-110 km/h	03	60	52	08	86

2) Detection Rate at Noon

The experiment conducted at 12 O'clock gave us low detection rate comparatively to the morning results due to high brightness in environment. The results are shown in table III.

TABLE III. DETECTION RATE IN THE EVENING

Speed	No. of Visits (A)	Total no. of Markers × (A)	No. of Detected Markers	No. of Missed Marker	Detection Rate (% age)
0-60 km/h	05	100	87	13	87
65-80 km/h	05	100	81	19	81
90-110 km/h	03	60	42	18	70

3) Detection Rate in Afternoon

The afternoon experiments gave similar results as in the morning. The results are shown in table IV.

TABLE IV. DETECTION RATE IN AFTERNOON

Speed	No. of Visits (A)	Total no. of Markers × (A)	No. of Detected Markers	No. of Missed Marker	Detection Rate (% age)
0-60 km/h	05	100	99	01	99
65-80 km/h	05	100	95	05	95
90-110 km/h	03	60	50	10	83

4) Detection Rate in the Evening

TABLE V. DETECTION RATE IN THE EVENING

Speed	No. of Visits (A)	Total no. of Markers × (A)	No. of Detected Markers	No. of Missed Marker	Detection Rate (% age)
0-60 km/h	05	100	93	07	93
65-80 km/h	05	100	90	10	90
90-110 km/h	03	60	48	12	80

The experiment conducted in the evening gave us good result but low than morning and afternoon as shown in table V.

The overall results of the percentage detection rate recorded at morning, noon, afternoon and evening, with different car speeds i.e. 0-60 km/h, 65-80 km/h and 90-110 km/h are shown in figure 5.

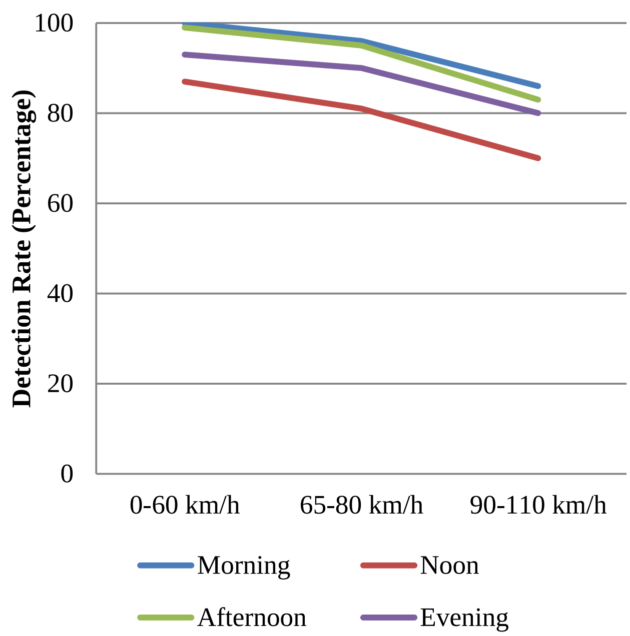


Fig. 5. Detection rate results graph.

B. Recognition Rate Results

In this part the results obtained for the recognition of markers from the experiments are elaborated.

1) Recognition Rate at Morning

The experiment conducted at morning gave good recognition results at all speeds i.e. 100%, 96%, and 86% for speed limits 0-60km/h, 65-80km/h and 90-110km/h respectively as shown in table VI.

TABLE VI. RECOGNITION RATE IN MORNING

Speed	No. of Visits (A)	Total no. of Markers × (A)	No. of True Positive	No. of False Recognitio n	Recognitio n Rate (% age)
0-60 km/h	05	100	98	2	98
65-80 km/h	05	100	93	3	93
90-110 km/h	03	60	48	4	80

2) Recognition Rate at Noon

The experiment conducted at noon gave us low recognition rate comparatively to the morning results due to high brightness level. The results are shown in table VII.

TABLE VII. RECOGNITION RATE AT NOON

Speed	No. of Visits (A)	Total no. of Markers × (A)	No. of True Positive	No. of False Recognition	Recognition Rate (% age)
0-60 km/h	05	100	84	3	84
65-80 km/h	05	100	78	3	78
90-110 km/h	03	60	37	5	61.6

3) Recognition Rate in Afternoon

The experiment conducted in afternoon gave good results as in the morning. The results are shown in table VIII.

TABLE VIII. RECOGNITION RATE IN AFTERNOON

Speed	No. of Visits (A)	Total no. of Markers × (A)	No. of True Positive	No. of False Recognition	Recognition Rate (% age)
0-60 km/h	05	100	97	2	97
65-80 km/h	05	100	93	2	93
90-110 km/h	03	60	47	4	78.3

4) Recognition Rate in the Evening

The experiment conducted in the evening gave us good result but low than morning and afternoon. The results are shown in Table IX.

TABLE IX. RECOGNITION RATE IN THE EVENING

Speed	No. of Visits (A)	Total no. of Markers × (A)	No. of True Positive	No. of False Recognition	Recognition Rate (% age)
0-60 km/h	05	100	90	3	90
65-80 km/h	05	100	87	3	87
90-110 km/h	03	60	43	5	71.6

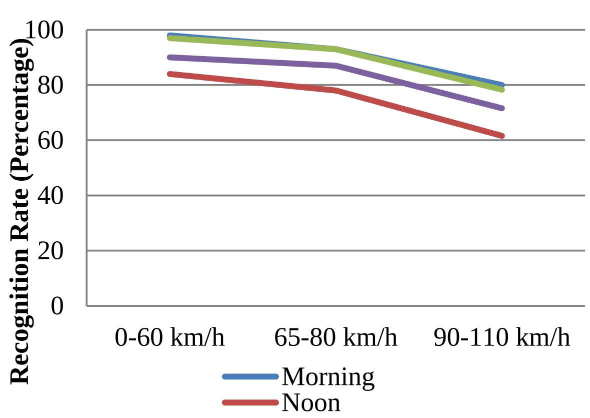


Fig. 6. Recognition rate results graph.

Figure 6 shows the overall results of the percentage recognition rate recorded at morning, noon, afternoon and evening with different car speeds i.e. 0-60 km/h, 65-80 km/h and 90-110 km/h.

The comparison of both the detaction and recognitation rates on the spacified speed is given in figure 7:

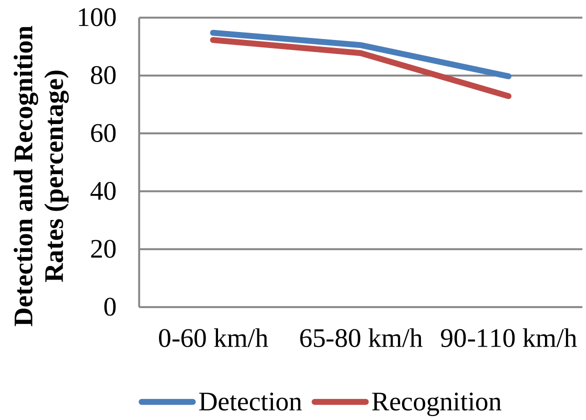


Fig. 7. Detection and Recognition rate comparison.

C. Subjective Evaluation of the System

In this section the proposed system is evaluated subjectively in the form of graph of a survey performed on Pakistani drivers. The system is evaluated on different perameters such as, knowledge, training tool, information/learning, ease of use, interesting and usability.

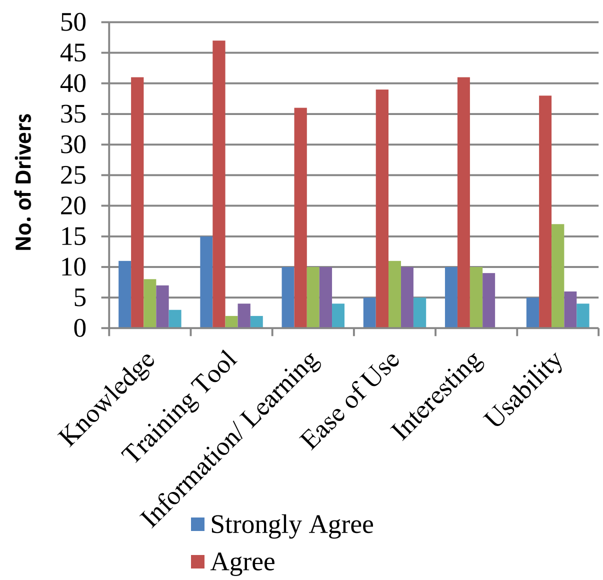


Fig. 8. Subjective evaluation of the system.

In this survay 70 drivers were given the questionnaire after giving them demo of the proposed system. There was variety of drivers for example experienced drivers & Non-experienced (New) drivers and young drivers & old-age drivers. The graph is given in Fig. 8:

D. Limitations of this work

The first challenge in marker-based approach is its tracking process which is extremely sensitive to light and brightness levels. The tracking failure ordinarily happens in outdoor environment where the light and brightness of environment constantly varies depending upon the weather conditions. Fiducial markers with varying lighting and brightness background cause tracking failure. Similarly, the varying sunlight on the camera or fiducial marker also causes tacking failure.

The marker tracking distance is the second challenge that should be tackled. In existing methodology, every marker has its own constrained challenge range. Due the short tracking range, the, marker-based approach is infrequently used in the development of large indoor or outdoor AR applications. There is a need to extend the tracking distance of markers to improve its usability. These issues confine its uses in uncontrolled environment and long range AR applications. Hence, marker based approach of tracking needs change to probe these issues.

V. Conclusion and Future Work

This work analyzed the current marker-based techniques and pointed out some important issues associated to it. The proposed system enhanced the functionality of marker-based tracking techniques in outdoor environment by developing an application capable of detecting the markers designed for Pakistani traffic signs.

This work is primarily concern with Pakistani road traffic signs. Focus on developing an augmented reality application that can detect and recognizes the markers created for Pakistani traffic signs. When the car's camera see the marker it identify it and then generate a voice command associated with that particular traffic sign to the driver so that he may prepare for the hurdle on the road.

During our experiments, we learnt that the speed of car on a rough road produces motion blur in the environment that can affect the marker detection. Therefore, the future work will focus on the design of algorithms that provides a robust and accurate detection and recognition solution for blurred markers. While performing out experiments, we noticed that this our application produces false detection rate (false positive rate, false negative rate and inter-marker confusion rate). Therefore a procedure is needed to overcome this problem of false detection rate.

References

[1] WHO, Global status report on road safety 2015. 2015: World Health Organization.

[2] Gudigar, A., et al., An efficient traffic sign recognition based on graph embedding features. Neural Computing and Applications, 2017: p. 1-13.

[3] Mogelmose, A., M.M. Trivedi, and T.B. Moeslund, Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. IEEE Transactions on Intelligent Transportation Systems, 2012. 13(4): p. 1484-1497.

[4] Barnes, N., A. Zelinsky, and L.S. Fletcher, Real-time speed sign detection using the radial symmetry detector. IEEE Transactions on Intelligent Transportation Systems, 2008. 9(2): p. 322-332.

[5] Broggi, A., et al. Real time road signs recognition. in Intelligent Vehicles Symposium, 2007 IEEE. 2007: IEEE.

[6] Aminian, M.S., et al. Cost-Efficient Traffic Sign Detection Relying on Smart Mobile Devices. in International Conference on Computer Aided Systems Theory. 2017: Springer.

[7] Hsien, J.-C., Y.-S. Liou, and S.-Y. Chen, Road sign detection and recognition using hidden Markov model. Asian journal of health and Information Sciences, 2006. 1(1): p. 85-100.

[8] Zhao, J., Z. Bai, and H. Chen. Research on Road Traffic Sign Recognition Based on Video Image. in Intelligent Computation Technology and Automation (ICICTA), 2017 10th International Conference on. 2017: IEEE.

[9] Chiang, H.-H., et al. Road speed sign recognition using edge-voting principle and learning vector quantization network. in Computer Symposium (ICS), 2010 International. 2010: IEEE.

[10] Nguwi, Y.-Y. and A.Z. Kouzani, Detection and classification of road signs in natural environments. Neural computing and applications, 2008. 17(3): p. 265-289.

[11] Qingsong, X., S. Juan, and L. Tiantian. A detection and recognition method for prohibition traffic signs. in Image Analysis and Signal Processing (IASP), 2010 International Conference on. 2010: IEEE.

[12] Xu, S., Robust traffic sign shape recognition using geometric matching. IET Intelligent Transport Systems, 2009. 3(1): p. 10-18.

[13] Gao, X., et al., Colour vision model-based approach for segmentation of traffic signs. Journal on Image and Video Processing, 2008. 2008: p. 6.

[14] Gao, X.W., et al., Recognition of traffic signs based on their colour and shape features extracted using human vision models. Journal of Visual Communication and Image Representation, 2006. 17(4): p. 675-685.

[15] De La Escalera, A., et al., Visual sign information extraction and identification by deformable models for intelligent vehicles. IEEE Transactions on Intelligent Transportation Systems, 2004. 5(2): p. 57-68.

[16] Escalera, A.d.l., et al., Road traffic sign detection and classification. 1997.

[17] Escalera, S., P. Radeva, and O. Pujol, Traffic sign classification using error correcting techniques. VISAPP (2), 2007. 2007: p. 281-285.

[18] García-Garrido, M.A., et al., Complete vision-based traffic sign recognition supported by an I2V communication system. Sensors, 2012. 12(2): p. 1148-1169.

[19] Keller, C.G., et al. Real-time recognition of US speed signs. in Intelligent Vehicles Symposium, 2008 IEEE. 2008: IEEE.

[20] Loy, G. and N. Barnes. Fast shape-based road sign detection for a driver assistance system. in Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on. 2004: IEEE.

[21] Torresen, J., J.W. Bakke, and L. Sekanina. Efficient recognition of speed limit signs. in Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on. 2004: IEEE.

[22] Bascón, S.M., et al., An optimization on pictogram identification for the road-sign recognition task using SVMs. Computer Vision and Image Understanding, 2010. 114(3): p. 373-383.

[23] Gil-Jimenez, P., et al. Traffic sign shape classification based on Support Vector Machines and the FFT of the signature of blobs. in Intelligent Vehicles Symposium, 2007 IEEE. 2007: IEEE.

[24] Maldonado-Bascón, S., et al., Road-sign detection and recognition based on support vector machines. IEEE Transactions on Intelligent Transportation Systems, 2007. 8(2): p. 264-278.

[25] Messelodi, S., et al., i-Street: Detection, Identification, Augmentation of Street Plates in a Touristic Mobile Application. 2015, Springer International Publishing: Cham. p. 194-204.

[26] Tokusho, Y. and S. Feiner, Prototyping an outdoor mobile augmented reality street application, in Proceedings of the 8th IEEE Symposium on Mixed and Augmented Reality. 2009 p. 3-5.

[27] Chianese, A., et al., A location-based smart application applied to cultural heritage environments, in Proceedings of the 22nd Italian Symposium on Advanced Database Systems. 2012: Sorrento Coast, Italy. p. 335-344.

[28] Briones, G.R. EyeRing voice-activated augmented reality device for the blind. 2012; Available from: http://www.ubergizmo.com/2012/08/eyering-voice-activated-augmented-reality-device-for-the-blind.

DDoS Classification Using Neural Network and Naïve Bayes Methods for Network Forensics

Anton Yudhana[1]
Department of Electrical Engineering
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

Imam Riadi[2]
Department of Information Systems
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

Faizin Ridho[3]
Department of Informatics Engineering
Universitas Ahmad Dahlan
Yogyakarta, Indonesia

***Abstract*—Distributed Denial of Service (DDoS) is a network security problem that continues to grow dynamically and has increased significantly to date. DDoS is a type of attack that is carried out by draining the available resources in the network by flooding the package with a significant intensity so that the system becomes overloaded and stops. This attack resulted in enormous losses for institutions and companies engaged in online services. Prolonged deductions and substantial recovery costs are additional losses for the company due to loss of integrity. The activities of damaging, disrupting, stealing data, and everything that is detrimental to the system owner on a computer network is an illegal act and can be imposed legally in court. Criminals can be punished based on the evidence found with the Forensics network mechanism. DDoS attack classification is based on network traffic activity using the neural network and naïve Bayes methods. Based on the experiments conducted, it was found that the results of accuracy in artificial neural networks were 95.23% and naïve Bayes were 99.9%. The experimental results show that the naïve Bayes method is better than the neural network. The results of the experiment and analysis can be used as evidence in the trial process.**

Keywords—DDoS; IDS; neural network; naïve bayes; network forensics

I. Introduction

The increasing number of internet users has caused many sectors to use online systems to provide services to their clients. This online service is utilized by several sectors such as education, government, and E-Commerce. Vulnerability in online service systems has the potential to be attacked by hackers. Attacks on online services can occur at any time and need solutions to improve them. The attack that is often carried out by hackers is distributed denial of service (DDoS). Kaspersky labs[1] has issued a report on DDoS attacks using botnets that have occurred in the first quarter of 2018. Researcher Kaspersky notes that attacks are often aimed at countries China, the United States, and South Korea because servers located in the country have the most number of online services. Based on the Benchmark Cisco 2018 study of the Asia Pacific Security Capabilities[2], which states that Indonesia has the highest percentage in Southeast Asia by getting attack warnings with the amount of 250,000 - 500.00 per day.

Long-term embezzlement and substantial recovery costs are additional losses for the company due to loss of integrity[3]. The activities of damaging, disrupting, stealing data, and anything that is detrimental to the system owner on a computer network is an illegal act and can be imposed legally in court[4]. Criminals can be punished based on the evidence found with the Forensics network mechanism.

Attack detection is often carried out using the intrusion detection system (IDS)[5] by monitoring the network traffic that is passed. Investigators usually utilize a network monitoring system such as IDS for forensics purposes, where analysis is performed using IDS logs and attack notification systems. Intrusion Detection System (IDS) works by monitoring and warning of suspicious activities that occur on the network and immediately reported as a warning. Using an intrusion detection system is usually done based on a signature. This causes a lot of errors in detecting attacks due to changes in network traffic that have an impact on the high volume of warnings that continue to increase because the data traffic in the network is not stationary to produce and respond to warnings that occur[6]. This error occurs due to a lack of protocol[7] which results in the attacker being able to send attacks more easily with the Ping, Hping, or LOIC tools[8]. The legacy of the syn protocol in network traffic allows IDS to frequently detect attacks because DDoS is done using syn packages.

Signature-based detection systems and attack notifications[9] are not strong enough to serve as evidence in trials. A new approach mechanism is needed to analyze and test the accuracy of DDoS attacks that have been detected by the intrusion detection system (IDS) to strengthen the evidence. Network packet classification is one mechanism that can be done to detect DDoS. Machine learning techniques, by validating network data provided to classify with legitimate observations based on anomalies, can be used in the network forensic process[10]. DDoS attacks through computer networks, especially Local Area Networks (LANs) can be detected using multi-classification techniques, which is by combining data mining methods to get better accuracy[11]. Classification using the Neural Network method in analyzing DDoS attacks can provide 99.6% results based on Hidden Neural Network Variations[12]. The similar analysis was also carried out with the Naïve Bayes method[13] using the KDD99 data set to find the highest accuracy of 99.7837%.

Based on the background above, the research was carried out to determine the process of a new approach in detecting and determining the accuracy of DDoS attacks for network forensic purposes. The study was carried out using a dataset of the Research Laboratory of the Master of Information Engineering of Ahmad Dahlan University (LRIS_MTIUAD). A new approach[14] in detecting DDoS attacks is expected to help develop the ability of Intrusion Detection System (IDS) to predict the presence of DDoS.

II. Basic Theory

A. Network Forensics

Network forensics[15][16] is the process of capturing, recording and analyzing network activities to find digital evidence of an attack or committed crime that carried out using a computer network so that the perpetrator can be prosecuted according to applicable law.

B. Artificial Neural Network

Artificial Neural Network[17] is a biologically inspired computing model consist of various processing elements (neurons). Neurons are connected to elements or weights that build the structure of neural networks. ANN has elements for processing information, namely transfer functions, weighted inputs, and output. ANN is composed of one layer or several layers of neurons as shown in Figure 1.

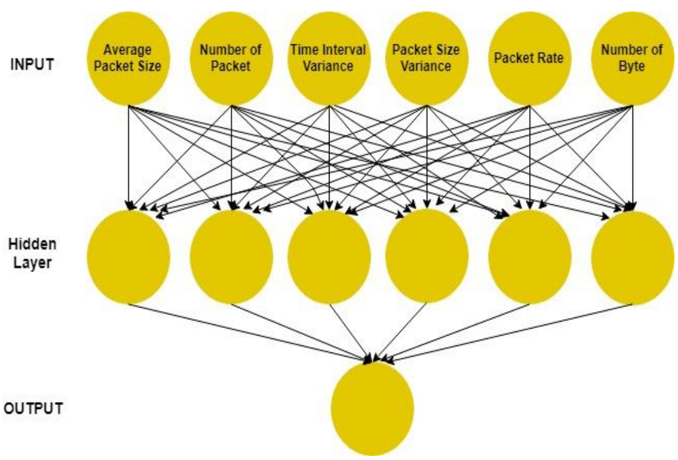


Fig. 1. Artificial Neural Networks

C. Naïve Bayes

Bayes method is used to calculate the probability of an event's occurrence based on the observed observation effect. The Naive Bayes method is a simple probabilistic-based prediction that rely on the application of the Bayes method with a sturdy independence assumption[11][18]. The equation for the Naive Bayes method is:

$$P(H|X) = \frac{P(X|H).\ P(H)}{P(X)} \tag{1}$$

Information:

X : Data with unknown classes

H : Hypothesis X data (a specific class)

P (H | X) : a probability of H hypothesis based on condition X (Posterior Probability)

P (H) : Probability of H Probability (Prior Probability)

P (X | H) : Probability of X based on the condition of hypothesis H

P (X) : Probability of X

III. Research Methodology

This study uses a classification method that consists of 4 stages as shown in Figure 2.

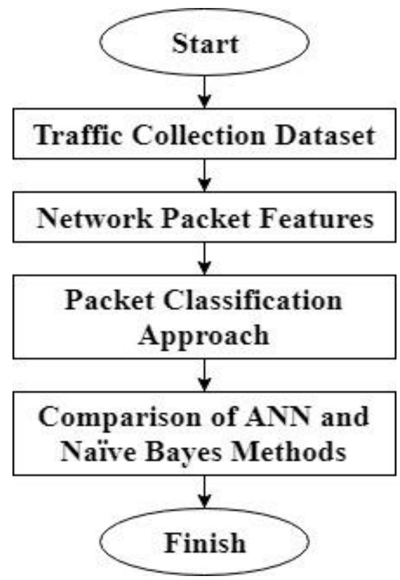


Fig. 2. Classification Method.

Figure 2 can be explained as follows:

A. Traffics Collection Dataset

Traffics Collection is a stage of generating normal datasets and attacks on the Ahmad Dahlan University Research Laboratory network (LRis-UAD) using the Wireshark monitoring application, then the information is stored in the .pcap format as in Figure 3.

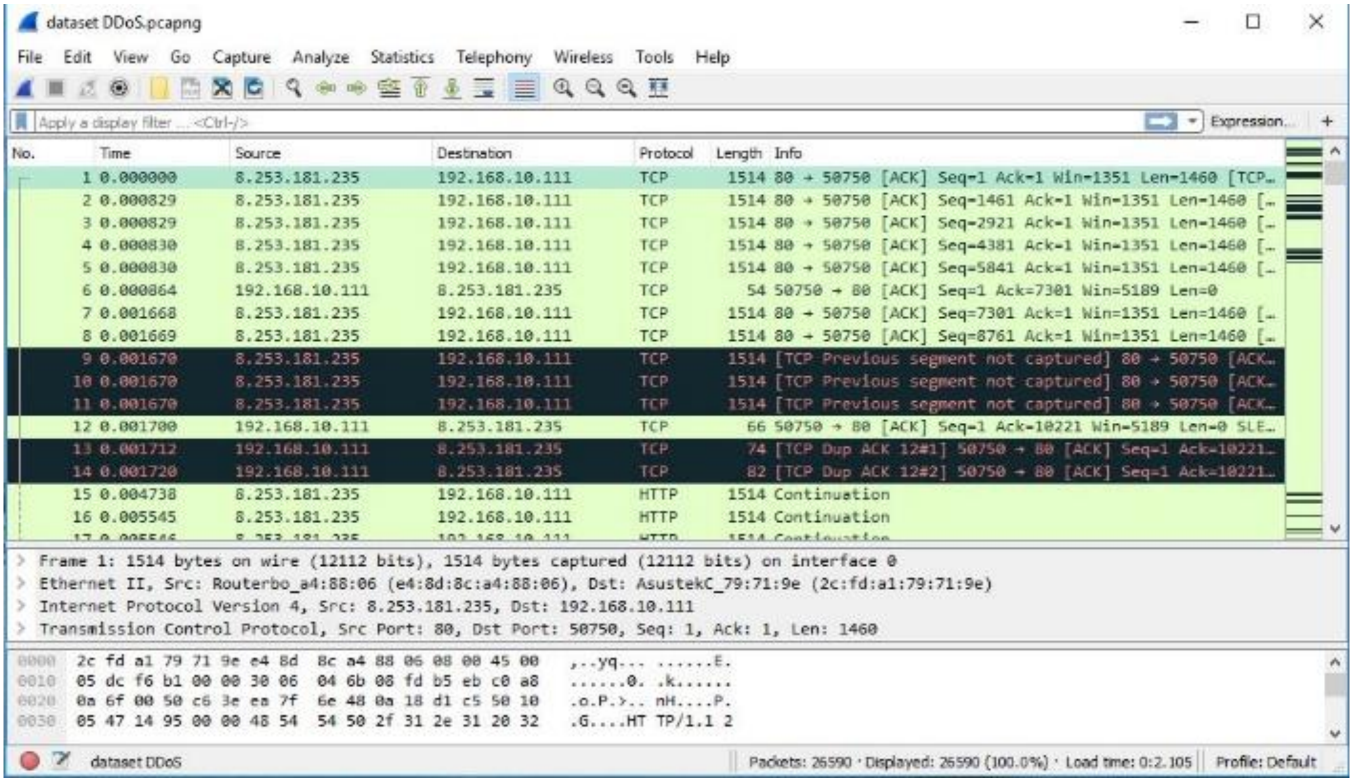


Fig. 3. Traffics Collection Dataset.

B. *Network Packet Features*

Network Packet Features are stages to extract network features in the dataset. The goal is to determine specific patterns in the data. In this case extraction program is carried out with six features using statistical methods. These features are:

a) Average value of network packet length in a predetermined time frame[12].

b) Value the total number of network packages in a predetermined time frame[12].

c) The value of the variance of the time lag variable for the arrival of the network package originating from a particular IP in a predetermined time frame. The value of the variance is generated from equation 2[12].

$$\text{Time Variation} = \sqrt{\frac{\sum(t_n - t)^2}{n}}$$

t_n = the time the package was received

t = average package time received

d) The variance value of the network packet length variable that originates from a particular IP in a predetermined time frame. The value of the variance is generated from equation 3[12][19].

$$\text{Packet size variance} = \sqrt{\frac{\sum(p_n - p)^2}{n}}$$

p_n = length of package received

p = the average length of the package is accepted

e) Package speed values in a predetermined time frame, calculated by equation 4[12].

$$\text{Package Speed} = \text{np} * \frac{1}{T.end - T.early}$$

f) Value the total number of data bits in a predetermined time frame.

C. *Packet Classification Approach*

1) Artificial neural networks: Classification process on artificial neural networks by applying hidden layers carried out by the steps as shown in Figure 4.

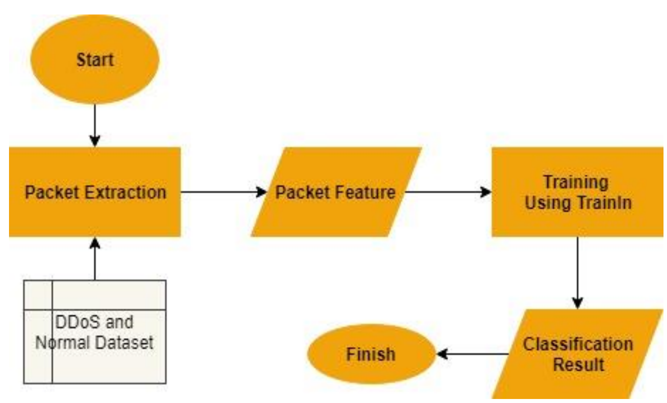


Fig. 4. Neural Network Process.

a) Take a normal dataset and DDoS dataset at the Ahmad Dahlan University Research Laboratory network.

b) Extract network packages using statistical methods to obtain network features, based on 6 inputs: average packet size, number of packages, variant time intervals, package size variants, packet levels, and number of bytes.

c) Use the Tansig, (Tangen Sigmoid), Purelin (Principal Components) and Trainlm (QuasiNewton), training function in Matlab.

d) Classification of classification results using accuracy, mean squared error (MSE), and iteration parameters.

2) Naïve Bayes method: The naïve Bayes classification process is carried out using the statistical method carried out in Figure 5.

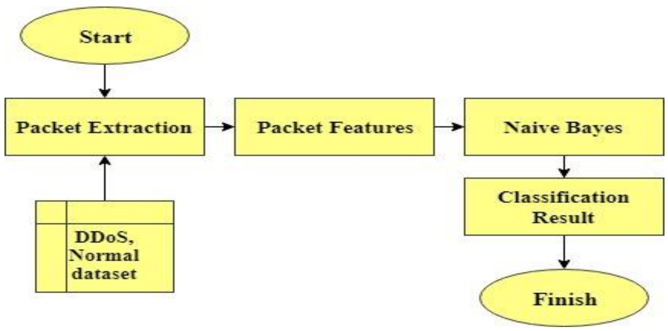


Fig. 5. Naïve Bayes Process.

a) Take a normal dataset and DDoS dataset at the Ahmad Dahlan University Research Laboratory network.

b) Extract network packages using statistical methods to obtain network features, namely the average packet size, number of packages, variant time intervals, package size variants, packet levels, and number of bytes.

c) Perform calculations by looking for standard deviation and threshold based on feature extraction package data.

d) The classification process for naïve Bayes using the Gaussian training phase was tested with test data to determine the success of the introduction of DDoS attacks as predictions.

D. Comparison of ANN and Naïve Bayes Methods

Comparison of artificial neural network and naïve Bayes methods is done to classify and test the best accuracy that can be used in the process of verification in the court. The use and comparison of methods can be carried out as a process of approach to network forensic analysis on DDoS attacks.

IV. Result and Discussion

A. Pre-Processing

Pre-processing data conducted by extracting normal and DDoS datasets with the format. pcap to be .csv so that the statistical calculation process can be performed. The extraction process can be presented in Figure 6.

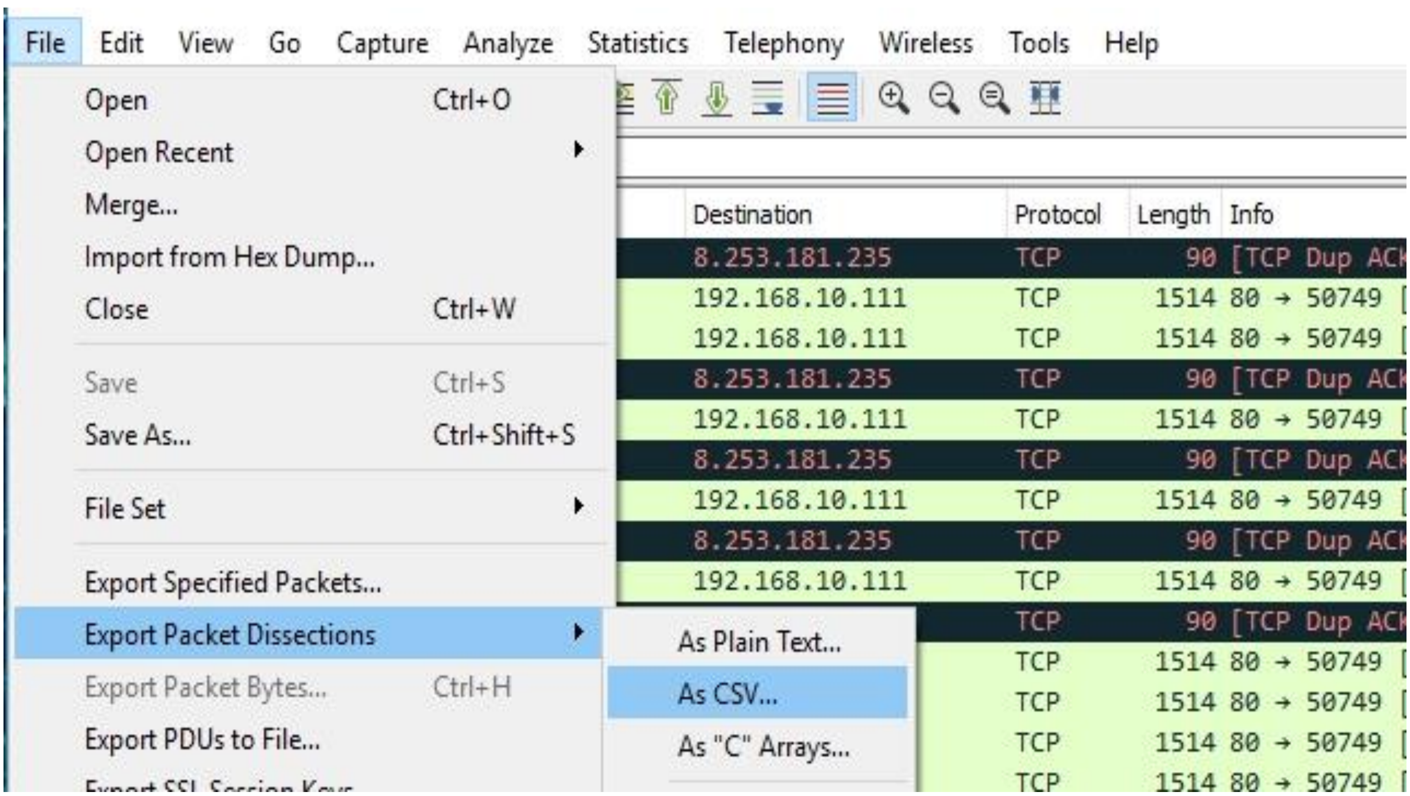


Fig. 6. Extracting .pcap Format Into .csv Format.

B. Packet Extraction

Packet extraction conducted by using 2014b Matlab environment that runs on Windows 10 64-bit. The training process on ANN and naïve Bayes was carried out using 70 DDoS data and 30 Normal Data. Testing process is conducted using 20 data logs on the intrusion detection system (IDS). Data processing is conducted by determining feature extraction based on statistical calculations[6][19]. The summation is conducted by using a fixed moving average window[20] with a duration of 3000 seconds and a 5-second pause with 6 inputs based on the average packet size, number of packages, variant time intervals, package size variants, packet levels, and number of bytes. The quantification process aims to characterize network activity characteristics within a span of time and facilitate the process of training and testing data classification with neural networks. Feature extraction results can be seen in Table 1.

TABLE I. FEATURE EXTRACTION USING STATISTIC

	Average packet size	**Number of packets**	**Time interval variance**	**Packet size variance**	**Packet rate**	**Number of bytes**
Normal	641561	2962	146844	685999	594251	190030
DDoS	553674	3199	142005	637087	778706	177120

Table 1 shows that the range of data values in the network features looks quite large. While the DDoS package after its feature extraction produces values that tend to be monotonous. It can be seen that the difference in data value on the feature extracted from the DDoS package looks quite small.

C. Training and Testing Process on Neural Networks

Training for each variation of artificial neural network architecture in this study is using the Tansig, (Tangen Sigmoid), Purelin (Principal Components) and Trainlm (Levenberg-Marquardt) functions. The purpose of a variety of training functions that provide the highest accuracy in recognizing normal traffic and attacks. Processing of the training process is done using the Matlab program.

The implementation of network package classification training from the method applied using the number of neurons (30-20-1) scheme with Epoch 100 (iteration) and with an MSE value of 0.001. Distribution of datasets for training, validation, and testing is done randomly to avoid bias in the sample pattern. The training process in the hidden layer is done using the sigmoid function. The basic parameters used in the training process are time = 100, function performance = mse, goal = 1e-6, maximum failure = 6, minimum gradient = 1.00e-7, 1.00e + 10. All variations of ANN are trained until the error performance function the mean square error (MSE) is less than 0.001. As presented in Figure 7.

The performance of neural networks after being trained is able to produce a regression value of R-test 0.99 which means that the connection weights between neurons in each layer of neural networks have been able to provide optimal results in recognizing input data patterns. The results of the regression value can be seen in Figure 8.

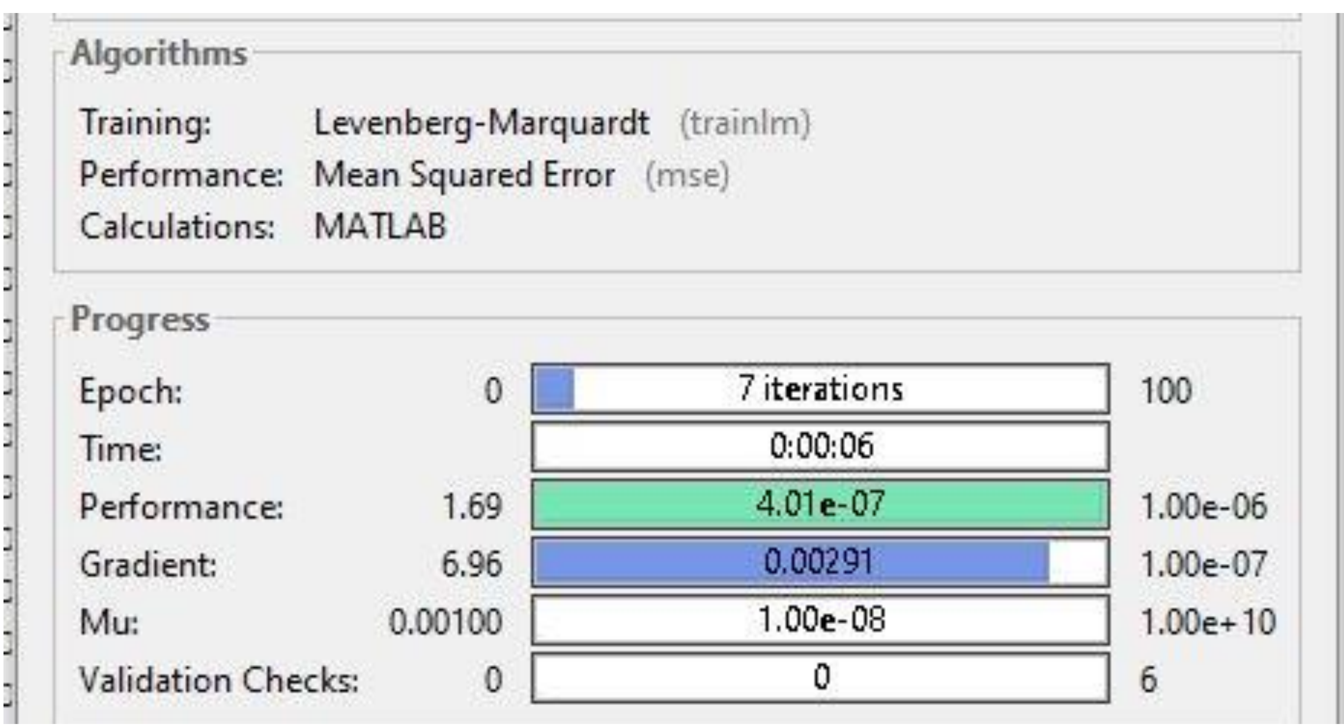


Fig. 7. Training Process on Neural Network.

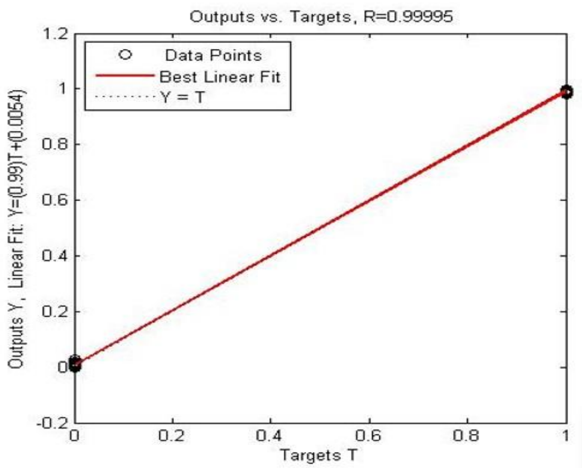


Fig. 8. Results of the Regression Value.

Testing on the neural network is carried out using 20 log data on the IDS to see the DDoS accuracy value. The results of the tests that have been conducted show that the logs stored on the IDS system are detected as DDoS attacks with an accuracy value of 95.23% as shown in Figure 9.

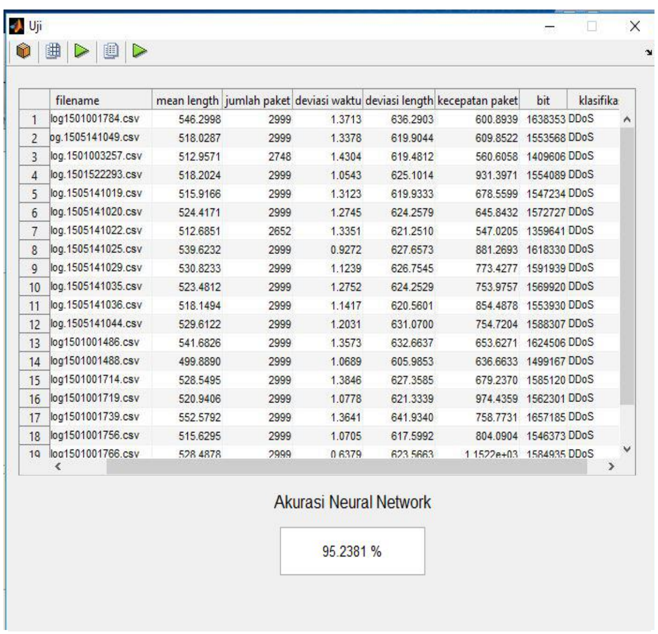


Fig. 9. Classification Results using Neural Network.

Figure 9. Shows that 20 IDS logs tested using artificial neural networks fall into the DDoS category.

D. *Training and Testing Process using Naïve Bayes*

The naïve Bayes extraction process is carried out by processing feature extraction on normal data and DDoS data using statistical calculations. Normal extraction and DDoS results are stored with file name *featureall* made as input data. Extraction results are processed by looking for probability values and standard deviations in DDoS data and normal data as shown in Figure 10. Data processing is done by loading feature all, where feature all contains the results of normal data extraction and DDoS data.

```
%% load feature
load featureall
% Train
%% prob
f=0;
j=0;
clast=featureall(:,end);
for k=1:size(clast,1)
   d=clast(k,:);
   if d==0
      f=f+1;
   else
      j=j+1;
   end
end
proba1=f/(f+j);
proba2=j/(f+j);
proba=[proba1,proba2];

%% mean & std | 0= Normal & 1= DDoS
m0=mean(featureall(1:5,1:end-1))
s0=std(featureall(1:5,1:end-1))
m1=mean(featureall(6:end,1:end-1))
s1=std(featureall(6:end,1:end-1))
```

Fig. 10. Probability Formulas and Standard Deviations.

The formula in Figure 10 shows the probability value in DDoS = 0.5232 and Normal = 0.4767 is found. Average values on DDoS and Normal can be seen in Table 2.

TABLE II. AVERAGE VALUE OF EXTRACTION PACKAGES

	Average packet size	**Number of packets**	**Time interval varianc e**	**Packet size varianc e**	**Packet rate**	**Number of bytes**
Nor mal	7,54541 E+14	1,6816E +14	1,02094 E+14	6,85827 E+14	3,43983 E+14	1,25262 E+14
DDo S	6,49411 E+14	2,30874 E+14	1,18789 E+14	6,54486 E+14	6,06509 E+14	1,38847 E+14

The results of the standard deviation values in Figure 10 are presented in Table 3.

TABLE III. STANDARD VALUE OF EXTRACTION PACKAGE DEVIATION

	Average packet size	**Number of packets**	**Time inter val varia nce**	**Packet size variance**	**Packet rate**	**Number of bytes**
Nor mal	9,40058 E+14	1,06763 E+14	0.425 49	1,51057 E+14	2,15221 E+14	8,01087 E+14
DDo S	1,40532 E+14	1,07386 E+14	0.320 42	3,39475 E+14	6,92154 E+14	5,89783 E+14

The test is carried out using 20 log data on IDS that have been extracted using statistical formulas. Extraction results are stored with the *featureuji* file name as test data. The training and testing were carried out using the naïve Bayes Gaussian method[11] with a formula which can be seen in Figure 11.

```
load featureuji
%% Test
for ii=1:size(featureuji,1)
   x=featureuji(ii,1:end-1);

   %% normal
   m = m0;
   s = s0;
   Y = normpdf(x,m,s)';
   y=1;
   for i=1:size(Y,1)
         y=y*Y(i);
   end
     y1=y*proba(:,1);

   %% serangan / DDoS
   m = m1;
   s = s1;

   Y = normpdf(x,m,s)';
   y=1;
   for i=1:size(Y,1)
         y=y*Y(i);
   end
     y2=y*proba(:,2);

     if y1>y2
        C1=0;
     else
        C1=1;
     end
   C2=y1/(y1+y2)
   C3=y2/(y1+y2)

   anew(ii,:)=C1;
end
```

Fig. 11. Naive Bayes Training and Testing.

The testing conducted shows an accuracy value of 99.9% was found. As presented in Figure 12.

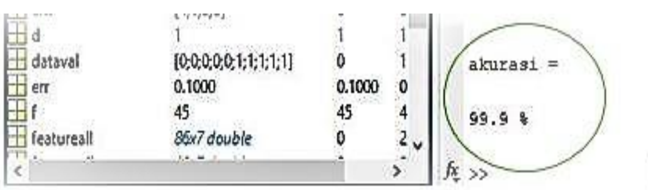


Fig. 12. Classification Results using Naïve Bayes.

V. CONCLUSION AND FUTURE WORK

The experiments carried out concluded that attack information that has been detected by signature-based IDS needs to be re-tested for accuracy using classification with statistical calculations. The test is done by an artificial neural network and naïve Bayes. Based on the analysis and testing conducted, it was found that the accuracy of the artificial neural network was 95.2381% and naïve Bayes was 99.9%. Based on experiments carried out shows that the naïve Bayes method is better than the neural network method. Methods of artificial neural networks and naïve Bayes can be applied in the field of network forensics in determining accurate results and help strengthen evidence in the court.

Further, research will be conducted improved on other parameters such as increasing sample size input patterns presented to the network, variations of hidden layers, reduce the target error and use more training methods. Perform testing using other classification methods such as Support Vector Machine (SVM), K-Means to get better accuracy so that it can be presented in court.

REFERENCES

[1] Symantec, “Kaspersky DDoS Attacks Report in Q1 2018,” 2018.

[2] R. B. Readiness, “Cisco 2018 Asia Pacific Security Capabilities Benchmark Study Regional Breach Readiness Table of Contents,” 2018.

[3] S. Geges and W. Wibisono, “Pengembangan Pencegahan Serangan Distributed Denial of Service (Ddos) Pada Sumber Daya Jaringan Dengan Integrasi Network Behavior Analysis Dan Client Puzzle,” JUTI J. Ilm. Teknol. Inf., vol. 13, no. 1, pp. 53–67, 2015.

[4] A. Fadlil, I. Riadi, and S. Aji, “Development Of Computer Network Security Systems So That Network Forensic Analysis,” J. Ilmu Tek. Elektro Komput. dan Inform., vol. 3, no. 1, pp. 11–18, 2017.

[5] G. Wang, “Neural Network Based Web Log Analysis for Web Intrusion Detection,” IEICE Trans. Inf. Syst., vol. E100D, no. 10, pp. 2265–2266, 2017.

[6] M. Chambali, A. W. Muhammad, and Harsono, “Classification of Network Packages Based on Statistical Analysis and Neural Network,” J. Pengemb. IT, vol. 03, no. 1, pp. 67–70, 2018.

[7] R. F. Fouladi, C. E. Kayatas, and E. Anarim, “Frequency based DDoS attack detection approach using naive Bayes classification,” 39th Telecommun. Signal Process. (TSP), Int. Conf., pp. 104–107, 2016.

[8] A. Iswardani and I. Riadi, “Denial of service log analysis using density k-means method,” J. Theor. Appl. Inf. Technol., vol. 83, no. 2, p. 2, 2016.

[9] J. Fahana, R. Umar, and F. Ridho, “Pemanfaatan Telegram sebagai Notifikasi Serangan untuk Jaringan Forensik,” Query J. Inf. Syst., vol. 1, no. 2, pp. 6–14, 2017.

[10] N. Koroniotis, N. Moustafa, E. Sitnikova, and J. Slay, “Towards Developing Network forensic mechanism for Botnet Activities in the IoT based on Machine Learning Techniques,” arXiv Prepr., 2017.

[11] A. Fadlil, I. Riadi, and S. Aji, “DDoS Attacks Classification using Numeric Attribute-based Gaussian Naive Bayes,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 8, pp. 42–50, 2017.

[12] I. Riadi and A. W. Muhammad, “Network Packet Classification using Neural Network based on Training Function and Hidden Layer Neuron Number Variation,” IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 6, pp. 248–252, 2017.

[13] V. D. Katkar and S. V. Kulkarni, “Experiments on detection of Denial of Service attacks using ensemble of classifiers,” 2013 Int. Conf. Green Comput. Commun. Conserv. Energy, vol. 2, no. 1, pp. 837–842, 2013.

[14] A. Fadlil, I. Riadi, and S. Aji, “Review of detection DDOS attack detection using naive bayes classifier for network forensics,” Bull. Electr. Eng. Informatics, vol. 6, no. 2, pp. 140–148, 2017.

[15] D. Mualfah and I. Riadi, “Network Forensics For Detecting Flooding Attack On Web Server,” IJCSIS) Int. J. Comput. Sci. Inf. Secur., vol. 15, no. 2, pp. 326–331, 2017.

[16] M. Elavarasi, “Network Forensics And Its Investigation Methodology,” Int. J. Emerg. Trends Sci. Technol., vol. 03, no. 05, pp. 852–859, 2016.

[17] A. S. Rawat, A. Rana, A. Kumar, and A. Bagwari, “Application of Multi Layer Artificial Neural Network in the Diagnosis System : A Systematic Review,” vol. 7, no. 3, pp. 138–142, 2018.

[18] M. I. Rahman, N. A. Samsudin, A. Mustapha, and A. Abdullahi, “Comparative Analysis for Topic Classification in Juz Al-Baqarah,” vol. 12, no. 1, pp. 406–411, 2018.

[19] I. Riadi, A. W. Muhammad, and Sunardi, “Neural network-based ddos detection regarding hidden layer variation,” J. Theor. Appl. Inf. Technol., vol. 95, no. 15, pp. 3684–3691, 2017.

[20] J. Ali, S.H Furutani, N Ozawa, sei-ichi Ban, & T Shimamura, “Distributed Denial of Service (DDoS) Backscatter Detection System Using Resource Allocating Network with Data Selection,” Mem. Grad. Sch. Eng. Syst. Informatics Kobe Univ., no. 7, pp. 8–13, 2015.

A Context-Sensitive Approach to Find Optimum Language Model for Automatic Bangla Spelling Correction

Muhammad Ifte Khairul Islam[1], Md. Tarek Habib[2],
Md. Sadekur Rahman[3], Md. Riazur Rahman[4]
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh

Farruk Ahmed[5]
Department of Computer Science and Engineering,
Independent University, Bangladesh

***Abstract*—Automated spelling correction is an important phenomenon in typing that has intense effect on aiding both literate and semi-literate people while using keyboard or other similar devices. Such automated spelling correction technique also helps students significantly in learning process through applying proper words during word processing. A lot of work has been conducted for English language, but for Bangla, it is still not adequate. All work done so far in Bangla is context-free. Bangla is one of the mostly spoken languages (3.05% of world population) and considered seventh language of all languages in the world. In this paper, we propose a context-sensitive approach for automated spelling correction in Bangla. We make combined use of edit distance and stochastic, i.e. *N*-gram language model. We use six *N*-gram models in total. A novel approach is deployed in order to find the optimum language model in terms of performance. In addition, for finding out better performance, a large Bangla corpus of different word types is used. We have achieved a satisfactory and promising accuracy of 87.58%.**

***Keywords*—Spelling correction; non-word error; N-gram; edit distance; magnifying search; accuracy**

I. Introduction

Spelling error is a common problem in every language whether it is in handwritten or in typing form. Therefore spelling checking and correction is always in the focus of computational linguistics for almost in every language. As a result significant efforts on this area have been observed in various languages like English, Chinese and Arabic. Though Bangla is one of the most widely spoken languages (3.05% of world population) and considered seventh language of all languages in the world [1], not so many notable works were found on automated spelling correction. However, it is also observed that in case of all works of spelling correction in Bangla, context-free spelling checking has been deployed by the researchers. Thus context-sensitive spelling checking remains out of focus in Bangla. Therefore, main focus of this research is to propose a context-sensitive language model. The language model consists of stochastic, i.e. *N*-gram language model and edit distance. Here *N*-gram contributes context-sensitive assessment and edit distance contributes context-free assessment. So, we take advantages of both context-free and context-sensitive approaches in our model. Six *N*-gram based stochastic language models are used. We propose a novel approach for finding the optimum language model. The corpora used so far in the works of automated Bangla spelling correction are not so large. The corpus that we use have surpassed all other so far used Bangla corpus in terms of size.

Rest of the paper is organized as follows: Section II highlights the ongoing researches those were targeted to solve the problems related to spelling mistakes. Next section discusses about the types of spelling errors and also about fundamental ideas of the stochastic language models those are used in this research. Then section IV describes the solution approaches of our research and also about the proposed algorithms. Section V includes the experimentations and results of our findings. Then VI presents the comparisons of our findings with the findings of other researches. Finally section VII concludes the paper mentioning the contribution and limitation of this research.

II. Literature Review

A number of research efforts has been performed in order to solve automated spell checking and correction in different international languages. English, Bangla, Arabic and Chinese are some most spoken languages in the world[1].Notable work on spelling checking in English has been done in [2], [3] and[4]. Bangla is the seventh (7^{th}) most spoken language in the world [1]. Some efforts on automated spell checking in Bangla have been reported in [5], [6], [7] and [8]. Likewise, some efforts on automated spell checking in Chinese [9] and Arabic [10] have been reported. Although almost all of them have concentrated on context-free spell checking, very few of them focused on Context-Dependent spell checking, where a lot of potential of ray of success is lying in.

Automated spell checking in Bangla has been experienced by a small number of papers [5], [6], [7], [8], [11], [12], [13], [14], and [15]. All of them concentrated automated context-free spell checking and correction, whereas none has performed Context-Dependent spell checking and correction in Bangla. Although different techniques are deployed in [5], [6], [7], [8], [11], [12], [13], [14], and [15], one thing is common for them. It is the absence of a balanced, big and reputable corpus. P. Mondal and B.M.M. Hossain [5] have used clustering based on edit distance in order to solve the problem of automated Bangla spell checking. Although they have claimed to chance an accuracy of 99.8%, their findings are not performed done to the size of test data use, i.e. 2450 words only. They deal with phonetic and typographical errors. N.U. Zaman and M. Khan [8] have used mapping rule based on edit distance and double metaphone in order to deal with automated Bangla spell checking problem. Though they have claimed to have an accuracy 91.67%, their input data is only 1607 words. Bidyut Baran Chaudhuri [11] used string matching algorithm for identifying phonetic errors. At first, he mapped the phonetically similar single unit of character code in a dictionary. He also construct a reversed dictionary which was used to keep characters of each word in reverse order. Misspelled words were corrected using both dictionary. He claimed that his accuracy rate high with 5% false positive detection. But he dealt with mainly phonetical errors. He need double memory space for one dictionary and its' reversed dictionary. N. U. Zaman and M. Khan [7] modified phonetic encoding based on soundex algorithm for matching Bangla phonetic. They also focused only phonetic errors. M. Z. Islam, M. N. Uddin and M. Khan [12] applied stemming algorithm for spell checking. If the stem is not found, then it produces a suggestions list using suggestion generation process. They used edit distance algorithm to find best match. M. K. Munshi et. al. [13] proposed a probabilistic approach for generating the suggestion list of error words using finite state automaton. Authors of [14] used a direct dictionary look up and binary search for detecting error word and generate suggestions using recursive simulation method. Author of [15] used character-based *N*-gram model for checking correctness of a word in Bangla. But they did not correct incorrect words. As none of them focused the context of the sentence while correcting the incorrect word, their accuracy rate can be changed for the Context-Dependent correction in the test sentences. For example, “লোকটা অনাহারে মচা গেলো।” here “মচা” is incorrect word and corrected word is “মারা”. If we do not consider the context of the sentence their system may be generate words like “পচা”, “মশা”, “মনা” etc. as suggestion in terms of edit distance and phonetical similarity. Their system may not suggest “মারা” word because it has less phonetically similarity and its' distance with incorrect word is more than other words. But these words are inappropriate with the context of this sentence. In this work the accuracy was calculated in terms of the context of the sentence. The program of this work correct the error word based on the context of the sentence which was never done before in Bangla.

Some papers [2], [3], [4] on English spell checking and correction have been studied. In one of the papers, i.e.in [2] direct dictionary lookup method was used to detect incorrect word and then suggestion list was created using edit distance and frequency of the word. They did not mention their corpus size, accuracy rate and test data size. Andrew Carlson and Ian Fette [3] use *N*-gram and confusion set to correct real-word and non-word error. They use Brown corpus and WSJ corpus and they got 96.1% accuracy for real-word and 96.4% for non-word error. Authors of [4] use tribayes (combination of trigram model and Bayes approach) to correct real-word error. They use brown corpus for train their data and Chinese Learners of English Corpus (ELEC) corpus for test data. They got 86.75% accuracy. Some paper on spell checking for other language have been presented. Author of [9] used edit distance algorithm, soundex algorithm, and combined them with pinyin to check and correct Chinese language spelling. They did not mention their accuracy and test data size. Authors of [10] proposed a system for checking Arabic language spelling using context words and *N*-gram language models. Their corpus size is 41,170,678 words. They used twenty-eight confusion sets for their experiment. Their average accuracy rate is 95.9%. They handle real-word errors and non-word errors both.

III. Types of Spelling Error and Stochastic Language Models

All spelling errors have been classified by Kukich [16] into two types. One is real-word error and the other one is non-word error. Real-word error means the word is not contextually appropriate though it is valid. For example, in the sentence “I eat water.”, “eat” is not contextually appropriate but it is a valid word. Similarly, in Bangla, in the sentence “আমি কাক বাড়ী যাব।”, কাক is not contextually appropriate but it is a valid word. So, a real-world error occurs here. Non-word error means the word is not valid lexically. For example, in the sentence “I wnta to go home”, 'wnta' is not a valid word. In the same way in Bangla, 'লারি' is a lexically invalid word in “আমি কাল লারি যাব।”. Kukich [16] has offered some more classification of non-word spelling errors. One is cognitive error and the other one is typographical error. Cognitive error occurs when user does not know the spelling of the erroneous word. Typographical error occurs due to typing mistake. For example, **“আবাসিক”** is a Bangla word, from which different errors **'আববাসিক'**, 'আসকি', 'আমাসকি' and **'আবাকসি'** are caused by insertion, deletion, substitution and transposition respectively.

To correct the non-word spelling error in a Context-Dependent way, we use stochastic language models, i.e. *N*-gram language models. *N*-gram language model is a type of probabilistic language model where the approximate matching of next item is very high. Probability is based on counting things or word in most cases. The probability of a word depends on the previous word which is called Markov assumption. First-order Markov model called bigram looks immediate previous one word and second-order Markov model is trigram looks immediate previous two words and similarly an *N*-1 Markov model is called *N*-gram language model which looks previous *N*-1 words [17]. Thus, the general equation for this forward *N*-gram approximation to the conditional probability of the next word in a word sequence, $w_1, w_2, \ldots.., w_n$, is:

$$P_f(w_n|w_1^{n-1}) \approx P_f(w_n|w_{n-N+1}^{n-1}) \qquad (1)$$

If N = 1, 2, 3 in (1), the model becomes forward unigram, bigram and trigram language model, respectively, and so on. If N=1, forward unigram probability is:

$$P_f(w_n|w_1^{n-1}) \approx P_f(w_n|w_n^{n-1}) \qquad (2)$$

If N=2, forward bigram probability is:

$$P_f(w_n|w_1^{n-1}) \approx P_f(w_n|w_{n-1}^{n-1}) \qquad (3)$$

If N=3, forward trigram probability is:

$$P_f(w_n|w_1^{n-1}) \approx P_f(w_n|w_{n-2}^{n-1}) \quad (4)$$

As like (1), the general equation for backward *N*-gram approximation to the conditional probability of the previous word in a word sequence, w_1, w_2, ……, w_n, ……, w_m is:

$$P_b(w_n|w_{n+1}^{m}) \approx P_b(w_n|w_{n+1}^{n+N-1}) \quad (5)$$

If N = 1, 2, 3 in (4), the model becomes backward unigram, bigram and trigram language model, respectively, and so on.

If N=1, backward unigram probability is:

$$P_b(w_n|w_{n+1}^{m}) \approx P_b(w_n|w_{n+1}^{n}) \quad (6)$$

If N=2, backward bigram probability is:

$$P_b(w_n|w_{n+1}^{m}) \approx P_b(w_n|w_{n+1}^{n+1} \quad (7)$$

If N=3, backward trigram probability is:

$$P_b(w_n|w_{n+1}^{m}) \approx P_b(w_n|w_{n+1}^{n+2}) \quad (8)$$

There is a more other type of *N*-gram based language models that takes the features of forward and backward *N*-gram into account. This a kind of hybrid of forward and backward *N*-gram, which looks immediate *N*-1 words backward and immediate *N*-1words forward. Thus, the general equation for this combined approximation to the conditional probability of the middle word in a word sequence, w_1,w_2, ...,w_n,…,w_m is:

$$P_c(w_n|(w_1^{n-1}w_m^{n+1})) \approx P(w_n|(w_{n-N+1}^{n-1}, w_{n+N-1}^{n+1}) \quad (9)$$

If N = 1, 2, 3 in (8), the model becomes combined unigram, bigram and trigram language model, respectively, and so on.

If N = 1, combined unigram probability is:

$$P_c(w_n|(w_1^{n-1}w_m^{n+1})) \approx P(w_n|(w_n^{n-1}, w_n^{n+1}) \quad (10)$$

If N = 2, combined bigram probability is:

$$P_c(w_n|(w_1^{n-1}w_m^{n+1})) \approx P(w_n|(w_{n-1}^{n-1}, w_{n+1}^{n+1}) \quad (11)$$

If N = 3, combined trigram probability is:

$$P_c(w_n|(w_1^{n-1}w_m^{n+1})) \approx P(w_n|(w_{n-2}^{n-1}, w_{n+2}^{n+1}) \quad (12)$$

IV. Research Methodology

Our proposed approach handles all kinds of non-word errors. Direct dictionary lookup method is used to detect a non-word error. To correct the misspelled word, minimum edit distance method and *N*-gram language model are combinedly used. Six *N*-gram language models, forward bigram, forward trigram, combined bigram, combined trigram, backward bigram, backward trigram, are used separately. After detecting a misspelled word, *N*-gram probability and minimum edit distance for a candidate correction are calculated. *N*-gram probability will contribute for context in further calculations, on the other hand to estimate structural similarity between misspelled word and candidate corrections minimum edit distance is used. It measures the minimum number of total operations required to transform one string into the other. The operations can be insertion, deletion and/or substitution. To calculate edit distance, we use the minimum edit distance dynamic programming algorithm [17] as written in Algorithm 1.Algorithm 1 works by creating a distance matrix with one column for each symbol in the predicted word sequence and one row for each symbol in the error word sequence in order to compare sequence. By using dynamic programming, Algorithm 1 calculates the minimum edit distance, i.e. Levenshtein distance [17], where it is assumed that insertion and deletion each has a cost of 1 and substitution has a cost of 2.

Algorithm. 1. Algorithm for calculating minimum edit distance

```
min_distance(misspelled_word ,
candidate_correction )

m ← length(misspelled_word)
n ← length(candidate_correction)
create distance matrix dis[n+1,m+1]
for each column i ← 0 to n
for each row j ← 0 to m
    dis[i,j]←min(dis[i-1,j]+ins-
        cost(candidate_correction_j),
        distance[i-1,j-1]+
        subst-cost(misspelled_word,
        candidate_correction_i),
        distance[i,j-1)+
        ins-cost(misspelled_word_j))
```

After finding the minimum edit distance $d(\widetilde{w}, w_n)$ between the misspelled word $\widetilde{w}$ and a candidate correction w_n, we normalize the distance using (13).

$$\bar{d}(\widetilde{w}, w_n) = \frac{d_{max}+1-d(\widetilde{w},w_n)}{d_{max}} \quad (13)$$

After normalization, the value of distance ($\bar{d}$) ranges in [$1/d_{max}$, 1]. If the distance d is maximum then the value of normalized distance $\bar{d}$ is $1/d_{max}$ and if the distance d is minimum then the value of normalized distance $\bar{d}$ is 1. In our work, maximum distance is 9 and minimum distance is 1. Thus, *N*-gram probability and minimum edit distance of candidate corrections are calculated, where *N*-gram probability takes context into account and minimum edit distance works context-independently. So, the final score $S_c(w_n)$ of a candidate correction w_n considers both the effects of context dependence, i.e. *N*-gram probability $P(w_n)$ and context independence, i.e. minimum edit distance $D(w_n)$ in the way shown in (14).

$$S_c(w) = (1-\alpha)P(w_n) + \alpha D(w_n), where\ 0 \leq \alpha \leq 1 \quad (14)$$

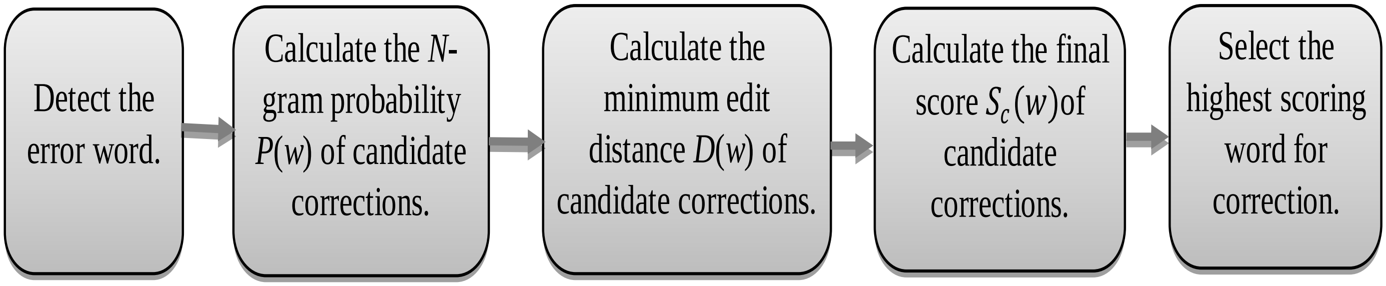


Fig. 1. The Approach for Detecting and Correcting Misspelled Word.

After scoring all candidate corrections the system predicts the word with the highest score as the correct. Suppose the predicted word is $\breve{w}$, then the equation for this word can be written as

$$\breve{w} = \text{argmax}_w S_c(w) \quad (15)$$

The entire process of detection and correction of error word is shown in Fig. 1.

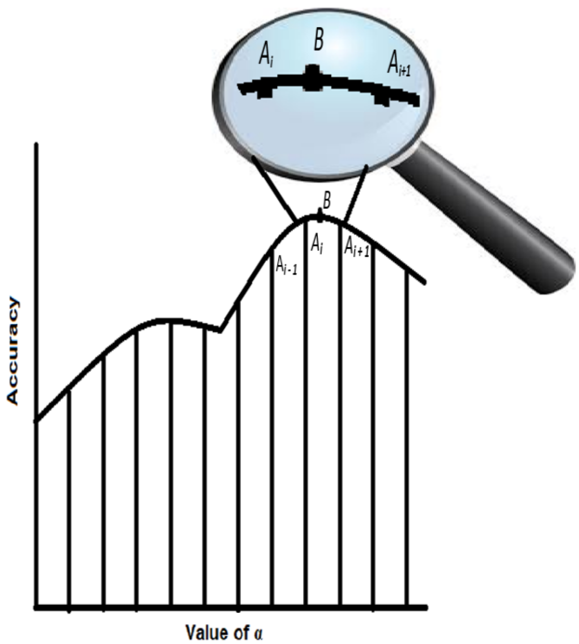


Fig. 2. Underlying Justification Ground for Magnifying Search.

It is easily palpable from (14) that the value of $S_c(w)$ is between 0.0 to 1.0 inclusive since the value of α ranges between 0.0 to 1.0 inclusive. The issue arises from (14) is that what the optimum value (α^*) of α is; that means what the value of α is for which the maximum accuracy is obtained. For this reason, we develop Algorithm 2. We named this Algorithm 2 magnifying search Algorithm. Let us discuss the justification of naming as well as working principle of this algorithm. Suppose that accuracy obtained is represented by A. Hence $A = f(\alpha)$. If we plot these two quantities A and α along x-axis and y-axis, we will obviously get a curve, namely accuracy curve, which will have one or more maximum points. For example, we get an accuracy curve as shown in Fig. 1. Now, we measure the A-values of some equally distant points in order to find the tentative maximum. Of course, we use very small distance; the final maximum will be the tentative one or a point left or right to this tentative point. This is the place where magnifying process comes into play. We magnify the curve fragments left and right to the tentative maximum in order to find more accurate value. We repeat this process until sufficient progress is not made. We can apply the same concept if more than tentative points are found. These entire scenarios are shown in Fig. 2, where A_i is the tentative maximum, $A_{i-1}A_i$ and A_iA_{i+1} are the two curve fragments to the left and right of A_i, respectively. The entire curve fragment $A_{i-1}A_{i+1}$ is magnified here.

Algorithm. 2. Magnifying search algorithm for finding α^*, the optimal value of α for each LM.

```
α* ← 0.0
acc_pre ← Accuracy of LM using α = α*
acc_max ← acc_pre
for i ← 0.01 to 1.0 increasing by .01
  acc_cur ← Accuracy of LM for α = i
      if acc_cur > acc_max
            acc_max ← acc_cur
            α*=i
      else if acc_cur = acc_max
              List.add(i)
End for loop

for each element x is in List
      t = Magnify(x, acc_max, 0.001)
      acc_t ← Accuracy of LM for α=t
      if acc_t > acc_max
            acc_max= acc_t
            α*=t
End for loop
```

After calculating α^* and accuracy for each of the six language models, i.e. forward bigram, forward trigram, backward bigram, backward trigram, combined bigram and combined trigram, the language model with the highest accuracy is considered as the optimum language model.

Algorithm. 3. Magnify algorithm for magnifying search area for accurate value

```
Magnify(α*,acc_max,δ)
  acc_initial ← acc_max
   for i ← α*+δ, j ← α*-δ;i< α*+(δ*10),
         j<α*-(δ*10); i=i+δ,j=j-δ
   acc_i ← Accuracy of LM for α=i
   if acc_i > acc_max
         acc_max ← acc_i
         α* =i
   acc_j ← Accuracy of LM for α=j
if acc_j > acc_max
         acc_max ← acc_j
     α* = j
  End for loop
  Δ← |acc_initial - acc_max|
  if Δ≤ε
     α* = α*
     return α*
  else
   δ← δ/10
return Magnify(α*,acc_max,δ)
```

V. Experimentation

A set of training modules were developed to train the six candidate language models, namely forward bigram, forward trigram, combined bigram, combined trigram, backward bigram and backward trigram. All these models are trained based on a corpus. In our work, we have used a very large Bangla corpus, which was constructed from the popular Bangla newspaper the "Daily Prothom Alo." The corpus contains more than 11 million (11,203,790) words and about 1 million (937,349) sentences, where total number of unique word s is 294,371, average w word length ($|\overline{w}||$) is 7 and average sentence length is 12. During training, the entire corpus is divided into two parts, namely training part and testing part. The holdout method [18] is used to split the corpus at the proportion of two-third for training and one third for testing. Therefore, this work starts with a training corpus of size more than six (6) hundred thousand sentences. In order to avoid model over-fitting problem (i.e. to have lower training error but higher generalization error), a validation dataset is used. In accordance with this approach, the original training data is divided into two smaller subsets. One of the subsets is used for training, while the other one (i.e. the validation set) is used for calculating the generalization error. Two thirds of the training set are fixed for model building while the remaining one-third is used for error estimation. The test data size is more than 3 million (3,734,596) words and about 300 thousand (312,449) sentences where in every sentence there is a misspelled word in different position in the sentences. The holdout method is repeated for five times in order to find the best model for each candidate models. After finding out the best model, the accuracy of the model is computed using the test set, through which the optimum value (α*) is determined based on magnifying search algorithm as given in Algorithm 2 and Algorithm 3.Table 1 shows the values of α* and accuracy for the best model for each candidate models. The accuracy comparison of all the models is presented in Fig. 3, where the optimum value of α of each model is marked with α*.

From Table 1 and Fig. 3, it can be observed that forward bigram language model generates highest accuracy rate 87.578%, where the value of α^* is .75. So, we can claim the forward bigram to be the optimum model. Other models have also shown good accuracy except combined trigram, which gives an accuracy of only 39.68%. Backward bigram shows an accuracy of 85.964%, which is near to the highest obtained accuracy 87.59%. It is also observed from the Table 1 and Fig.3 that as the value of α increase, the accuracy also increases for all models except for backward trigram. The value of α starts increasing just before α equals α^* for backward trigram. After reaching α*, the accuracy starts decreasing slightly and then remains same for all models other than for backward trigram, for which accuracy starts decreasing.

TABLE I. OPTIMUM VALUE OF A (A*) AND ACCURACY RATE OF ALL LANGUAGE MODELS

Language Model	**Optimal value of α(α*)**	**Accuracy (%)**
Forward Bigram	0.750	87.59
Forward Trigram	0.656	78.64
Combined bigram	0.658	72.75
Combined trigram	0.90	39.68
Backward Bigram	0.750	85.96
Backward Trigram	0.505	60.54

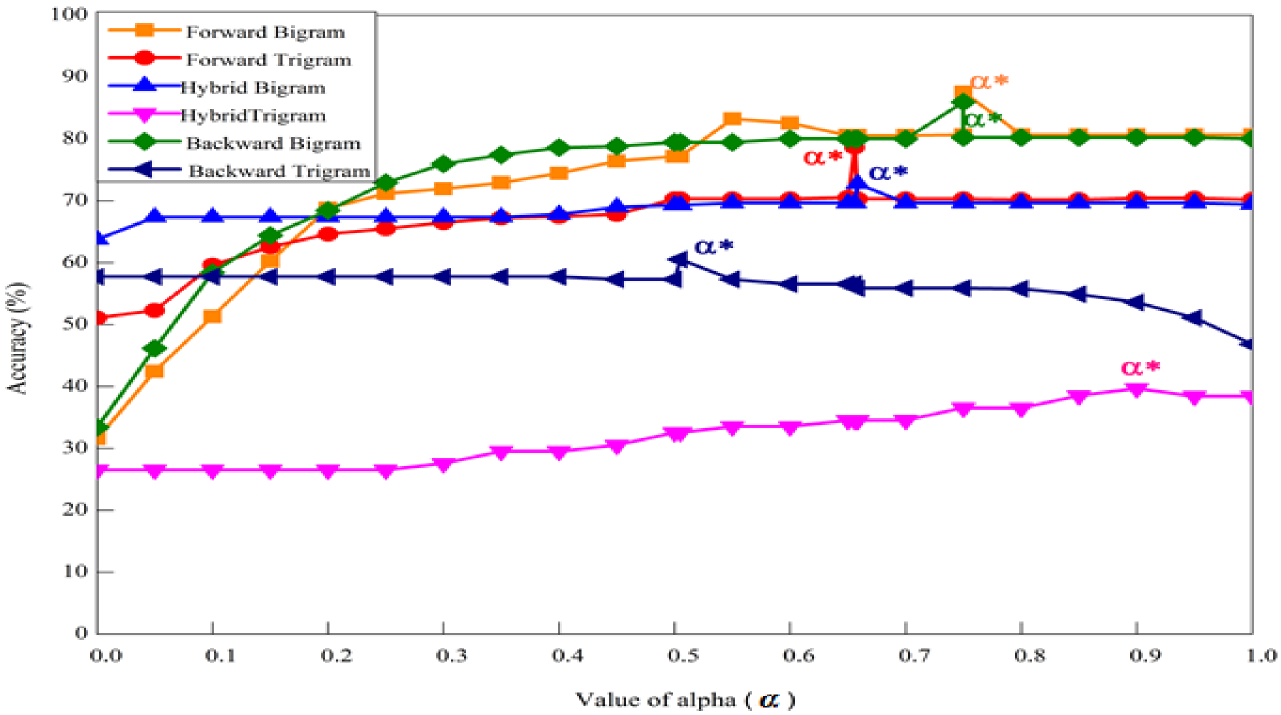


Fig. 3. All Models' Accuracy Comparison Against the Value of α.

TABLE II. ALL MODELS' ACCURACY ACROSS THE MISSPELLED WORD POSITION IN SENTENCES

Misspelled word position	**Forward Bigram**	**Forward Trigram**	**Combined bigram**	**Combined trigram**	**Backward Bigram**	**Backward Trigram**
1	20.24%	30.49%	20%	31.46%	57.28%	70.68%
2	55.36%	52.56%	50.95%	20.0%	53.21%	83.95%
3	55.56%	54.72%	64.03%	21.44%	78.70%	70.64%
4	80.54%	63.57%	59.83%	39.87%	82.36%	76.91%
5	85.634%	75.51%	44.06%	42.13%	87.35%	77.54%
6	68.53%	83.71%	81.96%	48.18%	86.33%	65.37%
7	67.46%	68.23%	86.84%	35.66%	77.70%	64.18%
8	71.85%	65.58%	79.23%	45.29%	79.99%	58.19%
9	73.77%	56.23%	77.05%	25.74%	69.43%	53.53%
10	87.52%	47.07%	54.56%	29.10%	91.72%	47.56%
11	94.66%	85.71%	85.71%	36.29%	45.65%	41.29%
12	95.33%	83.19%	28.51%	28.51%	25.79%	27.66%
Total	87.58%	78.64%	72.75%	39.68%	85.96%	60.54%

In addition, a detailed investigation is conducted, as shown in Table 2, in order to assess the rigorousness of performances of each best candidate language model by varying the misspelled words position in test sentences. The comparison of the six language model's accuracy against the misspelled words position in the test sentences is shown in Fig. 4. From the Fig. 4, it is seen that if misspelled word position is towards the beginning of the sentence then backward bigram, backward trigram and combined trigram show good accuracy rate, but if word position is towards the end forward bigram, forward trigram and combined bigram show better accuracy rate. For middle positions of the sentence all model show good accuracy rate. Combined bigram language model shows almost same accuracy for all positions in the sentence. It can be easily comprehend that if we average the accuracy for all positions, misspelled word forward bigram gives highest accuracy rate of 87.58%.

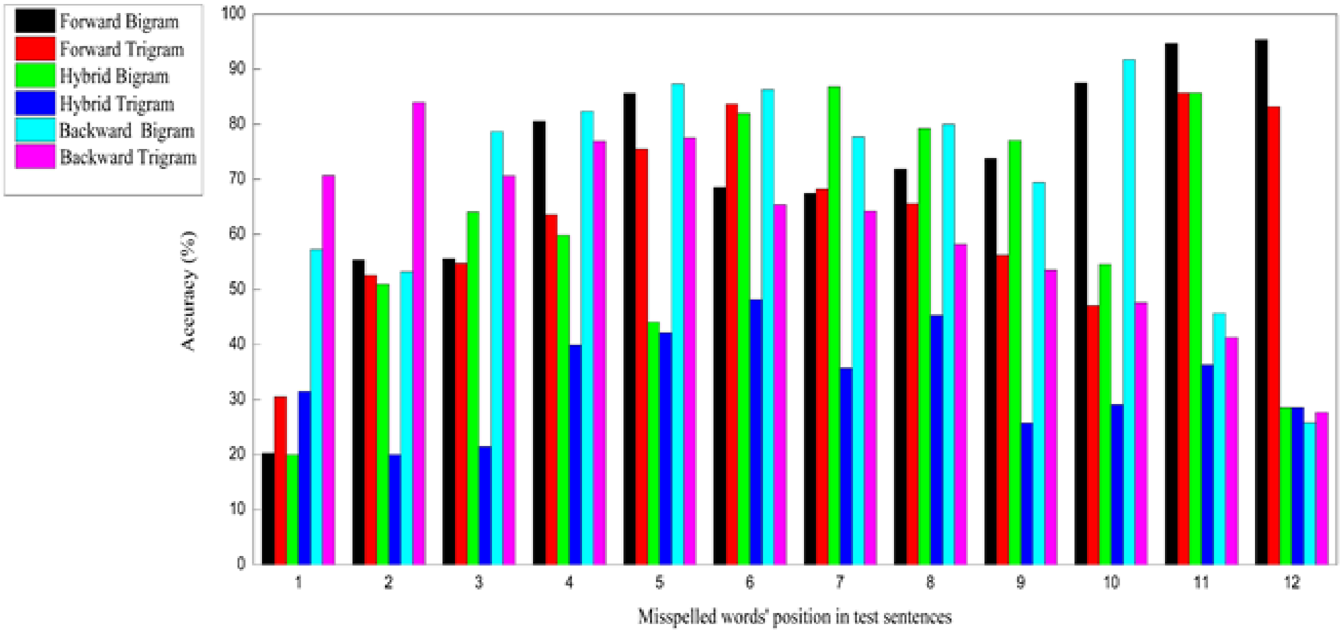


Fig. 4. All Models' Accuracy Comparison Against the Misspelled Words Position.

TABLE III. THE COMPARATIVE NITTY-GRITTY DETAILS OF ALL WORKS REPORTED

Work/Article	Algorithm	Test Data Size	Accuracy	Type of Errors Handled
This work	Context-sensitive technique based on *N*-gram and edit distance	3,734,596 words and 312,449 sentences	87.58%	Non word error
[5]	Clustering	2450	99.8%	Phonetic, typographical
[6]	2-edit distance and phonetic encoding	1607	91.67%	phonetic, typographical, OCR generated
[7]	Phonetic encoding	**NM*	More than 80%	Phonetic, Typographical
[8]	1.Double metaphone 2. Mapping rule.	1607	Edit Distance 0: Correct:1473 Error:134 Accuracy:91.67% Error: 8.33% Error: 107 for edit distance1 Error: 27 for edit distance 2	orthographic rules in Bangla
[11]	String matching algorithm	25,000 words	high accuracy with 5% false positive detection	Phonetic error
[12]	Stemming algorithm and Edit distance	13,000 words	90.8% for correcting single error misspellings and 67% for correcting multiple error misspellings	complex orthographic rules
[13]	Finite state automaton	291 words	92% for correcting single character misspellings and 70% for correcting multiple character misspellings	substitution errors, insertion errors
[14]	Direct dictionary look up method and Recursive Simulation algorithm	*NM*	*NM*	typographical errors and cognitive phonetic errors
[15]	*N*-gram Model (character based)	50,000 correct words and 50,000 incorrect words	96.17%	Non word error

**NM Not mentioned*

VI. Comparative Analysis of Results

It is a matter of fact that automated spell checking in Bangla has been performed in a small number of works. Moreover, all of them concentrated automated context-free spell checking and correction, but none of them has performed context-dependent spell checking and correction in Bangla. The size of test data they used is not so big. Some of them achieved good results, whereas some achieved results, which are not up to the mark. Although different context-free techniques are deployed by them, one thing is common for them. It is the absence of a balanced, big and reputable corpus. In such situation, it is difficult to compare performances obtained by all. In this circumstance, it will not be a callow statement that achieving an accuracy of 87.58% by applying a context-sensitive technique with a training and test data set of big size is quite satisfactory as well as promising. Table 3 shows the comparative nitty-gritty details of all works reported.

VII. Conclusion and Future work

The aim of the research was to find the optimum language model that can assist to overcome Bangla spelling error based on the context. For the purpose of the research a rich and large Bangla Corpus has been used and by applying machine learning techniques on that corpus six language models have been trained for finding the optimum language model for automatic Bangla spelling correction. Finding this language model is the main contribution of the research. Moreover, the approach used for finding the optimum solution is quite novel. Another notable feature of the research is using a large data set for training and testing the model. The accuracy of the model is 87.58% which is good as well as promising. There remains future work for offering a set of corrections rather than offering a single word. Work is in progress to come up with this feature.

References

[1] List of languages by number of native speakers, Available at: https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers. (Last Accessed: March 10, 2018).

[2] Shashi Pal Singh, Ajai Kumar, Lenali Singh, Mahesh Bhargava, KritikaGoyal, Bhanu Sharma, “Frequency based Spell Checking and Rule based Grammar Checking,” in International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) 2016.IEEE,2017, pp 4435 – 4439.

[3] Andrew Carlson,Ian Fette, “Memory-Based Context-Dependent Spelling Correction at Web Scale,” in Sixth International Conference on Machine Learning and Applications (ICMLA 2007), IEEE 2007.

[4] Ya Zhou, Shenghao Jing, Guimin Huang, Shaozhong Liu, Yan Zhang ,“ A Correcting Model Based on Tribayes for Real-word Errors in English Essays,” 2012 Fifth International Symposium on Computational Intelligence and Design.

[5] Prianka Mandal, B M Mainul Hossain, “Clustering-based Bangla Spell Checker,” in 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR). IEEE 2017, pp 1-5.

[6] N. UzZaman and M. Khan, “A comprehensive bangla spelling checker,” In the Proceeding of the International Conference on Computer Processing on Bengali (ICCPB), Dhaka, Bangladesh, 2006.

[7] N. UzZaman and M. Khan, “A bangla phonetic encoding for better spelling suggesions,” in Proc. 7th International Conference on Computer and Information Technology, Dhaka, 2004.

[8] N. UzZaman and M. Khan, “A double metaphone encoding for bangla and its application in spelling checker,” in 2005 International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2005, pp. 705–710.

[9] DongHui Li, DeWeiPeng. “Spelling Correction for Chinese LanguageBased on Pinyin- Soundex Algorithm,” Internet Technology and Applications (iTAP), 2011.

[10] Majed M. Al-Jefri, Sabri A. Mahmoud. “Context-Dependent Arabic Spell Checker using Context Words and N-gram Language Models,” 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, 2013.

[11] Chaudhuri, BidyutBaran, “Reversed Word Dictionary and Phonetically Similar Word Grouping based Spell-checker to Bangla Text,” Proc. LESAL Workshop, Mumbai. 2001.

[12] Md. Zahurul Islam, Md. Nizam Uddin and Mumit Khan, “A Light Weight Stemmer for Bengali and its Use in Spelling Checker,” Proc. 1st Intl. Conf. on Digital Comm and Computer Applications (DCCA07), Irbid, Jordan, March 19-23, 2007.

[13] M. K. Munshi, M. Z. Islam, and M. Khan. "Error-tolerant Finite-state Recognizer and String Pattern Similarity Based Spelling-Checker for Bangla," Proceeding of 5th International Conference on Natural Language Processing (ICON). 2007.

[14] Abdullah, A. B. A., and Ashfaq Rahman. "A Generic Spell Checker Engine for South Asian Languages." Conference on Software Engineering and Applications (SEA 2003). 2003.

[15] Khan, Nur Hossain, et al. "Checking the Correctness of Bangla Words using N -Gram." International Journal of Computer Application 89.11 (2014).

[16] K. Kukich, “Techniques for automatically correcting words in text,” ACM Computing Surveys (CSUR), vol. 24, no. 4, pp. 377– 439, 1992.

[17] Daniel Jurafsky and James H. Martin, (2000), “Speech and Language processing,” USA: Prentice-Hall, Inc.

[18] P.-N. Tan, M. Steinbach, and V. Kumar, “Introduction to Data Mining,” Addison-Wesley, 2006.

Performance Comparison between Merge and Quick Sort Algorithms in Data Structure

Irfan Ali[1], Haque Nawaz[2], Imran Khan[3], M. Ameen Chhajro[5], M. Malook Rind[6]
Department of Computer Science
Sindh Madressatul Islam University (SMIU)
Karachi, Sindh, Pakistan

Abdullah Maitlo[4]
Department of Computer Science
Shah Abdul Latif University (SALU)
Khairpur, Sindh, Pakistan

***Abstract*—In computer science field, one of the basic operation is sorting. Many sorting operations use intermediate steps. Sorting is the procedure of ordering list of elements in ascending or descending with the help of key value in specific order. Many sorting algorithms have been designed and are being used. This paper presents performance comparisons among the two sorting algorithms, one of them merge sort another one is quick sort and produces evaluation based on the performances relating to time and space complexity. Both algorithms are vital and are being focused for long period but the query is still, which of them to use and when. Therefore this research study carried out. Each algorithm resolves the problem of sorting of data with a unique method. This study offers a complete learning that how both of the algorithms perform operation and then distinguish them based on various constraints to come with outcome.**

Keywords—Performance; analysis of algorithm; merge sort; quick sort; complexity; time and space

I. Introduction

An algorithm is any well- defined computational or step by step process for resolving a problem. It takes some values as input and produces some values as output; it will terminate finite number of steps. In mathematics and computer science, an algorithm usually means a logical depiction of the commands which must be performed significant activity [1]. The main factor of analysis of algorithm is to study about time and space and their relationship between the algorithms necessities and number of elements or items being executed or processed. Generally sorting is the method of reorganizing a given set of data and objects within a particular arrangement and that's why, to understand the purpose of the most valuable sorting algorithms have considered as more significant research area nowadays [1, 2] even though, there are many novel sorting algorithms being instigated and used. Many software engineers in their area of programming they are depending on the different sorting algorithms: i-e. Merge, Quick sort etc. sorting is universally globally sorting is performed and it is known as essential activity. Most important sorting application is used in daily life. For example: Phone book faster access to contacts, income tax files, contents of tables, libraries access, dictionaries, search engine [1-9].

A. Computational Complexity

Sorting algorithms, computational complexities, are based on:

- O (nlogn)
- O (log2n)

B. Memory Usage

Classifications of algorithms on the bases of memory

1) When data set is small; internal sorting preferred primary memory only, [2] [1]

2) External sorting uses primary and secondary memory.

II. Research Methods and Discussion

The Algorithms implemented by using java language and discussed the outcomes.

A. Computational Method of Merge Sort

Merge sort algorithm is used DAC (Divide and Conquer) prototype; for example, it split the list of records into two smallest units after that it compare each element with adjacent list and sort the two pieces or units of data sets recursively, consequently it merges and sorted the all the elements in the list. Theoretically, a merge sort perform operation as trails to split the disorder list into n elements sub units or lists, comparing every element of a list of single element observed sorted.

MER-SORT (Data_list, k, m)

1) k < m
[Check base case value]

2) Then x = FLOOR [(k + m)/2]
[Div step]

MergeData (Data_list, k, x)

[Conquer step]

3) Merge Data(Data_list, x + 1, m)
[Conquer step.]

4) Merge Data (Data_list, x, k, m)
[Conquer step]

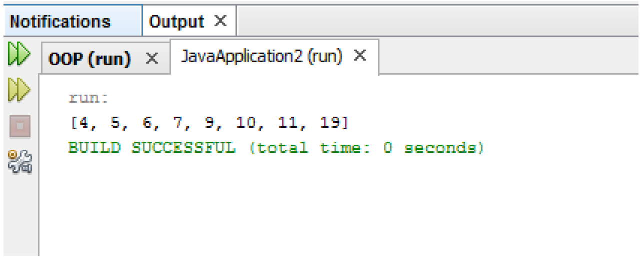


Fig. 1. Execution Output Console.

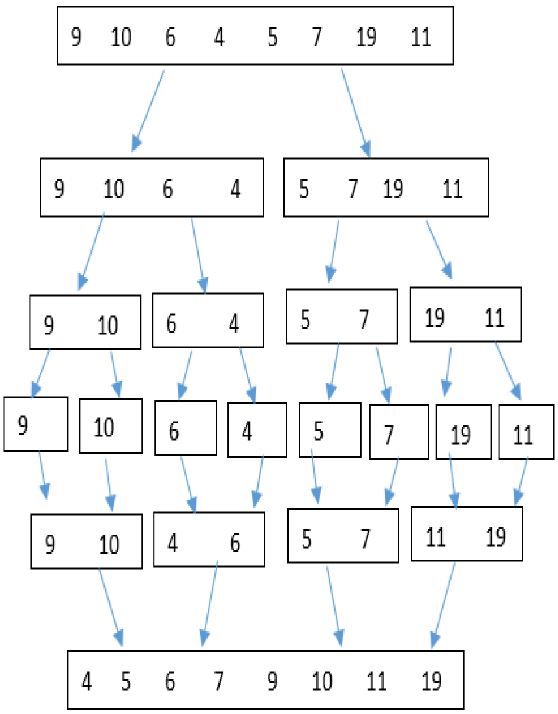


Fig. 2. Merge Sort Pictorial Presentation.

The Fig. 1. Shows the execution output console of merge sort.

To investigate the Merge Sort function, the two different processes that we need to reflect is to form its implementation. Fig.2 shows the list is divided into two parts. The first part of list is the length defined by, half log (n) times wherever n represents the no: of elements in the list. Another part of list is merging of the list. Where every element from the list will be computed, and positioned on the output list that is sorted. Hence, the merging procedure outcomes in a list require n procedures and size n. The analysis gives the outcome, split of log n, and each costs n for a whole of n (log n) processes.

- Best Case Complexity O(nlogn)
- Average Case Complexity O(nlogn)
- Worst Case Complexity O(nlogn)

Advantages:

- Time complexity O(nlogn).
- Used both internal and external sorting
- Stable sort algorithm

Disadvantages:

- As a minimum double the memory necessities of the further sorts since it is recursive.
- Required high space complexity

B. Computational Method of Quick Sort

Quick sort algorithm is used DAC (Divide and Conquer) prototype. Quick sort initial splits a list with in two small sub units: one having low item and another having the high items. Quick sort perform operation to sort sub lists recursively. The implementation activities are: to pick element from the list is called a pivot, from the data list.

Recursively type the sub-list of smaller components and the sub-list of larger components.

Quick sort could be extremely economical algorithmic rule and is predicated on dividing of an array of knowledge with in smaller arrays. An oversized array is divided into 2 arrays one among that hold prices smaller than the desired value, say pivot, supported that the divider is created and another array holds prices larger than the pivot value.

Quick type divides an array and so calls itself recursively doubly to type the 2 ensuing sub arrays. This algorithmic program is sort of economical for large-sized information sets as its average and worst case complexness square measure of O(nlogn), wherever n is that the variety of things.

In fact, it isn't essential to separate the list accurately; even though every pivot devides the weather, 99% one side and 1% on another side, the decision depth remains restricted to, therefore the total period of time remains O(n log n) [3].

Algorithm:

1. quick_sort(DATA, start, end):
2. if (start < end)
    a. set point = partition_list(DATA, start, end)
    b. quick_sort(DATA, start, point - 1)
    c. quick_sort (DATA, point + 1, end)
    [End if in step 2]
3. Exit

Function partition_list(Data_list, start_list, end_list)
1. Set pivotIndx = select Pivot(Data_list, start_list, end_list)
2. Set pivotElement = Data_list[pivotIndx]
3. Set Exchange Data_list [pivotIndx] and Data_list[end_list]
4. Set storeArrIndex = start_list
5. for k = start_list to end_list−1 a)
    if Data_list[k] <= pivotElement
    exchange Data_list[k] and Data_list[storeArrIndex]
    storeArrIndex = storeArrIndex + 1
        [End of if statement in step a]
        [End of for statement in step 5]
6. Exchange Data_list[storeArrIndex] and DATA[end_list]
7. return storeArrIndex

Fig. 3. Shows the output execution console of quick sort.

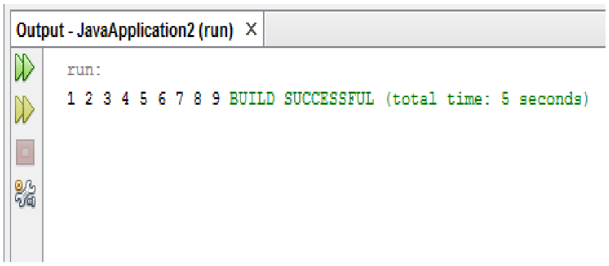


Fig. 3. Execution Output Console.

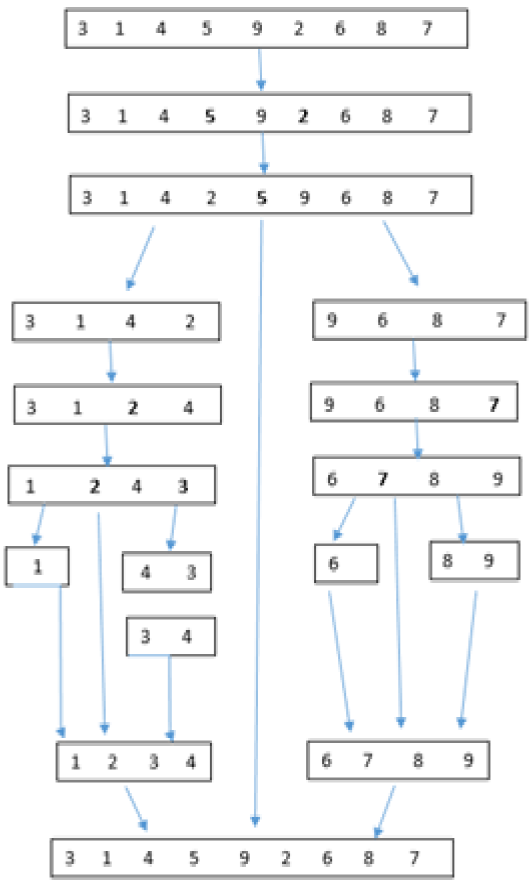


Fig. 4. Quick Sort Pictorial Presentation.

It has been analyzed by using a list has n length of elements, although the partition elements take place at mid of the list shows in fig. 4. More over the complexity will be the log(n) partitions carried out. However, to find themid-pointeach element of the list has been analyzed instead of pivot value. The result has been observed log (n). On other hand in worse case situation did not occur in the middle. It has been observed the mid-point in this case towards the left or right of the mid. That's why it has been observed irregular partition within this case list of n elements sorting split with 0 and n-1 element. After this by using:

n-1 split into 0 size and n-2 size and consequently. As it has been analyzed the recursion requires as $O(n^2)$ sort.

According to above algorithm the base case will be O(nlogn), average case O(nlogn) and worst case $O(n^2)$

Advantages

- No extra memory is required
- Fastest algorithms on average.
- When we have string and integer type of data compassion relatively cheap.
- The list has been traversed successively, and it has been produce very decent locality of location and the performance of cache for arrays

Disadvantages

- In case of recursive function, the average case space complexity is little bit costly, especially when we have large set of data
- Worst case complexity is $O(n^2)$
- Un stable sort algorithm

Comparison of merge and quick sort algorithms on bases on (Base, average and worst case) as showing in the subsequent table 1 [1, 4].

TABLE I. COMPLEXITY COMPARISON

Complexity	***Merge***	***Quick***
Best Case	O(nlogn)	O(nlogn)
Average Case	O(nlogn)	O(nlogn)
Worst Case	O(nlogn)	$O(n^2)$

III. CONCLUSION

From the above analysis, it has concluded that both the quick and merge sort uses DAC (Divide and Conquer) strategy. Both having the average time complexity of O(nlogn). However, both algorithms are quite different. The merge sort is usually required while sorting a too large set to hold or handle in internal memory. It divides the set into a few subsets of one element and then repeatedly merges the subsets into increasingly larger subsets with the elements sorted correctly until one set is left. Usually this method means that the sorting ultimately deals with only portions of the complete set.

In many cases, implementing the quick sort often yields a faster sort other than O(nlogn) sorting algorithms but the worst-case time is $O(n^2)$. The quick Sort operates by selecting a single element from the set and labeling it the pivot. The set is then reordering to ensure that all elements of lesser value than the pivot appear earlier than it and the entire elements of greater value come after it. This operation is recursively applied to the subsets on both sides of pivot until the entire set has been deemed and sorted.

REFERENCES

[1] M. P.K. Chhatwani and M. J. S. Somani ,"Comparative Analysis & Performance of Different Sorting Algorithm in Data Structure" International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, no.11, pp.500-507, 2013.

[2] M. Gul, N. Amin and M. Suliman "An Analytical Comparison of Different Sorting Algorithms in Data Structure" International Journal of Advanced Research in Computer Science and Software Engineein, Vol. 5, no.5, pp. 1289-1298, 2015.

[3] M. N. Kumar and M. R. Singh "Performance Comparison of Sorting Algorithms On The Basis Of Complexity" International Journal of Computer Science and Information Technology Research Vol. 2, no. 3, pp. 394-398, 2014.

[4] G. Kocher1 , N. Agrawal "Analysis and Review of Sorting Algorithms"International Journal of Scientific Engineering and Research (IJSER), Vol. 2, no. 3, pp. 81-84, 2014.

[5] K.K. Pandey, R.K. Bunkar, and K.K. Raghuvanshi. "A Comparative Study of Different Types of comparison Based Sorting Algorithms in Data Structure" International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, no. 2, pp. 304-309, 2014.

[6] Raju and D. Narasimha. "An efficient new approach mean based sorting." In Electrical Computer and Electronics (UPCON), UP Section Conference on, IEEE, pp. 1-5., 2015.

[7] Y. You, P. Yu, and Y. Gan. "Experimental study on the five sort algorithms" In Mechanic Automation and Control Engineering (MACE), Second International Conference, IEEE, pp. 1314-1317. 2011.

[8] D. Rajagopal, and K. Thilakavalli. "Comparison of different sorting algorithms in Data Structure based upon the Time Complexity" International Journal of Technology, Vol.6, no. 1, pp. 23-30, 2016.

[9] Z. Marszałek, "Parallelization of modified merge sort algorithm" Symmetry, Vol. 9, no. 9, pp.1-18, 2017.

Efficient Image Cipher using 2D Logistic Mapping and Singular Value Decomposition

Mohammed A. AlZain
Department of Information Technology
College of Computers and Information Technology, Taif University
P.O. Box 888, Al-Hawiya-Taif, 21974, KSA

***Abstract*—The research paper proposes an efficient image cryptosystem that depends on the utilization of two dimensional (2D) chaotic logistic map (CLM) and singular value decomposition (SVD). The encryption process starts by a confusion stage through applying the 2D-CLM to the input plainimage. Then, the resulted logistic transformed image is then decomposed using the SVD technique into three ciphered components; the horizontal, vertical, and diagonal components. The ciphered horizontal, vertical, and diagonal components are then transmitted to the destination which applies a reverse procedure to reconstruct the original plainimage. A matrix of encryption quality tests are performed for investigating the proposed 2D-CLM based SVD image cipher. The obtained test results confirmed and ensured the efficiency of the proposed 2D-CLM based SVD image cipher.**

Keywords*—*Image cipher; 2D-CL; SVD

I. Introduction

Recently, the unexpected growing in digital technology field witnessed a lot of attention. And this raises and makes the security issues of the data a critical important task especially during the transmission of sensitive data like images, audios, and videos [1-4]. Images may be utilized in different fields like medical, military, and communication fields, the images may carry important valuable information and may be shared over communication networks, so; it becomes a serious issue to maintain them secured as possible against all probable potential attacks. Plenty of works have been done in recent years and resulted in a many studies on image encryption [5-10].

Several image encoding methods were done and founded in literature like ones that depends on classical encryption like Advanced Encryption Standard (AES), Data Encryption Standard (DES), Double-DES, TDES, and International Data Encryption Algorithm (IDEA) [11-15]. Such techniques do not provide a satisfactory outcome because of image intrinsic characteristics like huge bulky capacity, high correlation, and redundancy [16-19].

On contrast, many reported methods in literature uses different transformations like Fractional Fourier Transform (FrFT) for image encoding which may be considered as a generalization of FT [20]. Other transformations that are actually utilized in image encryption include Discrete Wavelet transform (DWT) [21], Fractional Wavelet packet (FWP) [22], and 2D-CLM [23-24]. An image cryptosystem that utilizes a multi-channel/multi-stage FrFT is given in [25]. In [26], the authors presented a proposed an image cipher that performs encryption using an iterative FrFT. A color image cipher which applies FWP is presented in [22]. In [27], a color image cipher is proposed that employed encryption in dual FrFT-WT using random phases. In [28], a color image cipher that employed FrFT in conjunction with DWT is presented. In [29], a secure color image cipher which employed a DWT and SVD is proposed. In [30], an image encoding which utilizes SVD and AT in FrFT is presented. This method works as follows; the plainimage is firstly transformed with FrFT and the transformed image is decomposed with SVD into three components. The three components are AT transformed with different iterations, which resulted in the three ciphered images. In decryption procedure, the three ciphered images are inverse AT transformed using the iterations number and the three components are correctly multiplied. Also, uses the inverse FrFT with corrected parameters to achieve the correct decrypted image.

The 2D-CLM can be employed for pixel shuffling for several many image encryption methods [23]. The other method that may be utilized for improving security is the SVD in which the plainimage is decomposed into three components. Such components with the multiplication order are required during decryption and hence may serve as an additional key.

The main objective of this study is to present an efficient and secure image cipher method for transmitting over secure/unsecure communication network. At encoder end, the encryption process starts by a confusion stage through applying the 2D-CLM to the input plainimage. Then, the resulted logistic transformed image is then decomposed using the SVD technique into three ciphered components; the horizontal, vertical, and diagonal components. At destination, with the knowledge of parameters value, it is easy to retrieve the plainimage from the three ciphered images.

The paper reminder is marshaled as follows: Sect. II provides necessary background regarding the 2D-CLM and the SVD. Sect. III is dedicated for describing in details the proposed 2D-CLM based SVD image cipher. Experimental test results and the analysis of 2D-CLM based SVD image cipher are given in Sect. IV, Sect. V concludes the proposed 2D-CLM based SVD image cipher.

II. BACKGROUND

This section presents the fundamental tools which the proposed image cryptosystem depends on. These tools include the 2D-CLM and the SVD.

A. 2D-CLM

The chaotic logistic is based on several control parameters and this makes it commonly applied in chaos based encryption applications. This is because it is sensitive regarding to initial conditions. The 2D-CLM is derived from the 1D-CLM.

The 1D-CLM can be considered as a simple model that provides a chaotic manner and it can be represented mathematically as [31]:

$$Y_{n+1} = rY_n\,(1 - Y_n) \tag{1}$$

Where Y_n values are in the range [0,1], the r parameter is defined as positive and it allows values in the range [0,4]. Its utilized value controls and gives the manner of how logistic map can work.

The 2D-CLM is more complicated than the 1D-CLM since it has a complex chaotic manner which makes it efficiently sufficient for data encryption. The 2D-CLM can be mathematically described as [23-24]:

$$\text{2D Logistic map}: \begin{cases} x_{i+1} = r(3y_i + 1)x_i(1 - x_i) \\ y_{i+1} = r(3x_i + 1)y_i(1 - y_i) \end{cases} \tag{2}$$

Where r defines system control parameter and (x_i, y_i) represents pair-wise point at iteration i^{th}.

B. The SVD

The Singular value decomposition (SVD) is a commonly famous method in linear algebra, and it has many mathematical applications with respect to matrix inversion. The SVD can be considered as a powerful tool in numerical analysis that can be utilized in matric analysis. The SVD transformation method works by splitting the matrix into three equivalent sized matrices like the input matrix. So, any image may be considered like an array of positive scalar like a matrix. If Z is a square image, defined like $Z \in R$ with size of $n \times n$, where R defines the domain of real numbers, then SVD of Z can be mathematically described as [32]:

$$Z = USV^T \tag{3}$$

Where U and V define the orthogonal vertical and horizontal matrices with condition UT U = I, VT V = I, where I defines the identity matrix. S = diagonal($\sigma 1,\ldots,\sigma p$), p = min(m,n), $\sigma 1 \geq \sigma 2 \geq \ldots \geq \sigma p$ represent Z singular values. Diagonal entries may be considered as Z matrix singular values. The U columns and V columns are the Z left and right singular vectors. The SVD of Z is mathematically defined like given above in Eq. 3, where S, U and V are the diagonal, orthogonal vertical and horizontal matrices.

III. THE PROPOSED 2D-CLM BASED SVD IMAGE CIPHER

The proposed 2D-CLM based SVD image cipher is based on the utilization of 2D-CLM and SVD. The encryption procedure of 2D-CLM based SVD image cipher works through applying the 2D-CLM to the input plainimage. The logistic transformed image is then decomposed with the SVD into the final ciphered diagonal, vertical, and Horizontal orthogonal matrixes which then transmitted to the receiving end. The decryption procedure follows the inverse procedure of the encryption. It starts by applying the inverse SVD transformation to the received ciphered diagonal, vertical, and Horizontal orthogonal matrixes then followed by applying the inverse 2D-CLM to retrieve the final decrypted image. Figs. 1-2 show the encryption/decryption procedures of 2D-CLM based SVD image cipher, respectively.

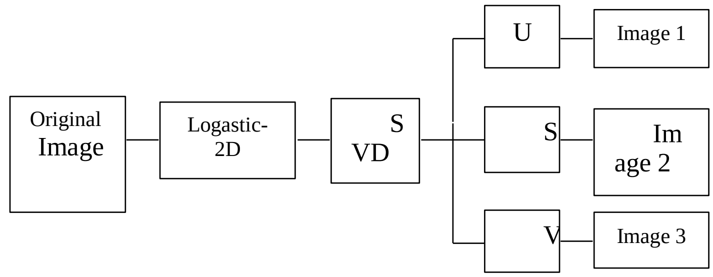


Fig. 1. Encryption Procedure of 2D-CLM based SVD Image Cipher.

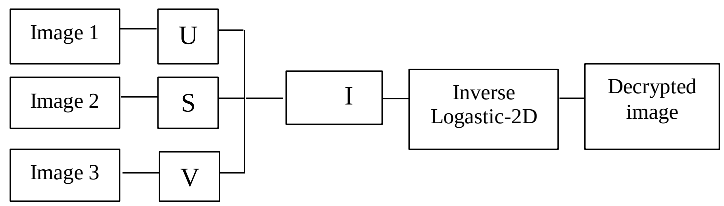


Fig. 2. Decryption Procedure of 2D-CLM based SVD Image Cipher.

IV. SECURITY STUDY

The security investigation of 2D-CLM based SVD image cipher is examined with a matrix of encryption measures and with the visual inspection. The proposed 2D-CLM based SVD image cipher is tested with a group of tests for investigating the performance of the proposed 2D-CLM based SVD image cipher for ciphering gray scale. These tests have been done using a set of different test images; such as Brabra, Boat and Pirate images as shown in Fig. 3.

a) Brabra b) Boat Pirate

Fig. 3. Different Samples of Original Images.

A. Visual Inspecting Results

The visual inspecting results of encrypting different test images; such as Brabra, Boat and Pirate images using the proposed 2D-CLM based SVD image cipher are in Fig. 4. The visual inspection results show that the proposed 2D-CLM based SVD image cipher can successfully hide all the details of the tested plainimages. These visually inspected results have verified and proved the efficiency of 2D-CLM based SVD image cipher.

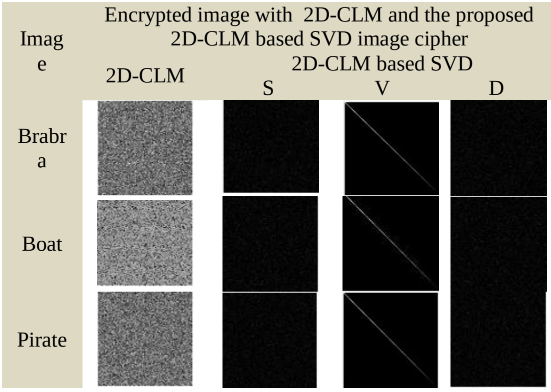


Fig. 4. Obtained Encryption Test Results of using the Proposed the Proposed 2D-CLM based SVD Image Cipher.

B. Histogram Test

The histogram results of examined plainimages and their corresponding cipherimages using the 2D-CLM and the proposed 2D-CLM based SVD image cipher are shown in Fig. 5. The cipherimages histogram results using the proposed 2D-CLM based SVD image cipher are completely different from their corresponding plainimages histogram results. On the other hand, it is noticed that the cipherimages histogram results using the 2D-CLM image cipher are equivalent to their corresponding plainimages histogram results. This is expected since the 2D-CLM just changes the location of pixels.

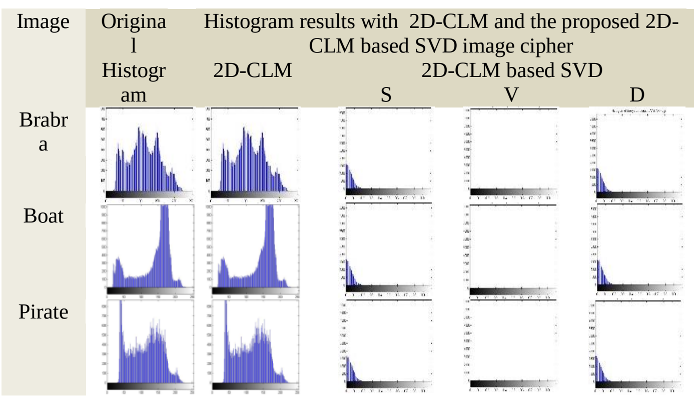


Fig. 5. Histogram test Results of Original/Encrypted Images using the Proposed 2D-CLM based SVD Image Cipher.

C. Entropy Metric

The entropy metric is performed for examining the cipherimages produced using the proposed 2D-CLM based SVD image cipher. The entropy metric can be defined as follows [33]:

$$E = -\sum_{i=1}^{n} P_r(x_i)\log P_r(x_i) \tag{4}$$

Where x_i defines the ith point intensity. High entropy values indicate a best ciphering. The obtained entropy test results using the proposed 2D-CLM based SVD image cipher are given in Table I. The entropy test results indicate that entropy of encrypted images with 2D-CLM is higher compared with their corresponding results using the proposed 2D-CLM based SVD image cipher. But this is expected due to the decomposition nature of the SVD.

TABLE I. ENCRYPTED IMAGES ENTROPY USING THE PROPOSED 2D-CLM BASED SVD IMAGE CIPHER

Image	Entropy values with 2D-CLM and the proposed 2D-CLM based SVD image cipher			
	2D-CLM	2D-CLM based SVD		
		S	V	D
Brabra	7.43567	3.439830	$44.3808*10^{-3}$	3.44555
Boat	7.09441	3.44494	$43.71455*10^{-3}$	3.43681
Pirate	7.2966	3.43596	$44.76*10^{-3}$	3.43

D. Encryption Quality Tests

A set of encryption quality tests are performed for investigating the quality of encryption for the obtained cipherimages using the proposed 2D-CLM based SVD image cipher. This set of encryption quality tests may contain correlation coefficients test (Cr), irregular deviation test (I_d) and histogram deviation test (H_d).

D1. Correlation Coefficients Metric (Cr)

The correlation coefficients test (Cr) is estimated among plainimage/cipherimage as follows [34]:

$$C_r = \frac{\operatorname{cov}(P,C)}{\sqrt{D(P)}\sqrt{D(C)}}, \tag{5}$$

Small Cr values shows high encryption quality.

TABLE II. CORRELATION COEFFICIENT TEST RESULTS USING THE PROPOSED 2D-CLM BASED SVD IMAGE CIPHER

Image	Entropy values with 2D-CLM and the proposed 2D-CLM based SVD image cipher			
	2D-CLM	2D-CLM		
		S	V	D
Brabra	$4.8728*10^{-3}$	$-11.786*10^{-3}$	$4.9116*10^{-3}$	$-2.93400*10^{-3}$
Boat	-1.4715	$796.61*10^{-6}$	$8.443*10^{-3}$	-4.85415
Pirate	$-919*10^{-6}$	$-1.95*10^{-3}$	$-2.73*10^{-3}$	$979.65*10^{-6}$

The achieved correlation coefficients test results for the 2D-CLM and the proposed 2D-CLM based SVD image cipher are given in Table II. The obtained test results show that Cr among plainimage/cipherimage pairs is very small and close to the ideal value of zero that demonstrate high encryption quality.

D2. Irregular Deviation Metric (ID)

The ID metric estimate the encryption quality through computing the abnormality resulted by encryption using the proposed 2D-CLM based SVD image cipher. The IR Metric can be calculated as follows [35]:

$$I_D = \frac{\left|\sum_{i=0}^{255} \left|h(i) - M_h\right|\right|}{MxN}, \tag{6}$$

TABLE III. IRREGULAR DEVIATION TEST RESULTS USING THE PROPOSED 2D-CLM BASED SVD IMAGE CIPHER

Image	Irregular deviation values with 2D-CLM and the proposed 2D-CLM based SVD image cipher			
	2D-CLM	2D-CLM		
		S	V	D
Brabra	$232.231*10^{-3}$	$168.86*10^{-3}$	$168.800*10^{-3}$	$168.861*10^{-3}$
Boat	$203.3*10^{-3}$	$220.92*10^{-3}$	$220.90*10^{-3}$	$220.924*10^{-3}$
Pirate	$234.3*10^{-3}$	$191.68*10^{-3}$	$191.67*10^{-3}$	$191.68*10^{-3}$

Where h(i) is cipherimage histogram at the i intensity value, and M_h is the ideal encrypted histogram. Small ID values demonstrate a good encryption quality.

The ID test results of both the 2D-CLM and the proposed 2D-CLM based SVD image cipher are given in Table III. The achieved testing results showed that the ID values using the proposed 2D-CLM based SVD image cipher are good when compared to their corresponding ID values using the 2D-CLM image cipher. This confirmed the efficiency of the proposed 2D-CLM based SVD compared with the 2D-CLM.

D3. Histogram Deviation Metric (HD)

The HD Metric estimate computes the encryption quality through measuring the variation increase between plainimage (P)/ cipherimage (C) pairs. The HD metric can be estimated as follows [36]:

$$H_D = \frac{\left|\sum_{i=0}^{255} d(i)\right|}{MxN}, \tag{7}$$

Where $d(i)$ defines the difference between plainimage/cipherimage pairs at pixel intensity i^{th}. M and N represent the image height and width. Higher ID values demonstrate good quality for encryption. The HD test results for both the 2D-CLM and the proposed 2D-CLM based SVD image cipher are given in Table IV. It is noticed that the obtained HD values using the 2D-CLM image cipher is zero. This can be interpreted since the 2D-CLM image cipher is just shuffle pixels position and does not change the histogram after employing the encryption. So, it is expected theoretically that the HD values using the 2D-CLM image cipher is zero. Also, results demonstrated that the values of HD using the proposed 2D-CLM based SVD image cipher are high compared with the obtained ID values using the 2D-CLM image cipher. This again confirmed the efficiency of the proposed 2D-CLM based SVD compared with the 2D-CLM.

TABLE IV. HISTOGRAM DEVIATION TEST RESULTS USING THE PROPOSED 2D-CLM BASED SVD IMAGE CIPHER

Image	Irregular deviation values with 2D-CLM and the proposed 2D-CLM based SVD image cipher			
	2D-CLM	2D-CLM		
		S	V	D
Brabra	0	$415.884*10^{-3}$	$372.537*10^{-3}$	$416.072*10^{-3}$
Boat	0	$396.661*10^{-3}$	$374.782*10^{-3}$	$396.657*10^{-3}$
Pirate	0	$427.85*10^{-3}$	$374.78*10^{-3}$	$428.04*10^{-3}$

E. Differential Metrics

Differential metrics are employed for testing the effect of one pixel modificatio in the input plainimage on the resulted cipherimage using 2D-CLM based SVD image cipher. Diffential metrics may include both the Unified Average Changing Intensity (UACI) and the Number of Pixels Change Rate (NPCR) metrics. The NPCR calculates the percentage of unequal pixels in two ciphered images E1 and E2. The NPCR can be computed as follows [37-39]:

$$NPCR(E_1, E_2) = \frac{\sum_{i,j} D(x_i, y_j)}{M \times N} \times 100\%, \tag{8}$$

$$D(x_i, y_j) = \begin{cases} 1 & if \quad E_1(x_i, y_j) = E_2(x_i, y_j) \\ 0 & \text{Otherwise} \end{cases} \tag{9}$$

Where M, N represent both the E1 and E2 height and width.

The UACI evaluates the variance average intensity among two encrypted images, CE1 and CE2 The UACI can be computed as [37-39]:

$$UACI(CE_1, CE_2) = \frac{1}{M \times N}\left[\sum_{x_i y_j} \frac{CE_1(x_i, y_j) - CE_1(x_i, y_j)}{255}\right] \times 100\%, \tag{10}$$

The results of NPCR and UACI between two cipherimages with a modification in one-pixel in their respected plainimages using the proposed 2D-CLM based SVD image cipher are given in Table V. The NPCR and UACI evaluations prove that the proposed 2D-CLM based SVD image cipher is sensitive to tiny changes in the tested images which indicate a good encryption.

TABLE V. NPCR/UACI TEST RESULTS USING THE PROPOSED 2D-CLM BASED SVD IMAGE CIPHER

Image		Irregular deviation values with 2D-CLM and the proposed 2D-CLM based SVD image cipher			
		2D-CLM	2D-CLM		
			S	V	D
Brabra	**NCPR**	**99.3606**	**100**	**100**	**100**
	UACI	**0**	**0**	**0**	**0**
Boat	**NCPR**	**99.0112**	**100**	**100**	**100**
	UACI	**0**	**0**	**0**	**0**
Pirate	**NCPR**	**99.2935**	**100**	**100**	**100**
	UACI	**0**	**0**	**0**	**0**

V. CONCLUSIONS

The paper presented an efficient and secure 2D-CLM based SVD image cipher. The proposed 2D-CLM based SVD image cipher is inspected, examined and investigated with a group of different encryption quality metrics such as visual inspecting, histogram examination, entropy testing, encryption quality measures and differential testing. The obtained test outcomes proved the superiority and affectivity of 2D-CLM based SVD image cipher in terms of different encryption quality metrics.

REFERENCES

[1] A. A. A. El-Latif, L. Li, and X. Niu, "A new image encryption scheme based on cyclic elliptic curve and chaotic system," Multimedia Tools and Applications, vol. 70, no. 3, pp. 1559–1584, 2014.

[2] Osama S. Faragallah, "Secure Audio Cryptosystem Using Hashed Image LSB watermarking and Encryption," Wireless Personal Communication, vol. 98(2), pp. 2009-2023, 2018, Springer (ISI JCR IF 2016: 0.951).

[3] Fuwen Liu, Hartmut Koenig ,"A Survey of Video Encryption Algorithms," Journal of Computers & Security, Vol. 19, No. 1, 2010, pp. 3-15.

[4] Esam Elsheh and Ben Hamza "Secret Sharing Approaches for 3D Object Encryption," journal of Expert Systems with Applications, Vol. 38 , No. 11, 2011, pp. 13906-13911.

[5] X. Zhang, X. Fan, J. Wang, and Z. Zhao, "A chaos-based image encryption scheme using 2D rectangular transform and dependent substitution," Multimedia Tools and Applications, vol. 75, no. 4, pp. 1745–1763, 2016.

[6] Xiaofeng Chen, Xinyi Huang, Jin Li, Jianfeng Ma, Wenjing Lou and Duncan Wong "New Algorithms for Secure Outsourcing of Large-scale Systems of Linear Equations," IEEE Transactions on Information and Forensics Security, Vol. 10, No. 1, 2015, pp. 69-78.

[7] X. Wu, H. Kan, and J. Kurths, "A new color image encryption scheme based on DNA sequences and multiple improved 1D chaotic maps," Applied Soft Computing, vol. 37, pp. 24–39, 2015.

[8] Ibrahim F. Elashry, Osama S. Faragallah, Alaa M. Abbas , S. El-Rabaie, Fathi E. Abd El-Samie, "Homomorphic image encryption," Journal of Electronic Imaging 18(3), 033002, 2009.

[9] Heba M. Elhoseny, Hossam E. H. Ahmed, Alaa M. Abbas, Hassan B. Kazemian, Osama S. Faragallah, Sayed M. El-Rabaie, Fathi E. Abd El-Samie, "Chaotic encryption of images in the fractional Fourier transform domain using different modes of operation," Signal, Image and Video Processing Journal, vol. 9(3), pp. 611-622, 2015.

[10] Fuyan Sun, Shutang Liu, Zhongqin Li, and Zongwang Lü, "A novel image encryption algorithm based on spatial chaos map," Chaos,Solitons and Fractals vol. 38, pp. 631-640, 2008.

[11] B. Schneier, "Applied Cryptography: Protocols, Algorithms, and Source Code in C," John Wiley and Sons, Indianapolis, IN, USA, 2015.

[12] S. Contini, R. L. Rivest, M. J. B. Robshaw, and Y. L. Yin, "The Security of The RC6TM Block Cipher", RSA Laboratories, M. I. T Laboratory for Computer Science, Version 1.0 - August 20, 1998.

[13] A. H. M. Ragab, N. A. Ismail, and O. S . FaragAllah, " Enhancements and Implementation of RC6 Block Cipher for Data Security", Proceedings of International Conference on Electrical and Electronic Technology, Vol. 1, pp. 133-137, 2001.

[14] Luanlan, "The AES Encryption and Decryption Realization Based on FPGA", Seventh International Conference on Computational Intelligence and Security (CIS), pp. 603 – 607, 2011.

[15] M. Feldhofer, J. Wolkerstorfer, and V. Rijmen, "AES implementation on a grain of sand," IEE Proceedings of Information Security, vol. 152, no. 1, pp. 13–20, 2005.

[16] Yuling Luo, Minghui Du," A Novel Digital Image Encryption Scheme Based on Spatial-chaos", Journal of Convergence Information Technology (JCIT), Vol. 7, N. 3, February 2012.

[17] Shan Gai, Guowei Yang, Sheng Zhang. Multiscale texture classification using reduced quaternion wavelet transform. International Journal of Electronics and Communications, 2013,67(3):233-241.

[18] Majid Khan, and Tariq Shah," A Literature Review on Image Encryption Techniques", 3D Res, vol. 29(5), 2014.

[19] G. A. Sathishkumar, K. Bhoopathybagan, N. Sriraam, S. P. Venkatachalam, R. Vignesh, "A Novel Image Encryption Algorithm Using Two Chaotic Maps for Medical Application", Advanced Computing Communication in Computer and Information Science, Berlin-Heidelberg: Springer-Verlag, vol.133, pp. 290-299, 2011.

[20] Li Y, Zhang F, Li Y, Tao R., "Asymmetric multiple-image encryption based on the cascaded fractional Fourier transform," Opt. Lasers Eng., vol. 72, pp. 18-25, 2015.

[21] Ensherah A. Naeem, Mustafa M. AbdElnaby, Hala S. El-sayed, Fathi E. Abd El-Samie, and Osama S. Faragallah, "Wavelet Fusion for Encrypting Images with a Few Details," Computers and Electrical Engineering, vol. 54, pp. 450-470, 2016.

[22] Chen L, Zhao D., "Image encryption with fractional wavelet packet method," Optik, vol. 119, pp.286-291, 2008.

[23] Y. Wu, G. Yang, H. Jin, and J. P. Noonan, "Image encryption using the two-dimensional logistic chaotic map," Journal of Electronic Imaging, vol. 21, no. 1, Article ID 013014, 2012.

[24] Sabry S. Nassar, Nabil M. Ayad, Hamdy M. Kelash, Hala S. El-sayed, Mohsen A. M. El-Bendary Fathi E. Abd El-Samie, Osama S. Faragallah, "Secure Wireless Image Communication Using LSB Steganography and Chaotic Baker Ciphering," Wireless Personal Communication, vol. 91(3), pp. 1023-1049, 2016.

[25] Liu S, Mi Q, Zhu B., "Optical image encryption with multistage and multichannel fractional Fourier-domain filtering," Opt. Lett., vol. 26, pp. 1242-1244, 2001.

[26] Zhang Y, Zheng CH, Tanno N., "Optical encryption based on iterative fractional Fourier transform," Opt. Commun., vol. 202, pp. 277-285, 2002.

[27] Chen L, Zhao D., "Color image encoding in dual fractional Fourier-wavelet domain with random phases," Opt. Commun. vol. 282, pp. 3433-3438, 2009.

[28] Prasad A, Kumar M, Choudhury DR., "Color image encoding using fractional Fourier transformation associated with wavelet transformation," Opt. Commun., vol. 285, pp. 1005-1009, 2012.

[29] Kumar M, Agrawal S., "Color image encoding in DOST domain using DWT and SVD," Opt. Laser Technol., vol. 75, pp. 138-145, 2015.

[30] Chen L, Zhao D, Ge F., "Image encryption based on singular value decomposition and Arnold transform in fractional domain," Opt. Commun. vol. 291, pp. 98-103, 2013.

[31] Chai Wah Wu, Nikolai F. Rul kov, "Studying Chaos via 1-D Maps", IEEE Transactions on Circuits and Systems: Fundamental Theory and Applications, vol. 40(10), 1993.

[32] B. Sarwar , G. Karypis, J. Konstan, J. Riedl, "Incremental Singular Value Decomposition Algorithms for Highly Scalable Recommender Systems," Fifth International Conference on Computer and Information Science, pp. 27-28, 2002.

[33] G. Chen, Y. Mao, and C. K. Chui, "A symmetric image encryption scheme based on 3D chaotic cat maps," Chaos, Solitons & Fractals, vol. 21, no. 3, pp. 749–761, 2004.

[34] L. Xu, X. Gou, Z. Li, and J. Li, "A novel chaotic image encryption algorithm using block scrambling and dynamic index based diffusion," Optics and Lasers in Engineering, vol. 91, pp. 41–52, 2017.

[35] H. Elkamchouchi and M. A. Makar, "Measuring encryption quality of Bitmap images encrypted with Rijndael and KAMKAR block ciphers," in Proceedings Twenty second National Radio Science Conference (NRSC 2005), pp. C11, Cairo, Egypt, Mar. 15,17, 2005.

[36] I. Ziedan, M. Fouad, and D. H. Salem, "Application of Data encryption standard to bitmap and JPEG images," Proceedings Twentieth National Radio Science Conference, pp. C16, Egypt, Mar. 2003.

[37] Osama S. Faragallah, Ashraf Afifi, "Optical Color Image Cryptosystem Using Chaotic Baker Mapping Based-Double Random Phase Encoding", Optical and Quantum Electronics, vol. 49(3):89, pp. 1-33, 2017.

[38] Fatma Elgendy, Amany M. Sarhan, Tarek E. Eltobely, S. F. El-Zoghdy, Hala S. El-sayed, Osama S. Faragallah, "Chaos-based model for encryption and decryption of digital images," Multimedia Tools and Applications, vol. 75(18), pp. 11529-11553, 2016, Springer (ISI JCR IF 2016: 1.530).

[39] Osama S. Faragallah, "Optical Double Color Image Encryption Scheme in the Fresnel-based Hartley Domain Using Arnold Transform and Chaotic Logistic Adjusted Sine Phase Masks," Optical and Quantum Electronics, vol. 50(3):118, pp. 1-27, 2018.

Agent-Based Co-Modeling of Information Society and Wealth Distribution

Fayçal Yahyaoui*, Mohamed Tkiouat
IFE-Lab, Laboratory of Research in Applied Mathematics, Ecole Mohammadia d'Ingenieurs,
Mohammed V University, Rabat, Morocco

Abstract*—With empirical studies suggesting that information technology influence wealth distribution in different ways, and with economic interactions and information technology adoption being two complex phenomena, there is a need for simulation approach that addresses the whole complexity of the issue without being too costly in terms of computations and without ignoring relevant empirical facts in defining the behavior of different agents.. While this problem seems to require a bottom-up approach using agent-based modeling, further complexity levels in managing the heterogeneous agents in space and time and an appropriate separation in domain areas show its limitations in practice. In this paper we illustrate the use of novel multi-level agent based concepts on this socio-economic issue, by considering our studied phenomenon as an interference of multiple simple other phenomena, namely a basic producer/consumer economy and a diffusion of information model. Such an approach involves writing models following a formalism allowing compatibility and exchange of variables, in addition to implementing appropriate synchronization algorithms. Our simulation used Levelspace, a recent extension project of Netlogo simulation tool combined with data exploration tools but the patterns described are generic and can be implemented in other simulation tools. Indeed, our case study offers a building block for a framework that can investigate wealth dynamics and other analogue cases with influence between models. Our approach successfully validates against empirical macro-trends in the distribution of wealth and other social patterns. Thanks to its flexibility in conducting experiments, we could reduce the hypotheses that restricted previous models from conducting a multi-dimensional analysis for the Gini index and enabled solving conflicting research issues.

Keywords*—*Agent-based modeling; computational economics, multi-level; complex systems; parallel computing

I. INTRODUCTION

Economics is considered a complex system where interactions and behavior of micro-elements produce macro phenomena that cannot be understood by a top-down linear decomposition[1]. Agent based modeling is an alternative to statistical modeling by trying to understand this micro-macro link with bottom-up experiments where simple actions of agents are defined with minimal assumptions based on empirical facts to study resulting phenomena. In its basic form agent-based simulation is still limited to address the complexity of economics and other complex phenomena that have different levels of representations. [7]

In fact, a level can be defined as a point of view in a system, integrated in a model as a specific abstraction [8]. We can note that agents can act on different levels, for instance policy makers generally interact with conceptual aggregations of agents (companies, sectors) rather than atomic agents. Some elements of agent-based simulations can also be influenced or constituted by interactions related to another phenomena represented by an agent-based approach. [8]

Many agent-based modelling tools exist to describe models and conduct such experiments [2], they differ mainly in user experience and software environment. Concurrently, research is continuously conducted on different aspects either in simulation theory to provide formalisms that can generalize and fit different modeling situations and also on improving the performance, flexibility, of the tools to be able to reproduce real phenomena

For wealth distribution, we can observe that a small change in a population can induce unpredictable change in wealth curves in the populations with simplest economic interactions [4], which explains the relevance of agent-based modelling for this issue as mathematical equations become quickly unsolvable if we consider the constraints and irregularities: an indicator that such a system falls into the complex category.

The aim of this paper is to show how recent multi-level agent based concept can offer effective ways and new perspective to study classical complex economic issues of interest rather than making the simulation more complicated or adding hypotheses that restrict the conduction of artificial experiments. We illustrate our paradigm through a multi-level simulation that assesses the impact of information society on wealth distribution where the design and performance outperform previous works.

The paper is organized as follows: In section 2, we review briefly the related literature, in section 3 we describe the methodology adopted and show why it's relevant for the phenomena addressed in this paper. In section 4, we describe the model and its specifications to finally discuss the results obtained and the approach in section 5

II. RELATED LITERATURE

Bottom-up approaches using agent based modeling and experiments on artificial societies to understand economic have offered an important alternative to classical economics in what is known as agent based computational economics [1] where many simulations manage to better understand phenomena traditionally explained by top down approach such as macroeconomic policies [11] .

The limitations of traditional models lie mainly in considering the agents as heterogeneous and perfectly rational while the empirical facts say otherwise. Those models also ignore the fact that a system can present emergent phenomena unexplained by summing its different parts.

Wealth distribution, which also appears in many agent based simulations is a particularly complex issue due to its interdisciplinary nature [3] where modeling research aims to understand how different social and economic factors influence wealth inequalities. For instance, the impact of technology on wealth distribution from an industrial perspective where technological advances affect productivity and employment was investigated in one of these models [2]. While the model showed how a behavioral paradigm outperformed traditional models by being able to define heterogeneous agents with a bounded rationality and reproducing emergent historical changes ,the structure of the model remained specific to the studied case and does not offer any design pattern that can be generalized to similar situations.

Also, with the large diffusion of information technology, many agent-based models try to simulate information society trends and other connected phenomena such as smart grids [15]. However, most research on the influence of modern technologies on wealth distribution remains based on descriptive statistics [10], mainly due the multi-level nature of the problem where the first one belongs to microeconomics and the second one belongs to complex network theory. In fact, recent literature acknowledges that multi-level agent based modeling is relevant to socioeconomic systems with different points of view which is exactly the case for our problematic. For this purpose, many generic or domain-specific frameworks using different approaches try to address the multi-level issue by considering the levels as either aggregations of agents or by connecting models or by defining a global system reaction to each agent action [Siebert, 2010]. Those concepts have been able to conduct successful simulations of multi-level complex phenomena such urban growth, cancer modeling, and many other complex phenomena [7].

However, there are no examples in the literature of applications of these recent concepts to simulations of socioeconomic phenomena. Limited examples of multi-level agent-based simulations can be found but they have in common that all the levels are included as agent interaction with other agents without a separation of scale or complexity thus increasing the computational cost of the simulation

III. METHODOLOGY

A. Formalism

Let **M** be a model defined by a group of agents interacting using specific procedures in an environment **E**. Let 's assume that we manage to isolate in the execution of the model some procedures involving only groups of agents without the others. The execution of the model would be equivalent to **(M1, R1, M2, R2,…Mi, Ri)** where **Mi** are the isolated sub models and **Ci** are model subroutines that could not be reified as a sub model. Depending on the complexity of these subroutines, they could be or not integrated as dynamic variables of the environment or approximated using statistical laws. Let's take for example the case of an agent that increases the price of a good periodically where the good is not an agent property, this can be easily translated into a variable of the environment following an arithmetic sequence. In other cases, we can also replace a decision mechanism in these subroutines with decisions based on observations of the environment, as in an inventory of sales of some category of products the results can suggest that some agents are doing financially well. While it is not always feasible to be able to view a phenomenon as an interference of other phenomenon's in case those subroutines have considerable complexity, our goal is to reach a situation describe by Fig. 1.

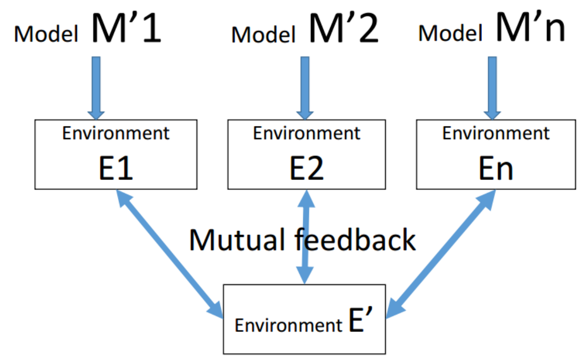


Fig. 1. Model Decomposition.

M'i describes domain specific versions of the sub models **Mi** with respective environments **Ei** that can interact with the dynamic environment. An analogy of this pattern exists in signal processing where a noisy signal can be decomposed to known simpler signals. In our case the signal is the wealth distribution. This distribution results from interactions of different agents in the economy (consumers, producers,) as specified by other models. We can note however that this socioeconomic factor can be influenced by many different dynamics of society whether behavioral or economic which can be themselves complex and can't be simple components of a model but require themselves to be modeled. The model mentioned in section 2 was constrained to make a simplistic formulation of technological progress to be able introduce it in the economic model.

For our study, modeling trends in a more connected society and wealth distribution dynamics are separate areas of expertise and we suppose a preexistence of model examples. We consider the connected society as a separate level that changes the environment of economic interactions where agents interact, which is reflected in a new kind of transactions, associations, goods and other economic possibilities due to distant communication and large availability of information

Linking such models however requires both conceptual and technical tools for information exchange and synchronization.

B. Tools

For our simulation we use Netlogo, one of the most used agent based platforms in different fields [6]

Experiments in Netlogo simply define agent properties and procedures then run simulations for different inputs while monitoring outputs of interests, which can result in dataset for later analysis or simply acknowledge the existence of an emergent phenomenon.

The growing research trends to take multi –level aspects in agent based simulations had led the founders of Netlogo project to conduct a research project that led to the development of Levelspace [9], an extension that allows users to take into consideration the multi-level aspects in different ways. This extension can call other simulations launched simultaneously to either connect different phenomena or reify agents in the simulation as resulting from another simulation.

A classic example is the wolf sheep predation model, where Levelspace can make the grass growth controlled by a climate change model that would have made the simulation unmanageable if included in the same model. Levelspace with other frameworks views the multi-level aspects in a model as a society of interacting models [9].

For our example, we consider informations society changes happening in a different level and we connect it to a classical economic interactions model. The “introduceICT” button in the first part of Fig. 2 below launches the model in the second part of the figure where we can see green nodes representing the environment getting connected by red segments to diffuse information.

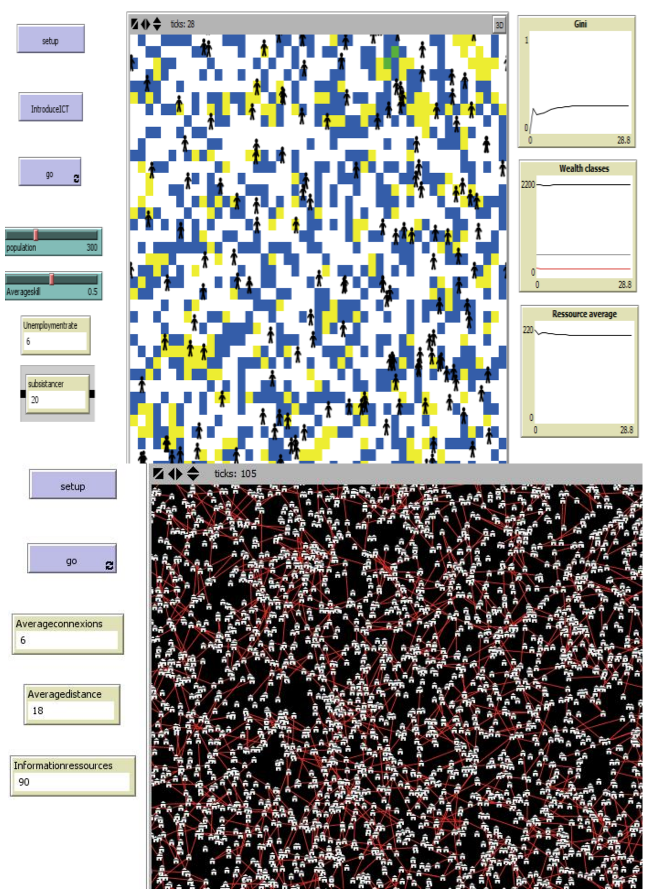


Fig. 2. Information Society Model Connected with Economy Model.

IV. The Model

A. Agents and Actions

This model is based on classical agent-based models of wealth distribution with individuals having basic economic interactions with each other and with the environment.

We consider however that the skill of an agent can be improved by a resource of type information. We list the properties of the agents below.

Resource agent: (represented by the patch)

- vision : represents how the radius of accessible resources
- sellable: Boolean variable that represents whether the resource can be sold or not
- resource type: either material or information or subsistence
- value: a quantification of the value of the resource

Individual agent:

- wealth: variable that represents the global wealth of the individual
- skill: skill of the individual which determines the ability to extract a resource
- material: quantity of material resources
- subsistence: quantity of subsistence resources

In this model according to the flow chart in Fig. 3 agents harvest the maximum resources that their location allows them to, giving priority to subsistence resources. An excess of resources allows the individuals to do other operations such as employing starving agents to harvest for them or selling and transforming resources. We can note that the execution of this model after significant steps shows an increase in Gini index and a relative stabilization in the long run which is consistent with other related models [2]

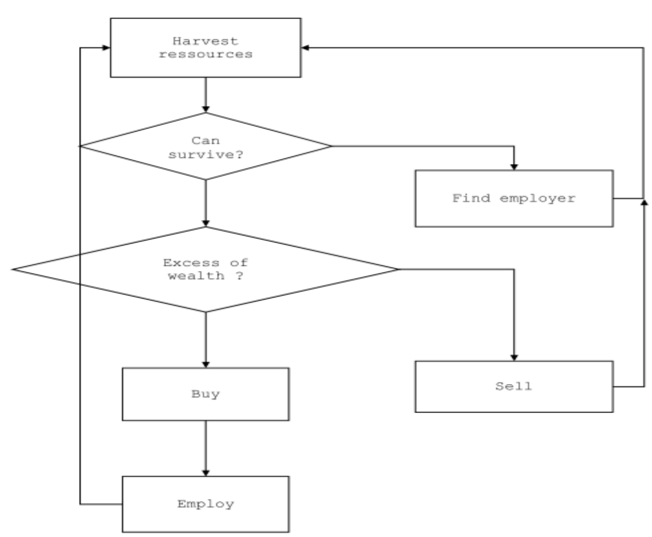


Fig. 3. Flow Chart of the Economy Model.

Parallel to this model is the information society model based of diffusion in a network [5]. In this case, it simulates how information can be more accessible when resource nodes connect to each other. The model has one node agent. In each step arbitrary nodes connect with each other and affect the vision of other nodes proportionally to numbers of continuously connected nodes and proximity of material resources.

Below is an algorithm of the procedure "updateinfo" that updates the properties of node in each execution step. Nodes with more than 5 neighbors pointing to material resources become sellable and the vision is incremented while traversing the connected neighbors that are connected to more than 10 nodes. We refer in the following algorithm to the resource who are considered connected to a parent resource as "link-neighbors" and to select a resource based on the maximality of a criteria by max-one-of.

```
to updeinfo
  let a nobody
  let x count link-neighbors with [resource type= "material"]
  if x > 5  [set sellable true]
  let xx max-one-of link-neighbors [count link-neighbors]
  if xx != nobody [ let y count [link-neighbors] of xx
  while [ y > 2] [  ask xx [set a max-one-of link-neighbors  [count
link-neighbors] ifelse a != nobody  [ set y count [link-neighbors]
of a  ] [set y 0]    set xx a]
    ifelse visiondist < 10 [set visiondist visiondist + 1]   [set y
0]] ]
end
```

This algorithm is not based on a graph structure but its complexity is not exponential due to the Netlogo graph methods implemented in the link agent.

B. Computational Considerations

To assess the impact of information technology by connecting the environment of the second model on wealth distribution of the first model, we assimilate that the connections will link resources of the environment, and the access to resources will become transitive in a defined radius .A diffusion of information will also impact the skills of the individuals. In each step, the properties of resources are sent to the second model as properties of the nodes, a step of connection is executed then the new values of the properties affected by this connection are sent back to the economy model as shown in Fig. 4, and so on.

One step of a model can refer to multiple runs depending on the time scale proportion of the two models. For instance, the exchange and spread of information occurs with more frequency than the economic transactions, and we set the number of runs of the latter to 3. Assuming the existence of appropriate datasets, a proportion factor of communication time and volume of transactions normalized by a time /information exchange indices would be a more accurate estimation of this number of runs.

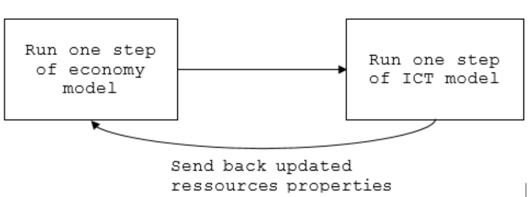


Fig. 4. Synchronization between Models.

Below is the synchronization algorithm that transfers the properties of resources to nodes in the second model to be influenced by connection.

```
to synchronize
  let i -24   let j -24 let k 0
 while [i <= 24] [ set j -24
 while [j <= 24]  [  let v [value] of ressource (i,j)
  ask "ICT Model"[ask ressources [set value v] ] set k k + 1 set j j
+ 1  ] set i i + 1 ]
 ask "ICT Model" ["execute one step"]
  set k 0 set i -24
  while [i <= 24] [ set j -24
 while [j <= 24] [  ls:let num k
   let v  [ [value] of ressource num] of "Economy model"
   ask ressource (i , j) [set value v]  set k k + 1  set j j + 1
 ] set i i + 1 ]
end
```

This algorithms loop on indices of the resources to transfer them to the other to address a limitation of Netlogo in concurrent programming, which does impact this specific model.

However in other cases, a sequential synchronization can make the model dysfunctional since the agents react to a state in the environment that was modified by another agent while they were supposed to react simultaneously to the unmodified state. The advantage of our design that is making the synchronization process part of the simulation rather than external is compatible with the transactional memory mechanism that places locks on variables mutually influenced by other agents as specified in Hlogo [12].

Nonetheless, computational efficiency issues might arise when the number of agents of transactions is high. The solution is to define "meso" entities in the model as groups of agents having the same behavior as in pedestrian simulation models , such a solution is also totally compatible with our design since the inter-model communication happens only in the environment and is not affected by a change of agents.

V. Numerical Analysis

A. Validation and Calibration

The values of the inputs of the model to conduct experiments were chosen according to different patterns suggested by different studies to ensure that the analysis is independent from input functions.

The experiments were conducted according to Table. 1 scenario:

TABLE I. Execution Scenario

	Minimum value	Step	Maximum value	Law
Population	100	20	1000	Linear
wealth	5	None	1000	Normal
Resource by individual ratio	2	0.5	10	Linear
Propagation speed	1 node /step		10 nodes /step	Time Linear
Unemployment rate	4		25	Random

Since the goal of the model is to search for macro patterns and not reproduce a real case study of a specific country or society, we opted for a level 2 validation [13] according to the constant following macro-properties in our executions:

-Existence of different wealth classes in a pyramidal structure

-Existence of economic cycles reflected in the wealth of the bottom classes

-Unemployment rate always above 4% to ensure

-Subsistence resources remaining the main form of resources by a 20% percentage.

B. Analysis

We conducted our analysis based on categories environment that launches the second model once its characteristics were detected by CART trees implemented in R statistical language which can communicate with Netlogo through the RNetlogo extension [14] . We considered the Gini index as the main measure of wealth distribution of a population. It is computed as the ratio of the cumulative wealth of each percentage of the population and the same wealth in a case of equal distribution.

We list below the categories of the our environments that were influenced by the introduction of information technology

- Non-egalitarian society: characterized by a transfer of wealth to the wealthier agents and an increase in the looking for survival pool.

We note in the plots of this case in Fig. 5 that the introduction of information technology to the environment inverses the tendency of the Gini curve for a short period then goes back to its initial shape. The same tendency can be observed in resource values where the average value increased temporarily then returned to its decreasing shape.

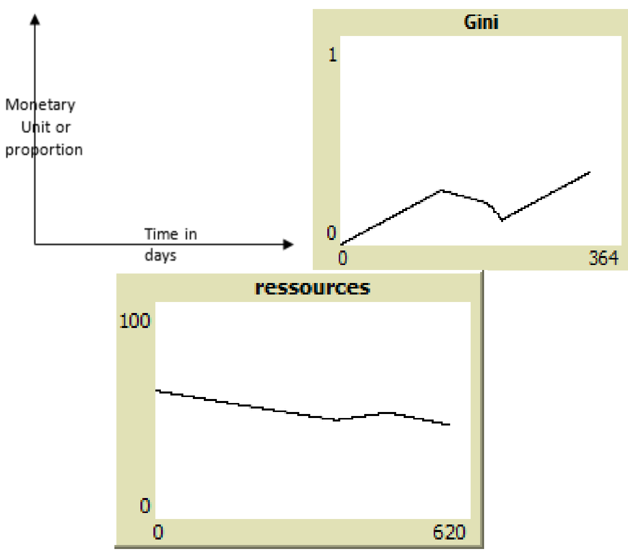


Fig. 5. Non-Egalitarian Society Plots.

- Egalitarian society: characterized by slight variations in cycles over an average of the Gini index."

In this society the introduction of information technology creates an immediate perturbation period shown in Fig. 6 then returns to its previous shape and the same thing can be said about resource values

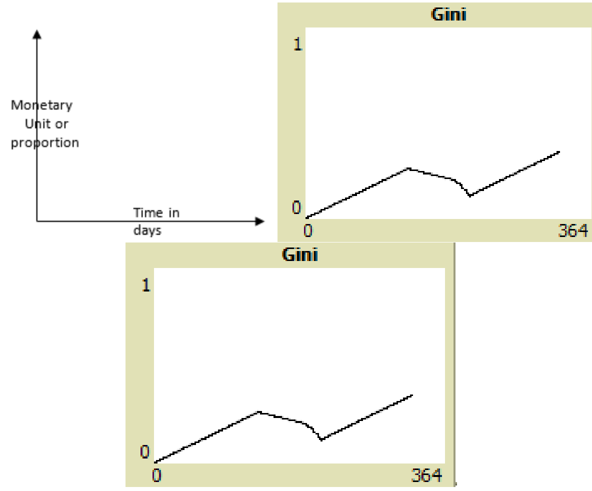


Fig. 6. Egalitarian Society Plots.

- Society with skilled individuals: characterized by a high skill average

The occurred mainly in the wealth pyramid where more than 50% of wealth became shared by more individuals. However this was not reflected in the Gini curve which suggests that the unskilled individuals got poorer.

- Society with abundance of material resources:

We noted in this category of society a tendency to more wealth repartition in the bottom of the pyramid but not in Gini curves which suggests that the richest individuals got even richer (Fig. 7).

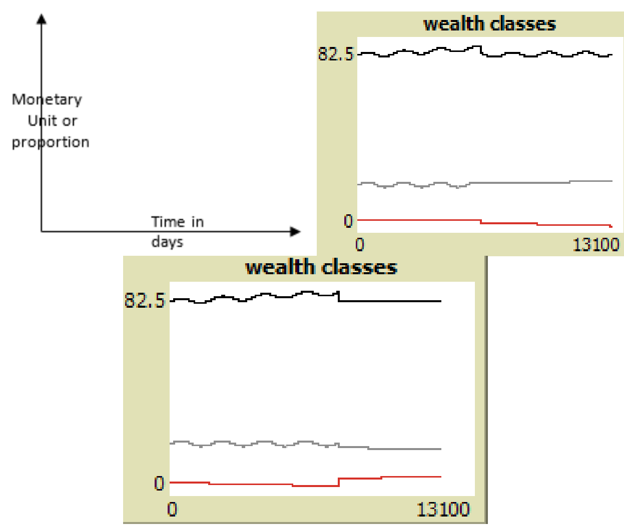


Fig. 7. Wealth Classes Plots.

VI. CONCLUSION

We can assert from the previous analysis that our simulation suggests that information technology has a negligible impact on wealth distribution. This is compatible with the fact that no conclusion can be drawn from empirical evidence where some case studies suggest an increase in inequalities while others suggest a decrease [4]. Undoubtedly, information society offers new opportunities for the less wealthy, but it also offer means to the wealthier for better exploitation of resources and a wider access to labor as suggested by reporters in our model. The stabilization of Gini curve is an emergent phenomena of our model and supports the theory that information technology in the long run contributes to the stabilization of wealth distribution which is still a positive effect. In comparison with models that have one level, our model being composed of two simple components allows change in any of the models without affecting many variables, it also encapsulates the non-relevant variables for our experiments and allows us to explore flexibly different possibilities of modern technology influence with a significant gain in computation costs.

REFERENCES

[1] *Naciri, Nawfal & Tkiouat, Mohamed.* Economic agent based models: Review. International Journal of Applied Engineering Research. 11. 5492-5502. 2016

[2] *Damaceanu, R.C.*. An Agent-based Computational Study of Wealth Distribution in Function of Technological Progress Using Netlogo. American Journal of Economics . 2018

[3] *Sanchari Goswami,* Agent based models for wealth distribution with preference in interaction ,Parongama Sen Physica A: Statistical Mechanics and its Applications 415 · August 2014 .

[4] *Davies, Jim & Fortin, Nicole & Lemieux, Thomas*. Wealth Inequality: Theory, Measurement, and Decomposition. Canadian Journal of Economics/Revue canadienne d'économique. 10.1111/caje.12313. 2017

[5] *D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins*, "Information diffusion through blogspace." In Proceedings of the 13th international conference on World Wide Web, pp. 491-501. ACM, 2004.

[6] *Naciri, Nawfal & Tkiouat, Mohamed*. Agent-based platforms comparison. 10. 34224-34229. 2015

[7] *Gil-Quijano, J., Louail, T., and Hutzler, G*. From biological to urban cells: Lessons from three multilevel agent-based models. In Desai, N., Liu, A., and Winikoff, M., editors, Principles and Practice of Multi-Agent Systems, volume 7057 of Lecture Notes in Computer Science, pages 620– 635. Springer. 2008

[8] *Soyez, J.-B., Morvan, G., Merzouki, R., Dupont, D., and Kubiak, P*. Multi-agent multi-level modeling – a methodology to simulate complex systems. In Proceedings of the 23rd European Modeling & Simulation Symposium. 2011

[9] *Hjorth, A., Head, B., & Wilensky, U*. LevelSpace NetLogo Extension. http://ccl.northwestern.edu/levelspace. Center for Connected Learning and Computer-Based Learning. Evanston, IL. Retrieved from www.github.com/NetLogo/LevelSpace .2015

[10] *Sanghamitra Bandyopadhyay* ,Are Mass Media and ICTs associated with Inequality and Poverty? Queen Mary University of London, Phd Thesis 2015.

[11] *Deissenberg, Christophe, Vanderhoog, Sander, & Dawid, Herbert*. EURACE: A massively parallel agent-based model of the European economy. Applied Mathematics and Computation, 204(2), 541–552. 2008

[12] *Bezirgiannis, Nikolaos & Prasetya, Wishnu & Sakellariou, Ilias*. HLogo: A Haskell STM- Based Parallel Variant of NetLogo. Advances in Intelligent Systems and Computing · July 201897- 119. 10.1007/978-3-319-69832-8_7. 2018

[13] *AXTELL, R.L. & Epstein, J. M.* . “Agent-Based Modeling: Understanding Our Creations”. Bulletin of the Santa Fe Institute. .page: 28–32, Winter 1994

[14] Thiele, J.R Marries NetLogo: Introduction to the RNetLogo Package. Journal of Statistical Software, 58(2), 1 - 41. doi:http://dx.doi.org/10.18637/jss.v058.i02 .2014

[15] *Vaubourg, Julien & Presse, Yannick & Camus, Benjamin & Bourjot, Christine & Ciarletta, Laurent & Chevrier, Vincent & Tavella, jean-philippe & Morais, Hugo & Deneuville, Boris & Chilard, O*.. Smart grids simulation with MECSYCO. Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). 9086. 10.1007/978-3-319-18944-4-37.2015

Data Flows Management and Control in Computer Networks

Ahmad AbdulQadir AlRababah
Faculty of Computing and Information Technology in Rabigh, Rabigh 21911, KSA.
King Abdulaziz University

***Abstract*—In computer networks, loss of data packets is inevitable, because of the buffer memory overflow of at least one of the nodes located on the path from the source to the receiver, including the latter. Such losses associated with overflows are hereinafter referred to as congestion of network nodes. There are many ways to prevent and eliminate overloads; these methods, in the majority, are based on the management of data flows. A special place is taken by the maintenance of packages, taking into account their priorities. The ideas of these solutions are quite simple for their implementation in the development of appropriate software and hardware for telecommunication devices. The article considers a number of original solutions to these problems at a level sufficient for the development of new generations of telecommunication devices and systems such as allowing interrupting transmission of the low-priority packet practically at any stage, then to transmit a high-priority packet and only then resume the interrupted transfer, moreover warning in time the data source about the threat of overloading one or several nodes along the route in the propagation of data packets.**

Keywords*—Data transmission; data stream; input output buffers; telecommunication devices; data packets; blocks of memory; switching matrix; high priority packets; bitstaffing

I. Introduction

Controlling data flows is one of the functions of the network layer, including load management and anti-blocking. There are several levels of management. Inter-node management is associated with the allocation of buffer memory at intermediate nodes (allocation to each direction of a certain number of buffers), which reduces to the restriction of the channel queues lengths [5]. Control "input-output" is aimed at preventing locks. It is realized by indicating the length of the message in the first packet, which allows the receiving node to predict the memory filling and to prohibit receiving certain messages datagrams, if a memory lock is predicted.

In the general case, the communication network is a distributed communication system serving to transmit information at a distance. These include television and radio broadcasting networks, telephone and cellular communication networks, cable television networks, etc. Synonym for communication - data transmission. The concept of telecommunications network implies a territorially distributed data transmission network.

A separate computer is an example of a centralized computer system. Unlike the centralized, the computer network is a distributed computing system. This is a combination of computer and communication equipment, communication channels and special software that manages the process of distributed computing between members of a given network.

Since the role of transferring non-numeric information via computer networks has recently increased, the term data network is now often used for them. To avoid confusion with the communication network, in which data is also transmitted, the term computer network is used. a program queue management scheme in computer networks, in which the scheduler serves queues with a higher priority to the detriment of low priority queues [1].

The maximum possible number of queues is four: High, Medium, Normal, and Low. The scheduler starts the service from the higher priority queue High. If there are no more waiting packets in its queue, it goes to the next less priority queue in which there are waiting packets. After each servicing of the Medium, Normal or Low queues, the scheduler returns to the High queue, that is, the process repeats. A low-priority Low queue is served only when there are no waiting packets in the High, Medium, Normal queues [2]. PQ planner has some obvious advantages and disadvantages. Packages from the high-priority High queue can claim 100% bandwidth with low latency and latency, while low-priority traffic is served much longer.

The management of external threads (access) is realized by giving priority to the transfer to internal threads before external ones, limiting the number of packets in the network (the packet is received if the node has the appropriate permission), sending warning packets-stubs to the source address, from which the packets go to the congested line connection [3].

II. Research Methods and Related Works

Method 1: Control the flow of data adjusting the length of pauses between packets

In this method, during the data transfer, the receiver notices a persistent shortage of arriving packets (for example, by tracking their sequence numbers) and sends a control packet containing the XOFF command to suspend the data stream to the data source A. The address of the data source is known to the receiver, since the data packets coming to it contain information about the addresses (or directly addresses) of devices A and B. Sending requests for retransmission of lost packets is also sent the node M2 is overloaded, its input buffer memory (in the following, for brevity, the input buffer) is completely or almost completely filled with incoming data packets [4]. New packages, at least some of them, are lost due to lack of free space in the buffer.

When the XOFF command is received, the data source completely stops sending packets and resumes it, either after some time specified in the data exchange protocol [5], or after receiving the renewal of transmission from the XON command receiver.

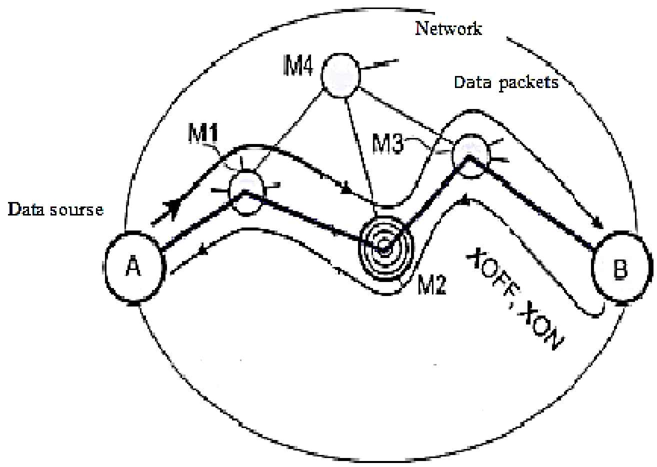


Fig. 1. Traditional Way to Control the Flow of Data.

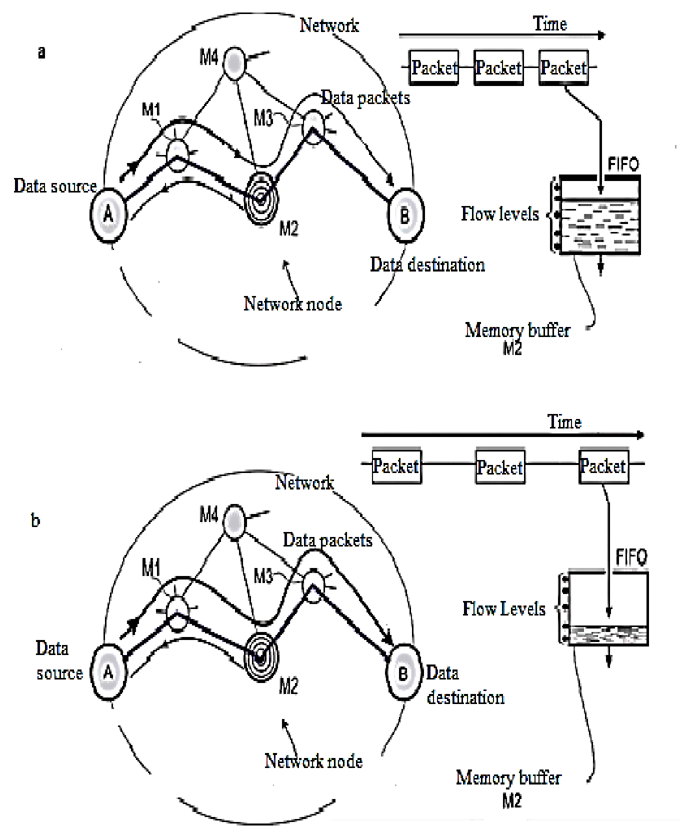


Fig. 2. Improved way to control the flow of data: a - if there is a danger of overflow of the buffer memory of the node M2; b - during normal operation.

The proposed solution (Figure 2) largely eliminates these shortcomings due to a smooth and "advanced event" adjustment of the data transmission rate by the source. The speed is controlled by changing the length of pauses between packets: the longer the pause, the lower the data transfer rate, and vice versa [3], [6]. Note that the presence of a pause does not mean that there is no signal in the communication line - the signal is present constantly, but there are no flag codes indicating the beginning of the packet, or vice versa - a continuous stream of these codes is transmitted.

In the Fig. 2, and the pause situations between packets transmitted on the route A-B are relatively small, or in other words [3], the data rate of the data placed in the packets is relatively large, in the sense that the buffer memory level of the intermediate node M2 is steadily increasing, which may result in buffer overflow [7]. Buffer memory for clarity is shown in the figure as a tank with liquid replenished by the input stream of packets, while the output stream tends to reduce the level of its filling.

In this case, the node M2 registers the operation of the second upper level sensor (the comparator of read and write addresses of the buffer memory block) [9]. This means that the level of filling is close to critical, therefore, it is necessary to reduce the rate of data flow to the buffer. To reduce the speed, the node M2 sends to the node A a service packet, a command to increase the pauses between packets.

In response to this command, node A increases the duration of pauses between packets (Figure 2, b) [5], [8]. The degree of increase can be stipulated in the protocol of data exchange between nodes of the network or in the explicit form indicated in the service package. After increasing the pauses, the buffer memory level of the M2 node starts to decrease, if there is no other reason for its increase[8]. Upon reaching the central or lower mark, the node M2 sends to the node A the command to reduce the duration of the pauses, the level of the buffer filling again begins to increase, etc.

Thus, in an ideal case, the buffer memory of node M2 does not overflow and does not emptied, the speed of data output from the buffer memory remains constant [10], the rate of data arrival adapts to it, making slow fluctuations inherent in conventional automatic control systems.

If there are several data sources, then to prevent overload the work of the most active one, but not the most priority, is slowed down; if the sources are equally active, then the impact on those with low priorities is primarily affected [11].

In the development of the described method, it is proposed to take into account, not only the level of the buffer completion, but also the dynamics of its change when forming commands for decreasing or increasing the intensity of the flow [12]. This allows eliminating unnecessary flow control commands when the buffer fill level is high, but the history of the process is such that there is a steady tendency to stop its growth and the subsequent decrease (and vice versa) [13]. Essentially, along with absolute reference, the rate of change in the rate (acceleration) of the motion of the level of buffer memory filling is considered.

Method 2: Managing the flow of data by notifying the packet source of causes of overload

Let's continue our consideration of the known methods of data flow control (Fig. 3) using the same network model as before (Figure 1, 2). The data source A transmits a series of data packets to the receiver B. In response to each packet or to a group of packets, the receiver B sends the ACK or NACK response packets to the source A. The ACK response acknowledges the successful reception; the NACK response is a request to retransmit a single packet or group of packets [10]. In principle, even such a simple feedback (using ACK or NACK response packets) allows detecting and eliminating network congestion on the A TO B path [2]. Indeed, if the data source is increasing the packet rate or at some fixed rate starts to receive an excessive number of retransmission requests, then, most likely, at least one of the nodes of the route entered the overload mode.

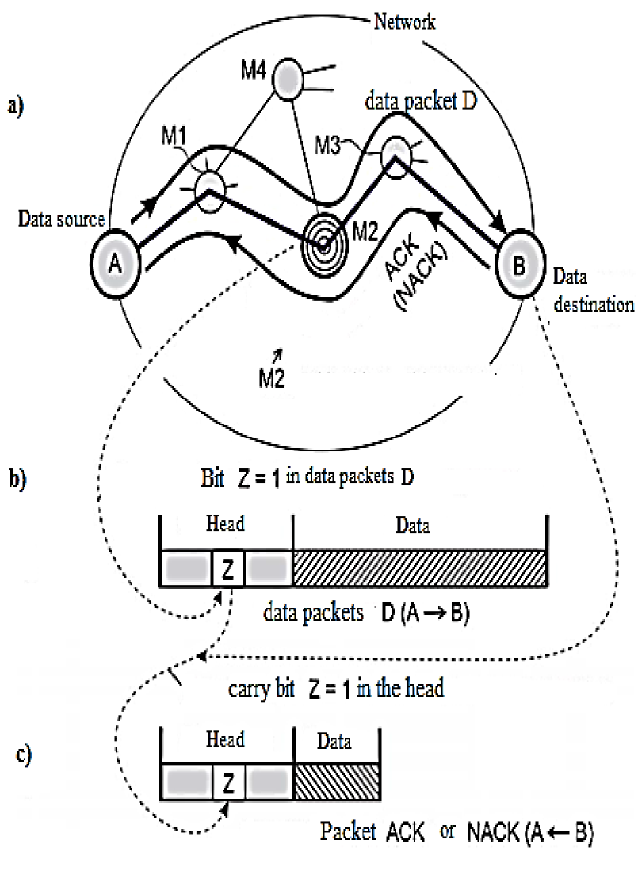


Fig. 3. Informing the data source A about the upcoming or available overload of the input buffer of the M2 node of the network: a - packet propagation paths b, c - packet structure D and ACK (NACK).

In this case, the data source drastically reduces the packet transmission rate or (and) increases their length to reduce the share of the overhead bits that make up the headers in the data stream[15]. In the future, the data source gradually either by random trial and error increases the data transmission speed, moving to the permissible upper limit, taking into account some permitted speed increase margin. Such a method is called a "slow start."

Of course, packet loss is possible, not related to the overload of network nodes, for example, due to uncorrectable errors caused by interference in the communication line, but in this case we are not interested in such losses [16]. The considered method of data flow control does not prevent the forthcoming loss of packets, but allows reacting only to the accomplished fact of overloading of the intermediate node of the network or the data receiver [14], [17]. This is its main defect.

The other main idea in this method is to warn in time the data source A about the threat of overloading one or several nodes along the route A in the propagation of data packets D. This warning is the bit Z included in the header of the ACK or NACK response packet [18] (Figure 3, c).

In the example shown in Fig. 3, the processor of node M2 anticipates overload, observing the steadily increasing level of buffer filling, as it was shown on its model, shown in the right part of Fig. 2, a; (other events are possible, such as predecessors of congestion.

In packages passing through the node M2, more precisely, in the header of each of them, there is information sufficient for its routing, for example, in the form of IP addresses of the source and the data receiver [19]. Viewing this information allows the M2 node to identify the "culprit" of the expected overload, from which the most intensive flow of packets originates.

Suppose that the main "culprit" of the impending congestion is the data source A. This source, like all others, transmitting data packets D, sets the Z bits to zero. With normal data transmission on the route A TO B, these bits remain in the zero state.

If the conditions for the upcoming overload are detected and knowing that the largest number of packets per unit of time originate from the source A, the node M2, when transmitted along the route A TO B, marks all packets or a part of them with their Z = 1 bits that inserts in the headers, as shown in Fig. 3, b. The data receiver B returns the received Z = 1 bits to source A, including them in the headers of the response packets ACK and NACK (Figure 3, c).

Finally, data source A receives bits Z = 1 and sharply reduces the data transfer rate to node M2[21]. Further, the data source A gradually restores the original data flow parameters [20] or even exceeds the previously reached data transmission rate until a new series of bits Z = 1, etc., is detected (here, too, the "slow start" mentioned earlier is applied) [22]. Having determined the allowable upper speed limit, the data source takes a small step down to create some margin, guaranteeing the route from overload.

This way of preventing or eliminating overloads is satisfactory, but its disadvantage is that, without knowing the reason for the overload of node M2, the source of data A is unable to adequately respond to it. So a sudden and sharp decrease in the data transfer rate - is unacceptable for many applications [23]. But if, for example, the data source A knew that the reason for the upcoming overload was that the processor of the M2 node could not cope with header stream processing, then it could, without reducing the transfer rate of payload data, increase the packet length to reduce the intensity of this flow.

The problem solved by the method 2 discussed below is thus not only to prevent the source of data on the impending overload, but also to inform him of its cause. Then the source could choose the most appropriate "line of behavior" in this situation[24]. The problem is solved by extending the single-digit sign Z to several bits. Let us explain what has been said by example, accepting some assumptions.

Suppose that route A-B (Figure 3) is a virtual telephone link between devices A and B, for example, between computers or IP telephones. The technology of VoIP (Voice over IP) is used [26]. Devices A and B contain codecs such as AMR (adaptive multi rate) [25]. The codec generates compressed speech fragments every 20 msec and encodes data from one of eight speeds in the range from 4.75 to 12.2 kbps. Further, as before, one-way data transfer from device A to device B is considered.

After the connection A-B is established, the data source generates packets, each of which contains a header and a data field. The data field of the packet is filled with fragments of speech from the codec output, and then the packet is sent along the communication line to node M2 [27]. The codec, if the bandwidth of the A-B channel allows, is initially set to the maximum coding rate to ensure the highest speech intelligibility recovered from the data input to receiver B. The Z bits of the sent packets are set to zero.

In the event of detection of the danger of overload by some node located along the A TO B route, this node (in our example, the M2 node) inserts some indication Z in the headers of packets originating from the most active source (A), as described earlier, taking into account that this feature contains not one, but at least two bits. This attribute is returned to the source; as a result, the processor of node A receives information about the reason for the upcoming overload. The node M2 may experience overloading for at least one of the following reasons.

1) Narrowing the bandwidth of the channel A TO B due to the appearance of a "bottleneck." This can happen, for example, because a part of the dedicated link A TO B of the linkage between the nodes M2 and M3 (Figure 3) has decreased. This decrease may be due to various reasons. Let's name two of them.

- The previously unobtrusive competing data flow along the route M4 M2 M3, which uses the same channel M2 to M3, as the route A TO B, has increased to a significant level earlier. As a result, the M2 node redistributed the strip of this channel to the detriment of the route A TO B.
- The M2 node has changed the type of signal modulation in the M2 to M3 channel, reducing the transmission rate due to the deterioration of the signal-to-noise ratio in this channel.

2) The M2 node processor for some reason or other has stopped coping with the volume of work on analyzing packet headers following the route A TO B.

The first and second reasons above for the approaching overload are displayed respectively by the codes Z = 012 and Z = 102, the absence of an overload hazard corresponds to the code Z = 002, both causes simultaneously generate the code Z = 112. The code Z = 112 can Form one node if it simultaneously observes both reasons for the upcoming overload, or by two or more nodes located along the A to V.

So, the node M2 can insert the Z = 102 codes into the headers of the A B packets that pass along the route, because the processor of this node cannot cope with the volume of work on the analysis of headers. These packets are transmitted to the M3 node, which is supposed to reveal a decrease in the M3 to B channel bandwidth allocated to the A to B route. In this case, the M3 node replaces the Z = 102 codes in the packets passing through it with Z = 112. These codes, as described, reach the receiver B and return to the data source A as part of the headers of the response packets (Figure 3, c).

The optimal response of the data source to the identified causes (1 or 2) of overloading may be this.

The narrowing of the channel bandwidth A to B (reason 1) should cause a corresponding decrease in the total data rate (both useful and service) of the source A. To estimate the rate reduction, it would be desirable to use a multi-bit code Z in which this degree is reflected. However, in this case there is no such possibility, therefore the processor of the data source A switches its codec to the mode of the lowest encoding speed (out of eight possible - from 4.75 to 12.2 kbps). If the packet length is unchanged, and the lowest

The frequency of their succession decreases due to the increase in pauses between them. At the same time, the delay in the formation of the packet increases due to the increase in the time it is filled with compressed fragments of speech. Thus, the data transfer rate (both useful and service data) is reduced by source A, and if the narrowing of the band is not too large, then there is no danger of overloads. To restore the high quality of voice transmission, the coding rate and, correspondingly, the packet repetition rate gradually increase to the experimentally detected limit, in which there is still no danger of overloading the network nodes on the A to B

Alternative response of the data source to the narrowing of the channel band A to B also provides for using the lowest encoding rate. When this keeps the packet repetition rate and their length decreases. The rate of transfer of useful data decreases, the service data flow remains unchanged.

Finally, the strongest reaction is possible, at which the coding rate is set to the minimum, and the length of the packets increases to such an extent that their average delay approaches the permissible limit (not more than 100 ms) [3], after which, during a telephone conversation, begins eavesdropped. Such a reaction is the maximum that can be done in this situation. After exiting the crisis, the coding rate gradually increases, and the length of the packets decreases with this in time (to reduce the delay of their transmission along the route A to B). This process of two-dimensional optimization of flow parameters is completed when the boundary is reached, after which the risk of overloading again arises.

Overloading the processor of one or more nodes on the A to B route (reason 2) is eliminated by reducing the intensity of the header stream that it (they) has to process. For this, while maintaining a high coding rate, the data source increases the length of the transmitted packets to such an extent that their average propagation delay along the A to B path does not exceed the previously mentioned allowable limit (100 ms). Thus, the correct response to the overload warning in many situations allows to eliminate the danger of overflow of input buffers and, what is essential, to maintain high quality of voice transmission.

Method 3: Control of the flow of data with compensation of the inertia of the feedback loop

One of the simplest ways to control the flow of data transmitted between the nodes of the network J1 and J2 (Figure 4, a) is as follows.

In steady state, data packets are accumulated in the output buffer of node J1 for transmission along a certain route, possibly through other network nodes (not shown in the figure) to the input buffer of node J2. Both buffers are executed in the form of blocks of memory of type FIFO.

The flow of data packets passing through the system from the left to the right has the character of "machine-gun queuing", since the series of packets are transmitted by the J1 node via the communication line only with the permission of the receiver, node J2, which "causes fire to the extent possible". The instantaneous packet transfer rate inside the series is C; the average speed is less than the instantaneous one and depends on the average ratio of the pauses between packets to the length of the series. The unevenness of the arrival of packets in the buffers of the nodes J1 and J2 causes fluctuations in the levels of their filling. The challenge is to protect these buffers from overflow or emptying.

Further, this task is solved only with respect to the input buffer of the node J2, however, the output buffer of the node J1 can be protected in a similar way by introducing feedback from the source of the packets sent to it (in the figure this source and its feedback are not shown). Such a successive chain with feedbacks between neighboring elements can be arbitrarily long. Each transmitting port thus issues a stream of packets to the communication line only if there is a transmission permission previously received from the destination of the XON command.

The input buffer of node J2 contains a pointer to the threshold level F of its filling. In this example, the input buffer of node J2 contains Q packets. At the moment the current level Q overcomes the threshold level filling in the F upward side (Q > F), the node J2 transfers to J1 the packet with the XOFF command of the transmission suspension. Similarly, at the moment overcoming the current level Q filling the threshold level F downwards (Q<= F), the J2 node sends a packet to the J1 node with the XON resumption command.

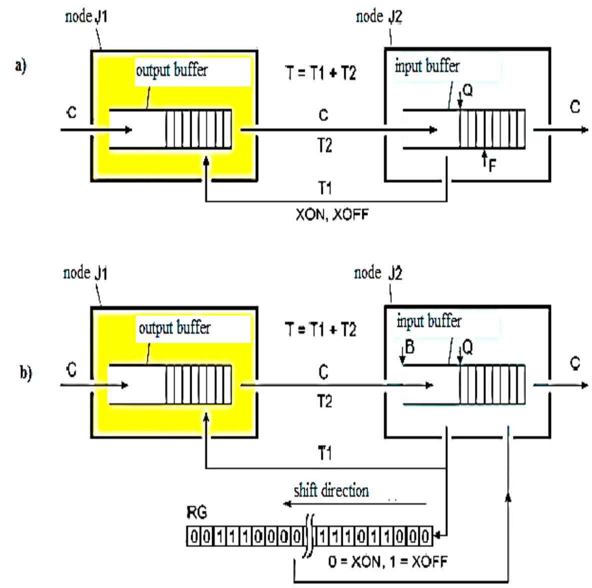


Fig. 4. Flow control scheme: a - traditional; b - the proposed.

The problem is that flow control can be very inertial.The response time of the system to the XON and XOFF commands is determined by the delay T = T1 + T2, where

T1 - the time from the instant the command is generated by the node J2 until the previously stopped process of sending packets by the node J1 resumes or the previously activated process of issuing packets by the node J1 is suspended;

T2 - the time of packet transmission from the output buffer of node J1 to the input buffer of node J2.

Thus, if the increasing filling level of the input buffer of the node J2 has overcome the threshold value F, then the generated XOFF command will stop the flow of packets at the input of the node J2 only through the time T. During this time, the input buffer of the node J2 continues "by inertia "To replenish. Similarly, the first packet after issuing the XON command to resume the previously-stopped stream will arrive at the input buffer no earlier than the time T. During this time, the level of filling the input buffer of the node J2 "by inertia" is reduced due to the outflow of data from it.

If the capacity of the input buffer of node J2 is small, then the inertia of the control can lead to overflow or emptying. In the worst case, after the moment of exceeding the threshold level F (Q > F, the command XOFF is issued) and at the time no outflow of data from the input buffer of the node J2 during the time T in this buffer "by inertia" will come with C*T packets. Similarly, if there was no inflow of data, after the moment of crossing the threshold level F in the direction of decrease (Q < = F, the command XON is issued) and with continuous data flow from the input buffer of node J2 during the time T from this buffer "on inertia "will be selected C*T packets. Thus, to protect against overflow and emptying, the input buffer of node J2 should be designed to store at least 2C*T packets; the threshold F must correspond to its middle.

The resulting estimate of the minimum buffer size is disappointing. Some switches contain several hundred buffers, In high-speed networks, the T value reaches tens and hundreds of microseconds. The value of C is of the order of 10 Gbit / s. As a result, the buffer size $2C \times T = 2 \times 10^{10} \times 10^{-4}$ is several megabits. The goal of the next solution is to reduce the buffer size by half thanks to smoother flow control.

Smoothness of control is achieved by fragmentation of series of packets and more intelligent algorithm of forming commands XON and XOFF to resume and stop transmission of the stream. The circuit shown in Fig. 4, b, [4] contains the same components and has the same parameters (T, C, Q), which have just been discussed. The volume of the input buffer of the node J2 is denoted by B. The new element of this node - the history memory of the control - is shown for clarity in the form of a shift register RG, although it can be executed programmatically using a set of memory cells.

For definiteness, suppose that the flow of ATM cells is transmitted via the communication channel [5]. (The term "cell" is equivalent to the term "packet".) This stream is continuous - after the last bit of the previous cell, the first bit of the next is transmitted. The length of the cell is 53 bytes. The cells follow the line of communication with a period of 40 ns. This does not mean that the proposed idea is applicable only to ATM technology - it is easy in the following description to operate with strictly prescribed quanta of time with duration of 40 ns.

Suspension of the flow in this case is conditional (a continuous stream of cells follows the connection line always) and means that the output of the nodes J1 accumulated in the output buffer really stops, but instead of them, bypassing this buffer, empty cells of the same length are output into the communication line, as well as cells with data. Empty cells can be inserted once or form more or less lengthy sequences. Blank cells are rejected by the J2 node and do not enter its input buffer.

Suppose that the time T = T1 + T2 = 2 μs, that is, corresponds to the passage of 50 cells. The rate of issuing commands XON or XOFF is equal to the rate of arrival of cells (empty and non-empty) at the input of node J2, that is, commands are issued every 40 ns. The commands issued by the node J2 in response to each incoming cell on the communication line affect the input stream after a time of 50 cells - this is the inertia of the control loop.

Simultaneously with issuing the XON or XOFF command from node J2 to node J1, it is stored as the corresponding bit (0 or 1) in the right-hand bit of the shift register RG, the remaining bits are shifted one position to the left, the leftmost bit is pushed out of the register. Thus, in the RG register, the history of issuing control commands for the next 50 cycles (the periods of succession of the cells) is displayed. Each XON or XOFF command when entering J1 is responsible for making a decision to issue one (regular) cell either from the output buffer of this node (when receiving the XON command) or from a source of empty cells to bypass the output buffer (upon receipt command XOFF).

The code in the RG register is analyzed by the J2 node. Counting the number of zeroes contained in it, the node predicts the number of cells with data that will go to its input buffer within the next 50 cycles. The single bits in this register correspond to the number of empty cells that will arrive at the input of node J2 during this period and will be destroyed by them. The formation of XON or XOFF commands is as follows. Let NON be the number of zero bits in the RG register, B the size of the input buffer of the node J2, Q the current size of the queue. Then: if Q + NON □□B, then the XOFF command is generated; otherwise, the XON command.

Indeed, in the worst case, when there is no outflow of data from the input buffer of the node J2, the expected level of its filling is equal to the current level of Q, increased by the number of NON cells that are actually already in transit and will surely be received in the next 50 cycles. The expected level of buffer filling Q + NON should not exceed its size B. If this condition is met, then the thread should not be suspended, so the XON command is generated. In the opposite situation, when the predicted level of buffer filling exceeds the volume of the buffer, stop the flow for at least one clock cycle, that is, generate the XOFF command. The commands, of course, will have an effect only after 50 clock cycles, but due to the "smallness" of their action and the integration of many commands in time, the total effect is expressed in that the fluctuations in the buffer fill level become smaller, and the necessary buffer memory capacity is reduced by two times.

So, in the steady state, the average level of buffer memory of the J2 node is close to V / 2, in the 50-bit RG register the average number of zeros and ones is approximately the same. Suppose that B = 50, the average level is 25. Then the stocks in relation to overflow and emptying the buffer will be 25 cells in each direction. This is consistent with the fact that the average number of arriving cells expected in the nearest time interval T is 50/2 = 25.

In the prototype (Figure 4, a), in the worst case (in the absence of data outflow from the buffer), at the time T, 50 cells arrive at the input of the buffer of node J2. Similarly, in the opposite situation, in the absence of data flow to the buffer, the level of its filling during the time T will decrease by 50 cells. Therefore, to create the necessary reserves of 50 cells in each direction, a buffer with a volume of 100 cells is needed, which is twice as large as when using method 3 (Figure 4, b).

As expanding the scope of the method 3, the idea of reducing the amount of buffer memory of the receiver when building a data transfer system between nodes of a computer network was considered. However, this idea can find wider application. As an example, consider the circuit of the commutator (Fig. 5). As usual, to simplify the description, we assume that the data streams propagate only in one direction - from left to right. In fact, to construct a switch operating with flows of both directions, it is necessary to apply the same circuit deployed in the opposite direction, superimpose the resulting circuit to the original one, and combine the corresponding external inputs with the outputs.

The switch contains three input buffers # 1 - # 3, a switching matrix, a processor and three output buffers ## 1 - ## 3. Comparing Fig. 5 with Fig. 4b, one can note the similarity between the block structures used in both schemes. Some designations also coincide, therefore further are not explained. The signals GO_1 - GO_3 from the rightmost cell of the corresponding input buffer of type FIFO are given a data packet, with the queue moving one position to the right. Data packets from independent sources, for example, from computer network nodes, enter the input buffers of the switch. As a result, buffers create queues of packets waiting to be sent to the output buffers. The directions of packet transmission are detected by the processor based on the analysis of address information contained in their headers.

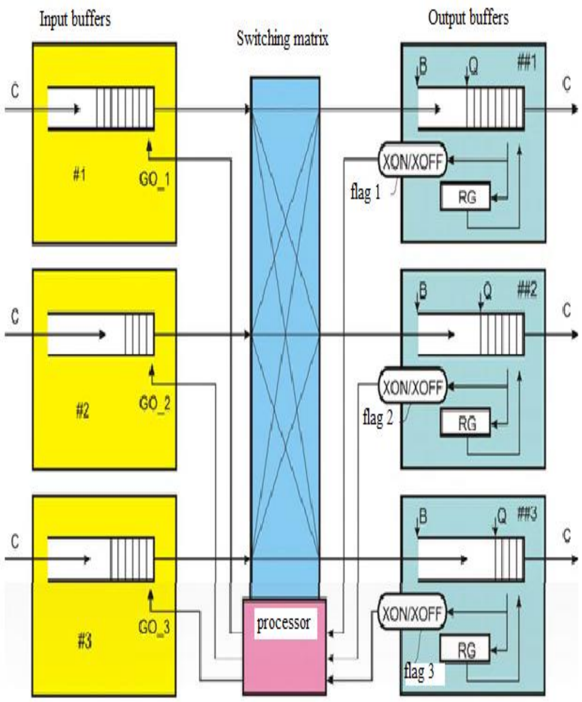


Fig. 5. Structure of the switch, the first option.

The packets are transferred from the input buffers to the output through the switching matrix under the control of the processor. Packets of some types are sent simultaneously to all output buffers or to some subset of them. The switching matrix allows simultaneous transmission of packets in different independent directions. For example, simultaneously with the transfer of a packet from the buffer # 1 to the buffer ## 3, transmissions along the directions # 2 ## 1 and # 3 ## 2 can be carried out.

In the output buffers, queues of packets awaiting delivery to the corresponding communication lines are also created. In each of these buffers, the previously discussed method of preventing overflows and devastations of the queue is applied (Fig. 4, b). However, in this case (Figure 5), the output buffer "does not know" from which directions and in what order the data is expected to arrive, i.e., it does not have information about which input buffers and which sequence should be sent the results of the queue state forecasting - the XON or XOFF commands. Therefore, the output buffers form the XON / XOFF flag bits (flag 1-flag 3), irrespective of which input buffer will be affected. The flags are polled by the processor and used by the processor to control the transmission of data through the switching matrix.

Looking through the outputs of buffers # 1 - # 3, the processor monitors a lot of packets, ready to be sent to the buffers ## 1 - ## 3. The decision to send each of these packets is accepted by the processor only if the flag of the corresponding output buffer is set to the enabling state - XON. Then the processor creates the required path through the switching matrix and initiates the issuance of the packet by the command (signal) GO_i (i = 1, 2, 3). The structure of the switch (see Figure 5) has a drawback that is not related to the application of the proposed method for managing data flows.

If the packet type provides its transfer to a group of several output buffers, the processor does not wait for the entire group to receive data at the same time to speed up the process. It transmits copies of this package sequentially, as the output buffers that make up the group appear. In this case, until the complete distribution of the packet across the whole group of output buffers, this packet is not removed from the input buffer and therefore prevents the progress of the queue in it.

A similar situation (blocking of the input queue) can be observed when sending a normal packet addressed to only one output buffer. If the output buffer is not ready for data reception for a relatively long time, then the packet remains at the output of the input buffer, and the queue in it does not advance, but only grows with the arrival of new packets. This queue may contain packages that could be serviced, since the corresponding output buffers are ready to receive data, but they are all prevented by the priority packet waiting for maintenance and blocking access to the rest of the packets to the switching matrix.

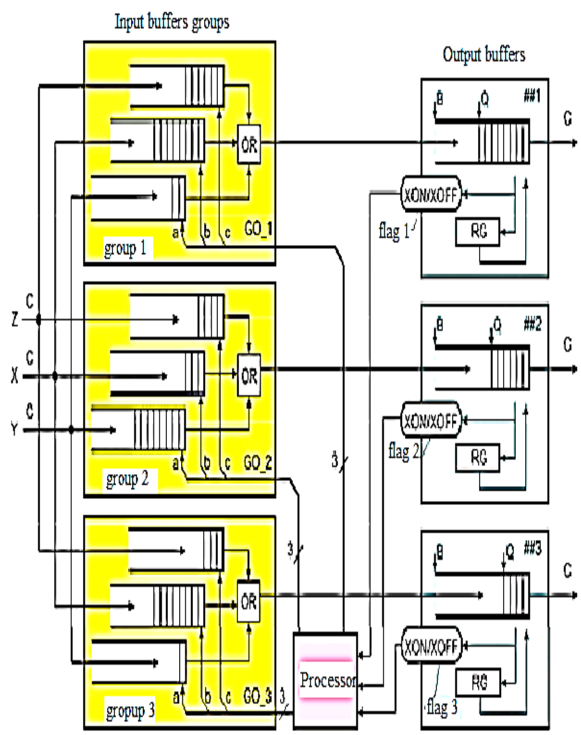


Fig. 6. Switch structure, second option.

Blocking of input queues is eliminated in the scheme shown in Fig. 6. In comparison with the previously considered circuit (Figure 5), the input buffers are replaced by buffer groups, the switching matrix is excluded. Each group of input buffers accumulates more than one queue for the number of input channels of the switch. Each group of input buffers transfers data to the corresponding output buffer.

Packets coming from the input channels Z, X and Y are sorted. Packets of channel Z, which should get into the output buffer ## 1, are written to the upper buffer of group # 1. Packets of the Z channel, intended for sending to the line through the buffer ## 2, are written to the upper buffer of group # 2. Packets of channel Z, which should be sent to the output buffer ## 3, are written to the upper buffer of group # 3. Packages from the input channels X and Y are sorted similarly.

The processor analyzes the flags 1 to 3 and, in the presence of the readiness of one or more output buffers, receives one or more GO_i signals (i = 1, 2, 3) to receive data. Each of these commands is addressed to one group of input buffers. Since in this example the group contains three buffers, the command contains three bits that indicate from which queue the next data packet should be issued via the OR gate. Commands (a, b, c) = (0, 0, 1), (a, b, c) = (0, 1, 0) and (a, b, c) = (1, 0, 0) correspond to the issuance the data packet from the upper, middle and lower case of the selected group. The queue number can be transmitted from the processor with binary code with its decoding in groups of input buffers, but this possibility is not considered to simplify the figure.

If one of the output buffers is not ready to receive data for a relatively long time, this does not affect the transmission of packet flows through other output buffers. For example, the output buffer ## 1 may not be ready to receive data (flag 1 in the XOFF state), then the GO_1 signal remains zero for this time (0, 0, 0), preventing the issuance of packets from group 1. Other groups remain in normal operating mode, i.e., as far as possible under the control of the processor, data is transferred to the corresponding output buffers.

Accelerated transmission of high priority packets through the switch. The switch shown in Fig. 7, is an improved version of the previously considered structure (Fig. 6). Comparing Fig. 6 and Fig. 7, one can note that some of the previously considered elements in Fig. 7 are not shown, although they may be present in the circuit. At the same time, new elements have been introduced, the functions of which do not violate the work of the previously considered schemes. The purpose of introducing new elements is to accelerate the transfer of high-priority packets through the switch.

Just like in the previous scheme, the switch contains three groups # 1 - # 3 input buffers of type FIFO. The outputs of these buffers in each group are connected through the first logical OR and the L1-L3 packet converters with the inputs of the output buffers ## 1 - ## 3. In each group of input buffers, the second logical OR is added, through which bypass paths (without queue) pass high-priority covenants, if they enter buffers. Switches SW1 to SW3 translate packets either from the corresponding queues located in the buffers ## 1 - ## 3, or from the workarounds. In the first case, the key is set to LP (low priority), in the second - to HP (high priority). Coordination of actions of all components of the multiplexer is performed by one or several processors (in Figure 7 processors are not shown).

As in the previous scheme (Figure 6), the packets arriving from the input channels Z, X and Y are sorted. Packets of channel Z, which are addressed to buffer ## 1, are written to the upper buffer of group # 1. Packets of channel Z, addressed to buffer ## 2, are written to the upper buffer of group # 2. Finally, the Z channel packets addressed to buffer ## 3 are written to the upper buffer of group # 3. Packages from the input channels X and Y are sorted similarly. Then the packets are moved along the corresponding input queues, through the first logical OR elements and the lower channels of the converters L1 to L3 are transmitted to the output buffers ## 1 - ## 3 and in the order of their arrival are output from them to the output lines Q, R and S via the keys SW1 to SW3, which are in the LP state.

This "natural" sequence of events is violated with the arrival of a high-priority packet, for example, in the upper buffer of group # 1. All new arrivals in the buffers, packets are checked for priority. Suppose first that the number of priority levels is two, and the high-priority packet came at a time when all other packets on the switch have low priorities. The priority level of the package is indicated in its header.

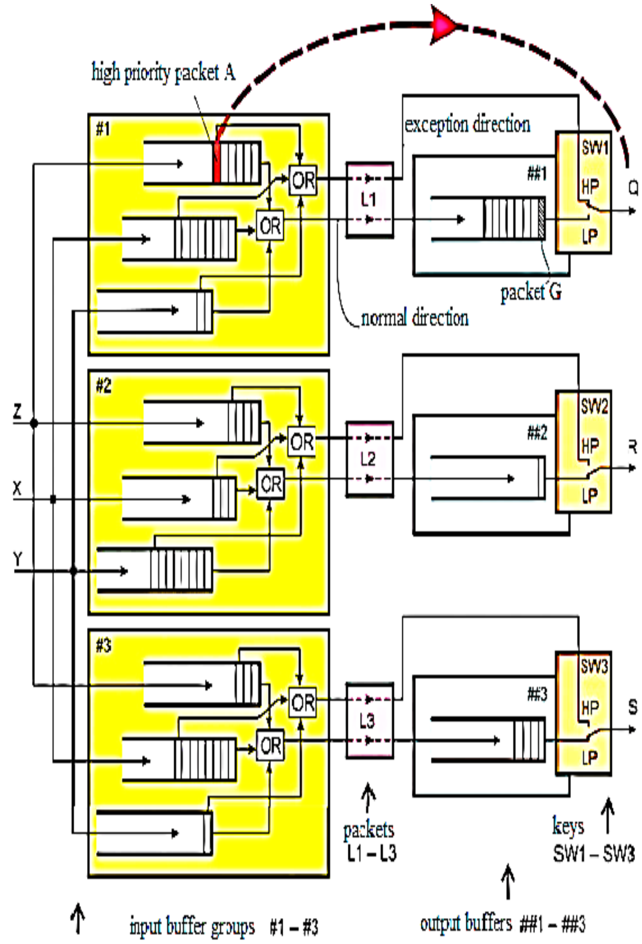


Fig. 7. Switch structure with accelerated maintenance of high-priority packets.

In the known switch structures, a simple and understandable reaction to the entry of a high-priority packet was adopted:

- If the desired output link is not used, then the high priority packet immediately, without delay, begins to be issued to it;
- If the communication line is busy transmitting a low-priority packet, the delivery of a high-priority packet is delayed until it is released.

The latter circumstance leads to delays in switching high-priority packets, which for some applications is highly undesirable or even unacceptable. In the worst case, a high-priority packet may be a little late at the time of issuing a low priority, which may have a significant length, for example, 1500 bytes.

The proposed solution allows interrupting transmission of the low-priority packet practically at any stage, then to transmit a high-priority packet and only then resume the interrupted transfer. Nested interrupts are possible if the number of priority levels exceeds two. Let us consider this solution in more detail.

Suppose that in each transmitted packet (Figure 8), in addition to the address and other information, its priority P and length N are indicated. The codes P and N can be located, for example, in two adjacent bytes, with three bits defining one of the eight priority- and the remaining 13 bits are the length of the packet (in bytes) in the range from some fixed minimum length U to the maximum, equal to (U + 213 - 1) bytes. All transmitted packets pass through converters L1-L3 (Figure 7), where each of them is bit-oriented and is preceded by a unique flag.

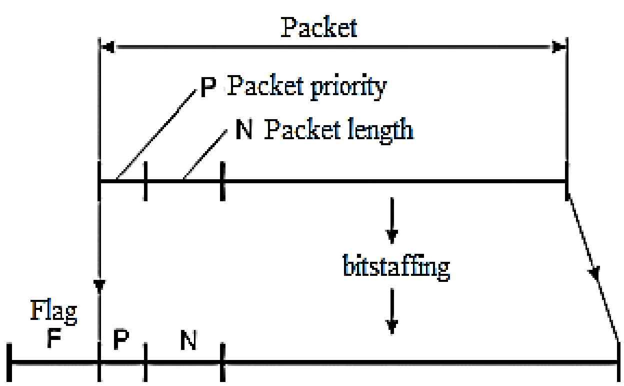


Fig. 8. Conversion of packets by blocks L1-L3 (Figure 7).

Recall that bit staffing allows you to exclude from the data stream a random copy of the unique code selected as the frame start flag F. In this example, F = 01111110.

In Fig. 9, and the "true" flag F of the beginning of the frame (circled in a rectangular frame) is inserted into some sequence of bits. The problem is that, most likely, this sequence also contains codes 01111110, which can be considered as false flags. In order to prevent the transmission of false flags to the far side of the communication channel, they are intentionally reversibly distorted, for example, according to the algorithm proposed in [7].

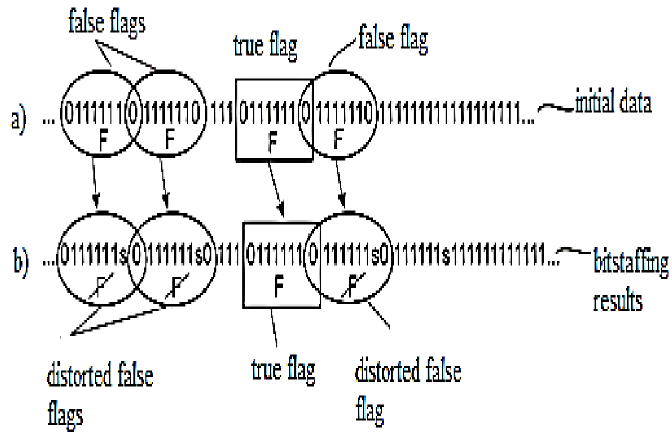


Fig. 9. Improved bitstaffing:.

a - the initial sequence of bits with the "true" flag of the beginning of the frame introduced into it; b - the same sequence after excluding false flags from it

This algorithm is as follows. The original sequence of bits with the "true" flag inserted into it is viewed through a sliding seven-bit window in order to detect in it the code 0111111, almost coincident with the flag. If such a code is detected and is not a component of the "true" flag, then it is supplemented by a single bit of s, regardless of the value of the subsequent bit (Figure 9, b). Such a procedure is called bitstaffing.

Bitstaffing does not apply to "true" flags, so they become unique, since all false flags are deliberately distorted by bits of s. On the far side of the communication channel, the reverse operation is performed - bits s (following the sequences 0111111, which are not constituent parts of the "true" flags) is destroyed.

In contrast to the classical bitstaffing used in the HDLC protocol, the variant proposed in [7] allows us to reduce the redundancy introduced into the initial bitstream by half. Indeed, for a single random sample, the probability of detecting a 7-bit code (0111111) in a random data stream is 1/27 = 1/128. In the classical version of bitstaffing, the probability of detecting a 6-bit code (011111) in a random data stream is 1/26 = 1/64. In other words, the insertion of redundant bits in the classical version of bitstaffing is carried out twice as often as in the version proposed in.

Suppose that in the initial state, a low-priority packet is sent to the line from the output buffer ## 1 of the switch (Figure 7). The SW1 switch is set to LP. As shown in Fig. 10, a, at some time T0, a high priority packet arrives from the upper channel of the packet transformer L1, bypassing the output queue. The transmission of the low priority packet terminates in the nearest bit interval, the SW1 switch goes to the HP state and the first flag bit of the high priority packet is placed in place of the not transmitted bit. Then all the bits of this packet are transmitted (Fig. 10, b).

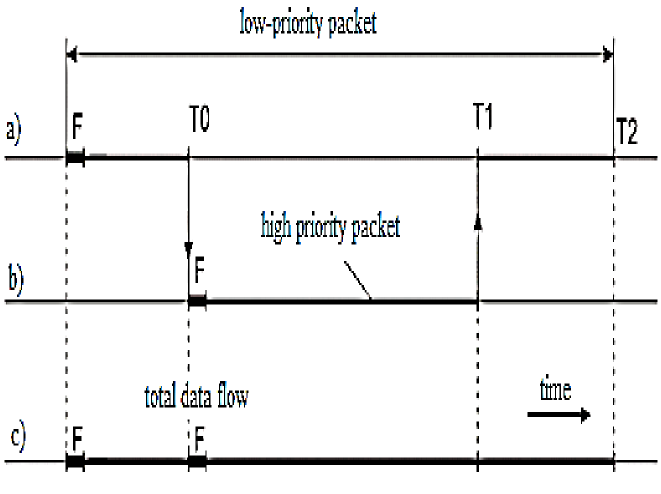


Fig. 10. Interruption of low-priority data stream high priority: a - low-priority data packet; b - high-priority data packet; c is the total data flow in the line.

At the time T1, the last bit of the high priority packet is transmitted. The key SW1 returns to the LP position. Following the last bit of the high-priority packet, all the bits of the previously suspended low-priority packet are transmitted. The total data flow (Fig. 10, c) can be divided on the far side of the communication channel into two components corresponding to Fig. 10, a and b, due to the uniqueness of the flags F and the presence of the P and N fields in the packet headers.

To simplify the analysis of code situations by the receiver, one can accept the condition that the low-priority packet flag is protected from interrupts, i.e., not crashed when switching to a high-priority packet transmission. In other words, if a high-priority packet has entered the SW1 key during the low priority packet transmission, it is delayed and its transmission begins only after the low-priority packet flag is fully transmitted. In the worst case, the delay is eight bit intervals. With a greater number of priority levels, the described process of switching data flows acquires the nature of nested interrupts widely used in microprocessor technology. As shown in Fig. 11, the transmission of packets can repeatedly go from one priority level to another and back.

In the period T0 - T1, the packet Y0 of the zero (lowest) priority level is transmitted to the line. At time T1, this transmission is interrupted due to the arrival of the Y1 packet of the first (higher) priority level. The transmission of the packet Y1, in turn, is interrupted at the time T2, after which the Y2 packet of the second priority level is fully transmitted. The end of the transmission of this packet is marked by a period.

At time T3, the switch returns to the transmission of the packet Y1, but at the time T4 the transmission is again interrupted by the higher priority packet Y3, which in turn is interrupted by the Y4 packet at the time T5. This packet has the highest priority; therefore its transfer cannot be interrupted under any circumstances.

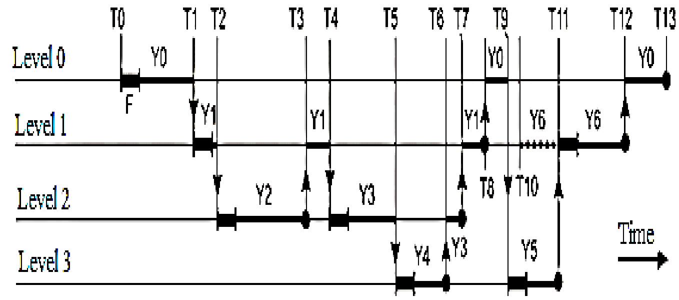


Fig. 11. Transmission of data packets Y0 to Y6 using a four-level priority system.

Further, at the moments T6 to T8, in the order of decreasing priorities, the transmissions of the packets Y4, Y3, Y1 are completed, and the transmission of the packet Y0 resumes. At time T9, this transmission is again interrupted by a Y5 packet having the highest priority. At time T10, the Y6 packet is ready for dispatch, but it is performed only starting from the moment T11, when the transmission of the Y5 packet is complete.

At the moments T12 and T13, the transmission of the packets Y6 and Y0 is completed.

III. CONCLUSION

As a summary, the article approved a number of original technical solutions to improve the quality of control and reduce the required amount of buffer memory of network nodes. High-priority packets are wedged into low-priority packets, without waiting for the end of their transmission. This allows reducing delays in high-priority packets and with low-priority packets of long length. The increase in the intelligence of telecommunication devices became possible to apply more sophisticated algorithms and original flow control schemes in comparison with the known ones. This allows us solving the following tasks:

- reduce the likelihood of overflow and emptying of buffer blocks located along the distribution routes of packets and, ultimately, improve the quality of computer networks;
- reduce the required amount of buffer memory; reducing the amount of buffer memory of the receiver when building a data transfer system between nodes of a computer network
- improve the efficiency of servicing high-priority packets.

ACKNOWLEDGMENT

I thank, King Abdulaziz University, for supporting and carrying out this work.

REFERENCES

[1] Archer, C.J. and G.R. Ricard, Administering registered virtual addresses in a hybrid computing environment including maintaining a cache of ranges of currently registered virtual addresses. 2016, Google Patents.

[2] Anderson, J.L. and T.J. Balph, Memory interface device with processing capability. 1981, Google Patents.

[3] Kim, S. and R. Lu, The Pseudo - Equivalent Groups Approach as an Alternative to Common - Item Equating. ETS Research Report Series, 2018.

[4] Sansyzbaevich, I.S., et al. Development of algorithm flow graph, mealy automaton graph and mathematical models of microprogram control mealy automaton for microprocessor control device. in Control and Communications (SIBCON), 2017 International Siberian Conference on. 2017. IEEE.

[5] Fujioka, Y., M. Kameyama, and M. Lukac. A dynamically reconfigurable VLSI processor with hierarchical structure based on a micropacket transfer scheme. in Information and Digital Technologies (IDT), 2017 International Conference on. 2017. IEEE.

[6] Kaushansky, D., et al., Programmable test instrument. 2017, Google Patents.

[7] Tan, C.J., et al. Review on Firmware. in Proceedings of the International Conference on Imaging, Signal Processing and Communication. 2017. ACM.

[8] Cabillic, G. and J.-P. Lesot, Selective compiling method, device, and corresponding computer program product. 2017, Google Patents.

[9] Wiśniewski, R., Prototyping of Concurrent Control Systems, in Prototyping of Concurrent Control Systems Implemented in FPGA Devices. 2017, Springer. p. 99-116.

[10] Durand, Y., et al. A Programmable Inbound Transfer Processor for Active Messages in Embedded Multicore Systems. in 2017 Euromicro Conference on Digital System Design (DSD). 2017. IEEE.

[11] Maeda, T. and R. Matsubara, Storage apparatus and failure location identifying method. 2017, Google Patents.

[12] Vladimirov, S. and R. Kirichek, The IoT Identification Procedure Based on the Degraded Flash Memory Sector, in Internet of Things, Smart Spaces, and Next Generation Networks and Systems. 2017, Springer. p. 66-74.

[13] Ye, J., A novel ship-borne positive pressure solid phase extraction device to enrich organo chlorinated and pyrethroid pesticides in seawater. Se pu= Chinese journal of chromatography, 2017. 35(9): p. 907-911.

[14] Vasumathi, B. and S. Moorthi, Implementation of hybrid ANN–PSO algorithm on FPGA for harmonic estimation. Engineering Applications of Artificial Intelligence, 2012. 25(3): p. 476-483.

[15] Wiśniewski, R., Modelling of Concurrent Systems in Hardware Languages, in Prototyping of Concurrent Control Systems Implemented in FPGA Devices. 2017, Springer. p. 117-137.

[16] Pearlson, K.E., C.S. Saunders, and D.F. Galletta, Managing and Using Information Systems, Binder Ready Version: A Strategic Approach. 2016: John Wiley & Sons.

[17] Ruiz, P.A.P., B. Kamsu-Foguem, and D. Noyes, Knowledge reuse integrating the collaboration from experts in industrial maintenance management. Knowledge-Based Systems, 2013. 50: p. 171-186.

[18] Han, Y.Y., et al., Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. Pediatrics, 2005. 116(6): p. 1506-1512.

[19] Jagadish, H., et al., Big data and its technical challenges. Communications of the ACM, 2014. 57(7): p. 86-94.

[20] Rafi, D.M., et al. Benefits and limitations of automated software testing: Systematic literature review and practitioner survey. in Proceedings of the 7th International Workshop on Automation of Software Test. 2012. IEEE Press.

[21] Al-Rababah, A. and N. Hani. Component linked based system. in Modern Problems of Radio Engineering, Telecommunications and Computer Science, 2004. Proceedings of the International Conference. 2004. IEEE.

[22] Rodríguez, P., et al., Continuous deployment of software intensive products and services: A systematic mapping study. Journal of Systems and Software, 2017. 123: p. 263-291.

[23] Taylor, S.J., R. Bogdan, and M. DeVault, Introduction to qualitative research methods: A guidebook and resource. 2015: John Wiley & Sons.

[24] Ciccozzi, F., et al., Model-Driven Engineering for Mission-Critical IoT Systems. IEEE Software, 2017. 34(1): p. 46-53.

[25] AlRababah, A.A., A new model of information systems efficiency based on key performance indicator (KPI). management, 2017. 4: p. 8.

[26] Al Ofeishat, H.A. and A.A. Al-Rababah, Real-time programming platforms in the mainstream environments. IJCSNS, 2009. 9(1): p. 197.

[27] Choi, J. and R.A. Rutenbar, Video-rate stereo matching using Markov random field TRW-S inference on a hybrid CPU+ FPGA computing platform. IEEE Transactions on Circuits and Systems for Video Technology, 2016. 26(2): p. 385-398.

Overview of Service and Deployment Models Offered by Cloud Computing, based on International Standard ISO/IEC 17788

Washington Garcia Quilachamin[1]
Student UNMSM - Lima, Perú
Professor Faculty of Engineer
ULEAM - Manta, Ecuador

Igor Aguilar Alonso[2]
Professor Faculty of Engineer
UNMSM Lima, Perú

Jorge Herrera-Tapia[3]
Professor of Computer Science
ULEAM - Manta, Ecuador

***Abstract*—Cloud Computing offers services over the Internet to support business processes, based on deployment models and service, meet business requirements in an efficient and cost-effective. A general context of the types of service models that it, as well as the models of deployment, are not known, so the following research questions are stated: Q1) How many studies refer to service models in Cloud Computing and Models of cloud computing deployment?, Q2) How the service models are classified in relation to the types of capabilities Application, Infrastructure y Platform in a Cloud?, and Q3) What types of cloud computing deployment models currently exist?. The objective of this paper is to investigate the service and deployment models that currently exist in cloud computing, for every which a process of systematic review of the literature has been used as a research methodology. The results show that 45 service models and 4 deployment models were found in Cloud Computing, this allows us to conclude that the offered models give a lot and diverse solutions for the business processes.**

***Keywords*—Cloud computing services models; IT demand management; deployment models; applications; platform; infraestructure**

I. Introduction

The important development that we are going through in the 21st century, according to [1] the use of Information Technology (IT), Cloud computing and the Internet of Things (IoT) have become popular with the exponential usage of smart devices in recent years. It is considered one of the key elements to achieve the objectives of an organization through the best practices of IT Governance [2]. Today, cloud computing is an example in technology business processes as resources must be available within seconds, or in minutes, compared to the traditional physical server acquisition process that could take weeks and years.

Cloud Computing is a business model that is designed to deliver services and IT resources through virtualization and distributed computing technologies; these are consumed as a service on demand and paid for their use. This new paradigm has the potential to change the relationships between IT service providers and consumers that influence competitiveness in various sectors, as well as changing the role of top executives and IT professionals in organizations [3].

Cloud Computing technologically offers services in an organization through the design of its architecture, applications, deployment models, and service models generally organized into service categories such as Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), Software-as-a-Service (SaaS), and Business Process-as-a-Service (BPaaS). These models are fundamental to any service management process in Cloud Computing [4], [5].

Technological development has allowed creating opportunities in the business processes, being this a problem in its implementation when new needs arise in aspects that influence in Knowledge, Data management, Web learning, Information Security, Business Process, Application, Platform, and Infrastructure, due to which Cloud Computing has created new applications and services for the end user or organization to meet that demand.

According to [6] ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. The main task of the joint technical committee is to prepare International Standards.

The objective of this paper is to investigate the service and deployment models that currently exist in cloud computing, for which a systematic review process of the literature has been used as a research methodology.

The new models of cloud computing services that are appearing according to the needs of the 21st century are based on SaaS, PaaS, and IaaS [7]. Therefore, to implement a service model, it is necessary to decide the type of Cloud deployment model as Public, Private, Hybrid, and Community [8].

Computer services are currently available on demand, as well as any public service available in society, which for the different management of services need to be implemented through various models of deployment in the cloud.

It is considered that all business process management is based on the observation that each product that a company offers to the market is the result of a series of activities performed [9]. These activities can be performed by humans, systems or a combination of both. By identifying and structuring these activities in workflows, companies get information about their business processes. By monitoring and reviewing their processes, companies can identify the problems within these processes and can achieve improvements in using IT.

This article is structured as follows: the first section, makes a reference to the introduction of Cloud Computing, its service models and deployment, the second section refers to the historical background of Cloud Computing, in the third Section the methodology used for the writing of this paper is explained, the fourth section shows an analysis of the results found on this research and the conclusions are presented in the fifth section.

II. Background

The concept of Cloud Computing was introduced in the 1960s by John McCarthy, who suggested that advances in information and communications would lead to that "one day, computing would be organized as a public service" just like the water or electricity businesses [10].

In [11] the authors state that the evolution of Cloud Computing began in 1980 from the complex and widespread roots of IT, and in the 1990s companies began to offer virtual private networks.

Cloud Computing begins to be considered as a business model in services from the twentieth century, in a scenario of IT platforms supported by new technologies such as Web 2.0 and distributed computing. In 2002, Amazon launched the Amazon Web Service market and in 2006 introduced the term Elastic Compute Cloud (EC2) as a commercial service that allows small and medium-sized businesses to rent servers where they can run their own applications [10].

From then on, in the following years, big companies like Google or IBM began to investigate cloud computing and, as a result of these investigations, in 2009, OpenNebula and Eucalyptus were born, an open source platform that allowed the creation of systems in Cloud Compatible with Amazon EC2 web services, [12]. Google with its own google apps began to offer useful and reliable applications for computers, smartphones, and tablets that run in Cloud computing.

According to Aguilar [13], in 2011 Apple launched its iCloud service, a cloud storage system - for documents, music, videos, photographs, applications, and calendars - that promised to change the way we use the computer.

A definition is provided by Armbrust et al. From the University of Berkeley [14], where Cloud Computing refers to applications used as a service over the Internet, known for a long time as "Software as a Service." National Institute of Standards and Technology (NIST) [15], [16] states that Cloud Computing is a technology model that allows ubiquitous, adaptive and on-demand network access to a shared set of configurable computing resources.

Therefore cloud computing is defined as a distributed consumer-oriented computing system, consisting of a collection of virtualized and interconnected computers [8]. Thus, the services that are in the network and offer their resources, highlight the services of hosting that allow us to save information outside our computers, that is to say, in servers that are in the cloud and to which we can accede through a communications network.

According to [17], the resources in a self-service demand allow access to the ubiquitous network to be scaled up and down quickly, [18] where resource use is measured, and service models are performed as a business through payment for consumer use. The diversity of business models through the use of IT, allows users to access a standardized service and respond to the needs of individuals and companies. In [9], [19], [20] the diverse applications are established, such as models of services that adapt to the necessities of computation in storage, speed, and availability, considering the most expensive part of the systems.

The IaaS provider in Cloud Computing allows the creation of a storage computing infrastructure and offering it for use in the business process, according to [21], [8], it requires an initial investment and established hardware maintenance in a service model. The utility of computing is based on service models and their implementation in Cloud, which are developed about the massive transformation of the entire computer industry of the twenty-first century.

At present, the emergence of Cloud Computing has drastically altered the perception of all infrastructure architectures, software design and development models within service management using IT.

ISO/IEC 17788 was prepared by Joint Technical Committee ISO/IEC JTC 1, Information technology, Subcommittee SC 38, Distributed application platforms and services (DAPS), in collaboration with ITU-T. The identical text is published as ITU-T Rec. Y.3500 (08/2014). ISO / IEC 17788: 2014 mentions service categories and emerging service categories as the service models and classifies them according to the types of capabilities in the cloud [6].

We consider this information as a source of research for its updating and expansion of the new models of cloud services found.

III. Research Methodology

The development of this work was done considering the guidelines used by [22], which have been adapted, focusing on main three processes: Research planning, development of information and results found..

A. Planning of the Research

According to the target of this investigation, the following research questions were elaborated: Q1: How many studies refer to service models in cloud computing and cloud computing deployment models?, Q2: How the service models are classified by the types of capabilities Application, Infrastructure, and Platform in a Cloud?, Q3: What types of cloud computing deployment models are currently available?.

In the construction of the search terms, the suggestions of [23] have been considered, such as the following:

- Keywords related to research questions.

- Identification of alternative terms and synonyms of keywords as search terms.

Searching for information by using the AND logical operator as a connector.

In this process, the inclusion and exclusion criteria are considered, elements which are very important in defining the results of the research.

The following are considered as necessary inclusion criteria: articles available in full text, the search range of the articles comprises from 2010 to 2018, articles related to research questions, limited search in the disciplines of Computer Science and engineering and sources of information from journals and congresses. The exclusion criteria considered are the following: articles that do not relate to service models in Cloud Computing, studies that are not related to research questions and articles that are not in the English language.

To obtain the results of the information needed for our research, we used the Internet and data sources in search of scientific knowledge that are accessible through a browser, a defined search string is applied taking into account the inclusion and Exclusion criteria, whose objective is to locate, select and obtain the documents that give answers to the questions asked.

In this paper, the following data sources were used: ACM Digital Library, IEEE Xplore Digital Library, Springer Link and Science Direct.

In the search chain, the protocols were defined, the sources of information were selected, and the search strategy was created. The following keywords were used for such purpose: Cloud Computing, Model Service, Model Types of services in Cloud Computing, Model service in Cloud Computing, Deployment Model in Cloud Computing, Implementation Model in Cloud Computing and the logical relationship was established using the operator AND, as detailed below: (Model Service AND Cloud Computing), (Model Types of services in Cloud Computing AND Model services in Cloud Computing), (Deployment Model AND Cloud Computing), (Implementation Model AND Cloud Computing)

The search was considered through specific fields (Title, Abstract, Key, Document title, Publication title) and delimited descriptors, dates, typology, etc.

B. Development of Information Search

In order to carry out the information search process shown in Figure 1, the criteria established in the previous section were considered. The information search process was performed by applying the search string to each of the data sources, and as a result, a total of 41857 articles were obtained, see Table 1.

A review and depuration of articles that did not meet the established criteria or were duplicated in different sources of information were made. Subsequently the abstract of the articles found was reviewed and a total of 40724 research work were eliminated, obtaining 1133 relevant studies from this process, from which the complete content of each of these articles was reviewed, obtaining 236 main studies and from the analysis of these articles, 45 service models offered by Cloud Computing and 4 deployment models were obtained.

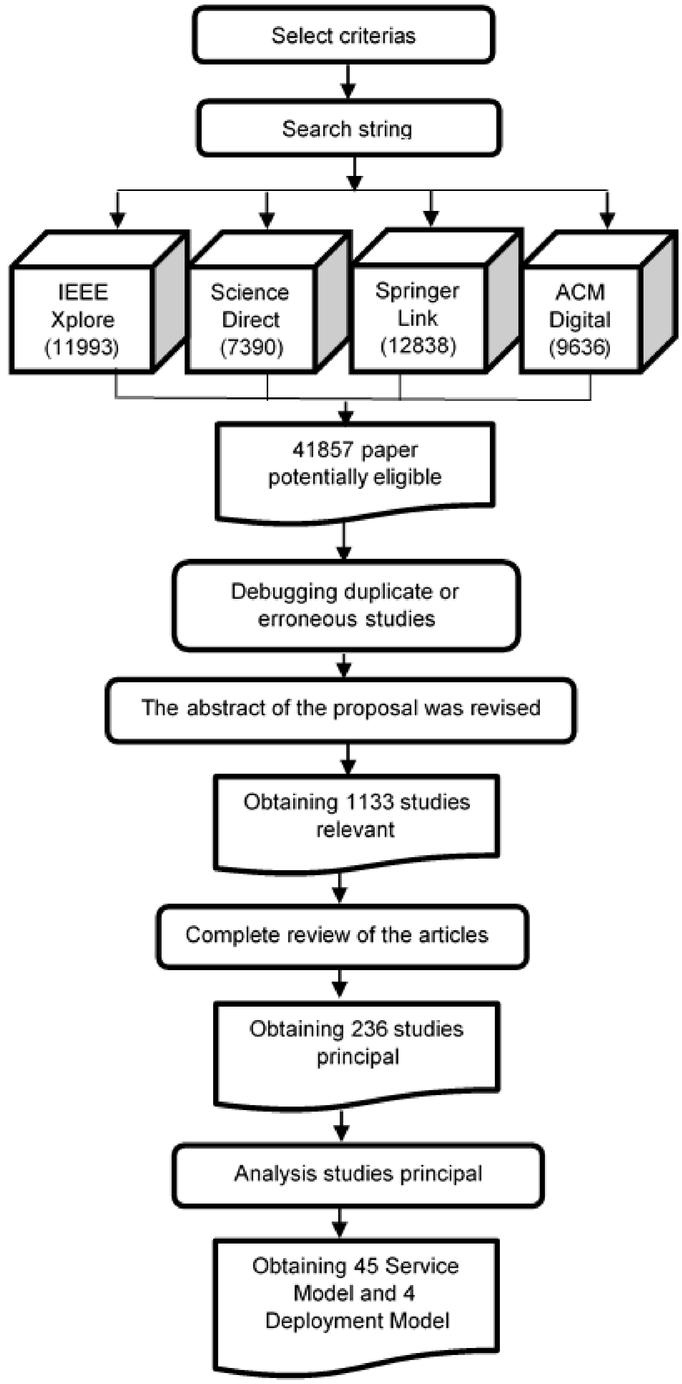


Fig. 1. Information Search Process Diagram.

TABLE I. NUMBER OF STUDIES OBTAINED THROUGH DATA SOURCES

Source	Potentially elegible studies	Studies Relevant	Studies principal	%
IEEE Xplore	11993	415	95	40,3%
Science Direct	7390	130	55	23,3%
Springer Link	12838	352	49	20,8%
ACM digital library	9636	236	37	15,7%
TOTAL	**41857**	**1133**	**236**	**100,0%**

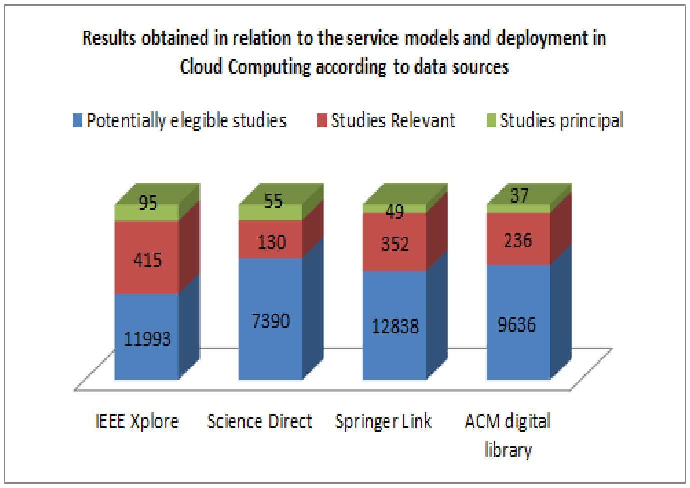


Fig. 2. Results Obtained In Relation to the Service Models and Deployment in Cloud Computing According to Data Sources

C. Results Found

In this section, the general results classified according to the data sources are shown, highlighting the potentially eligible studies, relevant studies and the primary studies with their respective percentage, as shown in Table 1.

Figure 2 shows the results obtained from the eligible potential studies in each of the data sources.

The following are the results found on the basis of the research questions:

a) Results that refer to the studies found in relation to the service and deployment models in Cloud Computing according to the data sources.

Results that refer to the studies found in relation to the 45 service models in Cloud Computing are diverse. Figure 3 shows the results of the number of studies found that refer to each of the cloud computing service models as detailed below:

The well-known models SaaS, PaaS, IaaS, BPaaS were the first service models that appeared for cloud computing. Later, from the combination of these 3 models new service models are looking, as well as the found in this research: Education Learning-as-a-Service (ELaaS), Hardware–as-a-Service (HaaS), Data-as-a-Service (DaaS), Application-as-a-Service (AaaS), Business Intelligence-as-a-Service (BIaaS), Testing-as- a-Service (TaaS), Communication–as-a-Service (CaaS), Knowledge-as-a–Service (KaaS), Actionable Knowledge–as-a-Service (AKaaS), Information Technology–as-a-Service (ITaaS), Broker–as-a-Service (BaaS), Modelling and Simulation-as-a-Service (MSaaS).

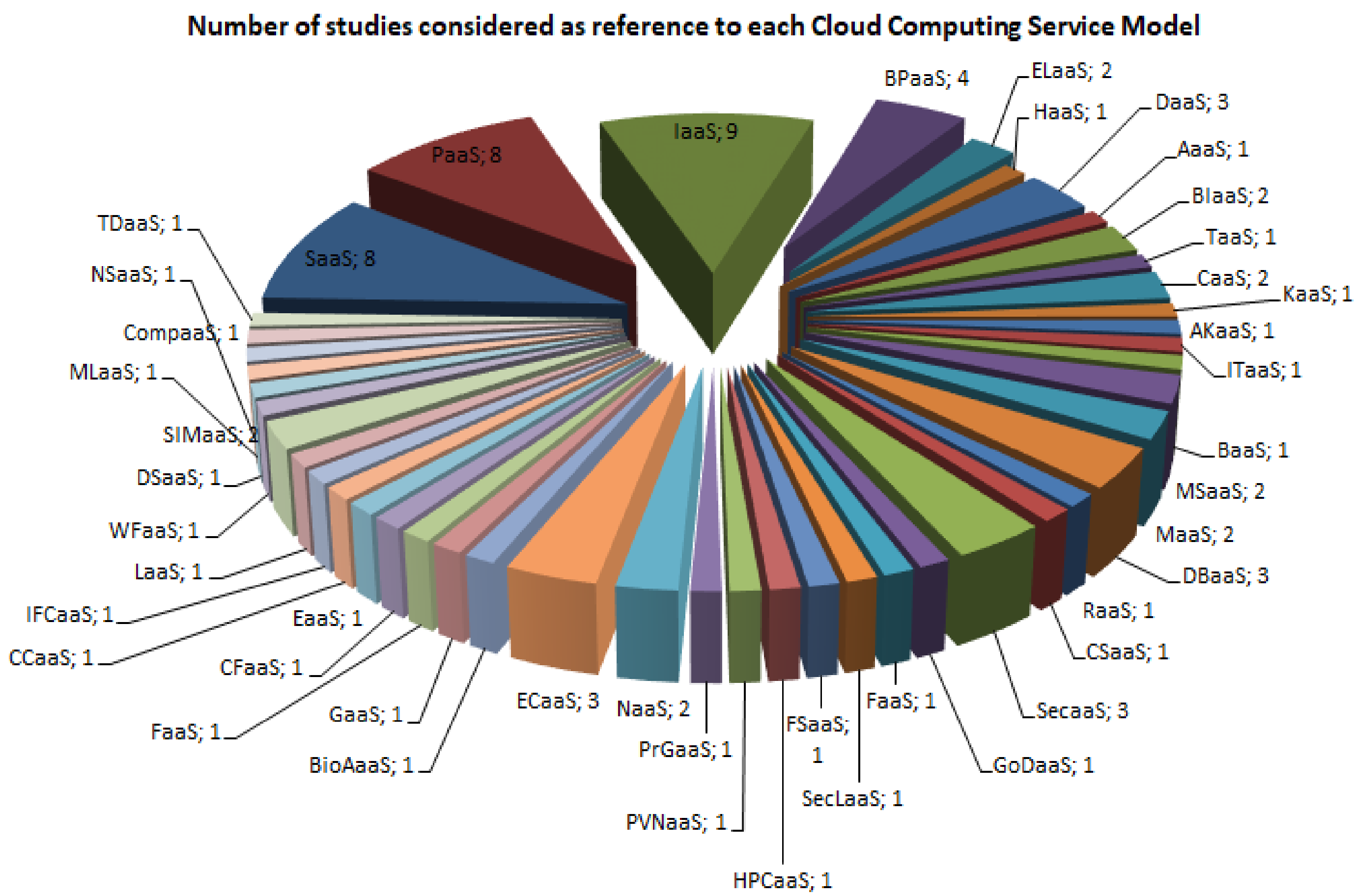


Fig. 3. Number of Studies Considered as Reference to Each Cloud Computing Service Model.

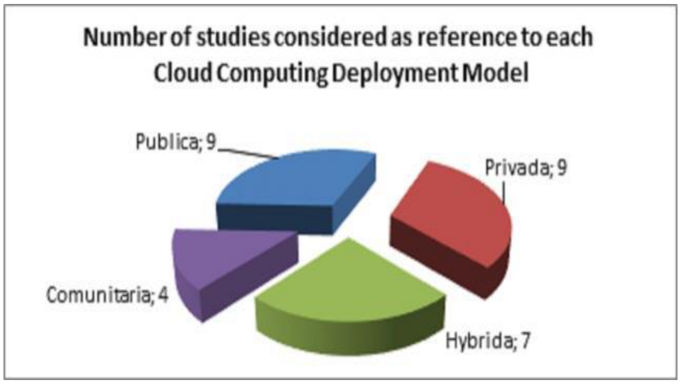


Fig. 4. Number of Studies Considered as a Reference to Each Cloud.

Modelling-as-a-Service (MaaS), Database-as-a-Service (DBaaS), Resource–as–a–Service (RaaS), Campus-System-as-a-Service (CSaaS), Security–as-a-Service, (SecaaS), (GoDaaS) Government Open Data-as-a-Service, Functions–as-a-Service (FaaS), Secure Logging–as-a-Service (SecLaaS), Failure Scenario-as-a-Service (FSaaS), High Performance Computing-as-a-Service (HPCaaS), Programmable virtual network-as-a-Service (PVNaaS), Prognostic-as-a-Service (PrGaaS), Network-as-a-Service (NaaS), E-Commerce-as-a-Service, (ECaaS), Biometric Authentication-as-a-Service (BioAaaS), Green-as-a-service (GaaS), Forensics-as-a-Service (FaaS), Cloud-Forensic-as-a-Service (CFaaS), Energy-as-a-Service (EaaS), Crowdsourced-Coverage-as-a-Service (CCaaS), Information-Flow-Control-as-a-Service (IFCaaS), Laboratory-as-a-Service (LaaS), Simulation-as-a-Service (SIMaaS), Workflow-as-a-Service (WFaaS), Data Security as a Service (DSaaS), Micro Learning as a Service (MLaaS), Network-Simulation-as-a-Service (NSaaS), Compute as a Service (CompaaS) and Traffic-Data-as-Service (TDaaS).

Figure 4 shows the results of the number of studies considered that refer to each of the cloud computing deployment models.

Computing Deployment Model

b) Results found from the cloud computing service models and relation with Cloud capabilities types.

In this section, we describe the service models found in Cloud Computing, and classification is made according to the type of capacity (Application, Infrastructure, and Platform), taking as reference the initial sorting in the International Standard ISO / IEC 17788: 2014 [6].

According to the studies conducted, the results found about the service models in the Cloud environment are 45, as shown in table 2.

To conclude the results section, Table 2 also shows the suppliers of the 45 service models found and the contribution of the service models in relation to technological, social and environmental aspects.

c) Results found from the cloud computing deployment models.

Depending on the relationship between the provider and the consumer, there are four different cloud computing deployment models within organizations as shown in table 3.

TABLE II. TYPES OF SERVICE MODELS FOUND IN CLOUD COMPUTING AND RELATION WITH CLOUD CAPABILITIES TYPES

N.	Services Models	Contribution			Cloud Capabilities Types			Providers	Author – reference	Total references
		Social	Environ-ment	Technology	Application	Infrastruc-ture	Platform			
1	SaaS	X			X			AppDirect, Concur, Ingram Micro, Net Suite, Parallels, Salesforce, GoogleApp, Office 365	[6], [52] , [53], [54], [55], [56], [57], [58]	8
2	PaaS			X			X	Amazon, Salesforce, long Jump, Windows Azure, IBM, OpenShift, CloudFoundry, GoogleApp, EngineYard	[6], [59], [60], [61], [62], [63], [64], [65].	8
3	IaaS			X		X		Ospero, Amazon, Microsoft Azure, , GoogleApp, Alibaba	[6], [66], [67], [68], [69], [70], [71], [72], [73].	9
4	BPaaS	X			X		X	BPM, Salesforce, BAM, ESB	[74], [75], [76], [77].	4
5	ELaaS	X			X		X	Avanzo, GoogleApp	[78], [79].	2
6	HaaS	X			X	X		Microsoft Azure, Amazon, Rackspace	[17].	1
7	DaaS	X			X	X	X	Oracle, Microsoft Azure, Forbes, Amazon	[6], [80], [81].	3
8	AaaS			X	X			GoogleApp, Microsoft	[82].	1
9	BIaaS	X			X		X	SAP, IBM, Oracle, Microsoft, Qlik, Tab leau, TeradataCorp.	[83], [84].	2
10	TaaS	X			X			Oracle,	[85].	1

11	CaaS	X			X		X	Oracle, IBM, Avoxi	[6], [86].	2
12	KaaS			X	X		X	GoogleApp	[87].	1
13	AKaaS	X		x	X		X	GoogleApp	[88].	1
14	ITaaS			X	X		X	IBM, Microsoft	[89].	1
15	BaaS	X			X		X	Amazon, Microsoft	[90].	1
16	MSaaS	X			X		X	IBM, CloudSME, Amazon	[91], [92].	2
17	MaaS			X	X		X	GoogleApp, Microsoft, Salesforce	[93], [94]	2
18	DBaaS	X		X	X	X	X	Azure SQL Database, DBaaS	[6], [95], [96].	3
19	RaaS	X			X	X	X	IBM, HP, Telcos	[97]	1
20	CSaaS	X		X	X		X	Windows	[98]	1
21	SecaaS	X		X	X	X	X	McAfee, Symantec, Trend Micro, VeriSign.	[6], [99], [100].	3
22	GoDaaS	X		X		X	X	Microsoft	[101].	1
23	FaaS			X	X			Microsoft Azure, Akamai	[102], [103]	2
24	SecLaaS	X		X	X	X	X	BackBlaze, Hightail, Akamai, SpiderOak	[104].	1
25	FSaaS			X	X			Windows, Amazon Web Services, Google App Engine	[105]	1
26	HPCaaS		X		X		X	IBM, Microsoft, AWS, ORACLE	[106]	1
27	PVNaaS	X			X	X	X	Microsoft, IBM, Office 365	[107]	1
28	PrGaaS	X			X			IBM, Amazon Web Services	[108]	1
29	NaaS	X		X	X	X	X	IBM, Akamai, Salesforce.com	[6], [109]	2
30	ECaaS	X			X		X	Inercia Digital, Microsoft, Amazon Web Services	[6], [110], [16]	3
31	BioAaaS			X	X	X	X	Symantec ES, Microsoft	[111]	1
32	GaaS		X		X		X	Amazon Web Services	[112]	1
33	FaaS	X			X			Amazon, Google App,	[113]	1
34	CFaaS	X		X	X		X	Google App	[114]	1
35	EaaS			X	X	X	X	IBM, Amazon Web Services aa	[115]	1
36	CCaaS			X	X		X	Microsoft, Salesforce.com	[116]	1
37	IFCaaS	X		X	X			Avoxi	[117]	1
38	LaaS	X			X	X	X	IBM	[118]	1
39	SIMaaS	X			X		X	Microsoft, Office 365	[119], [120]	2
40	WFaaS	X		X	X		X	Quantum, IBM, SddddAP	[119]	1
41	DSaaS			X	X	X	X	Akamai, Microsoft, Amazon	[121]	1
42	MLaaS			X	X	X	X	Microsoft, Windows, Google App, SAP	[122]	1
43	NSaaS		X		X	X	X	IBM, Salesforce	[123]	1
44	CompaaS			X		X	X	Windows, Oracle	[6]	1
45	TDaaS			X	X		X	Oracle, Microsoft Azure, IBM	[124]	1

TABLE III. CLOUD COMPUTING DEPLOYMENT MODEL

Deployment Models	**Description**	**Author – reference**	**Tot.Ref.**
Public cloud	Organization that sells services in the cloud.	[24], [25], [26], [27], [28], [29], [30], [31], [32].	9
Private cloud	Exclusive use for a single organization.	[33], [34], [35], [36], [37], [38], [39], [40], [41]	9
Community cloud	It's shared by various organizations-	[42], [43], [44], [45]	4
Hybrid cloud	Composition of two or more clouds (private, community or public)	[46], [47], [11], [48], [49], [50], [51]	7

IV. Analysis of Results

This section describes the report of the results obtained and gives an answer to each of the questions posed in section 3.1 and the analysis of the number of studies obtained through the data sources. IEEE Xplore found the largest number of main studies, 95 studies accounting for 40.3%, followed by Science Direct with 55 studies accounting for 23.3%, Springer Link with 49 studies accounting for 20.8% and ACM digital library with 37 studies representing 15.7%. of the 1133 relevant studies, 236 main studies were considered, of which reference is made to 45 service models and the 4 deployment models in Cloud Computing described in Table 1.

Table 2 shows the suppliers of the 45 service models found, and it is considered as a result of this analysis that IBM, Microsoft, Amazon suppliers are the most accepted in the business. The figure 5 shows the evolution of the cloud service models found from 2006 to 2018. These service models are suggested as proposed trends, from their first three initial models (IaaS, SaaS, PaaS).

A. *Analysis of the Studies that Refer to the Models of Services and Deployment According to the Data Sources*

This section analyses of the number of studies that refer to service and deployment models in Cloud computing.

Regarding the studies that refer to the Service models in Cloud Computing, we found 85 reviews, as can be seen in figure 3. SaaS and PaaS are two models of Cloud services referenced by authors in 8 articles; the IaaS model is referenced in 9 articles. IaaS, PaaS, and SaaS have more authors as a reference because they are the main service models in the cloud as a base structure for the construction of service models; BPaaS referenced by authors in 4 articles.

The DaaS, DBaaS, SecaaS, ECaaS, models are referenced each by three authors, the ELaaS, BIaaS, CaaS, MSaaS, MaaS, FaaS, NaaS and SIMaaS models are referenced each by two authors and Haas, AaaS, TaaS, Kaas, AKaaS, ITaaS, BaaS, RaaS, CSaaS, GoDaaS, SecLaaS, FSaaS, HPCaaS, PVNaaS, PrGaaS, BioAaaS, GaAs, FaaS, CFaaS, EaaS, CCaaS, IFCaaS, LaaS, WFaaS, DSaaS, MLaaS, NSaaS, CompaaS, TDaaS models are referenced by a single author, in which the services applied and used feasibly by different customers are described. of the 45 models found and displayed in Table 2, 17 models contribute to the social, 3 models contribute to the environment, 15 to technology, and 10 to the social and technological. It is considered that the contribution of these service models in the technical aspect refers to the storage of data in the cloud, the environmental aspect refers to the environment and social aspect is considered the human society.

In relation to the studies that refer to the models of Deployment in Cloud Computing, we found 29 studies, as can be seen in figure 4. Reference to Private Cloud and Public Cloud as deployment models is referenced by authors in 9 articles considering that they facilitate business operations in Cloud Computing, Hybrid Cloud is referenced by authors in 7 articles combines deployment models in Cloud and in a flexible, Agile and profitable manner and Community Cloud which is referenced in 4 articles.

Responding to Q1, the studies that refer to service models in Cloud Computing are 85 seen in Table 2. Also, the studies that touch the deployment models in Cloud Computing are 29 as can be seen in table 3.

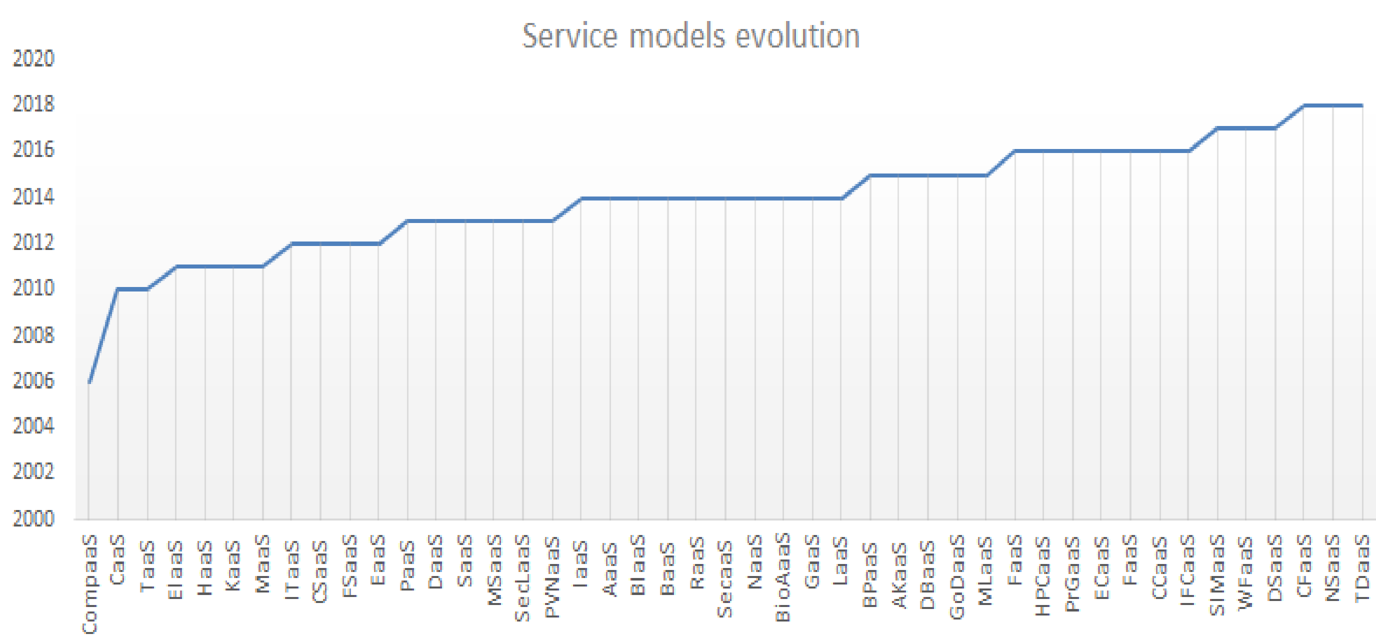


Fig. 5. Service Models Evolution.

B. *Analysis of the Service Models of Cloud Computing Found and Relation with Types of Cloud Capability*

As can be seen in the results found, there are 45 service models that appeared from the Infrastructure-as-a-Service, Platform-as-a-Service, Software-as-a-Service. Table 2 shows the service models found and classified according to their relationship with Cloud capability types (Application, Infrastructure, and Platform).

As it can be seen, the cloud computing services can have any of the combinations of the types of capacity in the cloud or just one class, which is indicated by an "X" and is detailed below: 8 have application type, 1 has a platform type, 1 has infrastructure type, 1 has application type and infrastructure, 19 have application type and platform, 2 have infrastructure type and platform, 13 have application type, infrastructure, and platform.

Responding to Q2, the service models found and classified in relation to the types of Application Capability, Infrastructure and Platform in a Cloud environment were 45 service models as indicated in table 2.

C. *Analysis of the Cloud Deployment Models Found*

As can be seen in the results found there are 4 models of deployment that have not changed since their appearance, as indicated in table 3.

Responding to Q3, the deployment models in Cloud Computing found were 4 models.

V. Conclusion

Considering the International Standard ISO / IEC 17788 as a basis, the classification of cloud service models was continued, and the search for information was done using search engines, obtaining a significant amount of information in IEEE Xplore that represents 42.2%.

The total result of the information search was 30173 articles, after performing a depuration of articles that did not meet the established criteria or were duplicated, obtaining 1050 relevant studies.

From the review of the relevant studies, we have found 211 main studies that reference the authors to 39 service models and 4 deployment models in Cloud Computing found and that help different users and organizations through the internet, which facilitate the development in businesses, storage, integration, scalability, and security, as well as the exchange of knowledge between developers of different applications.

The IaaS, SaaS and PaaS models are the basis for the creation of new service models that offer cloud computing and the types of models found in Cloud Computing generate solutions in strategic services to the business processes for companies, organizations, and users.

The types of capabilities in the cloud Applications, Infrastructure and Platform, allow the establishment of specific functions of the cloud service models considering aspects that influence the Virtualization, Knowledge, Data management, Web learning, Information Security and Business Process, which will continue to emerge as cloud computing continues to grow.

With the models of services found, suppliers, contribution and types of capabilities, it is suggested that ISO / IEC analyze and make modifications that cover these trends in their regularization.

It is concluded that each service model is aimed at different clients and their adoption is in relation to the time and present need, considering that the models of deployment in Cloud computing facilitate agile business processes.

As future work is being considered the analysis of the hypothesis related to the level of contribution and usability of cloud technologies in the Market. Where is considered the scalability of each service model found and its employability in business.

References

[1] R. Chaudhary, N. Kumar, and S. Zeadally, “Network Service Chaining in Fog and Cloud Computing for the 5G Environment : Data Management and Security Challenges,” no. November, pp. 114–122, 2017.

[2] I. Aguilar Alonso, J. Carrillo Verdún, and E. Tovar Caro, “Description of the framework’s structure of the process of IT demand management,” Int. J. Inf. Manage., 2016.

[3] I. A. Alonso, J. C. Verdún, and E. T. Caro, “IT, senior executives and board of directors contribute to the success of the business: Implicates on the it demand process - Life cycle,” in ICCIT 2009 - 4th International Conference on Computer Sciences and Convergence Information Technology, 2009, pp. 149–156.

[4] N. Suicimezov and M. R. Georgescu, “IT Governance in Cloud,” Procedia Econ. Financ., vol. 15, pp. 830–835, 2014.

[5] S. Srinivasan, “Cloud Computing Basics,” vol. 1, no. September 2011, pp. 1–17, 2014.

[6] I. Electrotechnical, “International Standard Iso / Iec,” vol. 1999, 2000.

[7] A. Prasad, P. Green, and J. Heales, “On governance structures for the cloud computing services and assessing their effectiveness,” Int. J. Account. Inf. Syst., vol. 15, no. 4, pp. 335–356, Dec. 2014.

[8] Z. Dvhg, H. E. Ssolfdwlrq, and Q. Hqd, “Implementation of cloud computing and big data with Java based web application,” pp. 1289–1293, 2016.

[9] A. M. Drivers and G. Implementation, “Desinge and Implementation of Government Cloud Computing Requirements : TOGAF,” 2017.

[10] Y. Jadeja and K. Modi, “Cloud computing - Concepts, architecture and challenges,” 2012 Int. Conf. Comput. Electron. Electr. Technol. ICCEET 2012, pp. 877–880, 2012.

[11] A. Srinivasan, M. A. Quadir, and V. Vijayakumar, “Era of Cloud Computing: A New Insight to Hybrid Cloud,” Procedia Comput. Sci., vol. 50, pp. 42–51, 2015.

[12] K. C. Valencia and S. Público, “Historia Del Cloud Computing,” Rev. Inf. Tecnol. y Soc. versión impresa, vol. 7, no. 1997–4044, pp. 51–52, 2012.

[13] L. J. Aguilar, “La Computación en Nube (Cloud Computing): El nuevo paradigma tecnológico para empresas y organizaciones en la Sociedad del Conocimiento,” Rev. Icade. Rev. las Fac. Derecho y Ciencias Económicas y Empres., no. 76, pp. 95–111, 2012.

[14] M. Armbrust et al., “A view of cloud computing,” Commun. ACM, vol. 53, no. 4, pp. 50–58, 2010.

[15] P. M. Mell and T. Grance, “The NIST definition of cloud computing,” 2011.

[16] M. Almubaddel and A. M. Elmogy, “Cloud Computing Antecedents , Challenges , and Directions,” 2016.

[17] P. Kalagiakos and P. Karampelas, “Cloud Computing learning,” 2011, pp. 1–4.

[18] A. W. Muzaffar and M. Waseem, “Application of Model Driven

Engineering in Cloud Computing – A Systematic Literature Review," pp. 7–12, 2017.

[19] P. T. Endo, M. Rodrigues, G. E. Gonçalves, J. Kelner, D. H. Sadok, and C. Curescu, "High availability in clouds: systematic review and research challenges," J. Cloud Comput., vol. 5, no. 1, p. 16, 2016.

[20] O. Pantelić, A. Pajić, and A. Nikolić, "Analysis of Available Cloud Computing Models to Support Cloud Adoption Decision Process in an Enterprise," no. Icccc, pp. 135–139, 2016.

[21] H. Jin, S. Ibrahim, T. Bell, W. Gao, D. Huang, and S. Wu, "Cloud Types and Services," pp. 335–355, 2010.

[22] B. Kitchenham, "Procedures for performing systematic reviews," Keele, UK, Keele Univ., vol. 33, no. 2004, pp. 1–26, 2004.

[23] M. Svahnberg, T. Gorschek, R. Feldt, R. Torkar, S. Bin Saleem, and M. U. Shafique, "A systematic review on strategic release planning models," Inf. Softw. Technol., vol. 52, no. 3, pp. 237–248, Mar. 2010.

[24] B. L. Muhammad-Bello and M. Aritsugi, "TCloud : A Transparent Framework for Public Cloud Service Comparison," 2016 IEEE/ACM 9th Int. Conf. Util. Cloud Comput., pp. 228–233, 2016.

[25] A. Antoniadis, Y. Gerbessiotis, M. Roussopoulos, and A. Delis, "Tossing NoSQL – Databases out to Public Clouds," Util. Cloud Comput. (UCC), 2014 IEEE/ACM 7th Int. Conf., pp. 223–232, 2014.

[26] I. Konstantinou, E. Floros, and N. Koziris, "Public vs Private Cloud Usage Costs: the StratusLab Case," CloudCP '12 Proc. 2nd Int. Work. Cloud Comput. Platforms , 2012.

[27] A. G. Kumbhare, Y. Simmhan, and V. Prasanna, "Designing a secure storage repository for sharing scientific datasets using public clouds," Int. Work. Data Intensive Comput. Clouds, p. 31, 2011.

[28] A. Zimba and C. Hongsong, "Analyzing Trust Concerns in Public Clouds Using Finite State Automata," pp. 25–29, 2016.

[29] S. Khatua, P. K. Sur, R. K. Das, and N. Mukherjee, "Heuristic-Based Resource Reservation Strategies for Public Cloud," IEEE Trans. Cloud Comput., vol. 4, no. 4, pp. 392–401, 2016.

[30] J. Weinman, "Migrating to - or away from - the Public Cloud," IEEE Cloud Comput., vol. 3, no. 2, pp. 6–10, 2016.

[31] J. Wang, P. Varman, and C. Xie, "Optimizing storage performance in public cloud platforms," J. Zhejiang Univ. Sci. C, vol. 12, no. 12, pp. 951–964, 2011.

[32] K. Xue and P. Hong, "A dynamic secure group sharing framework in public cloud computing," IEEE Trans. Cloud Comput., vol. 2, no. 4, pp. 459–470, 2014.

[33] M. LaPointe, L. Walker, M. Nelson, J. Shananaquet, and X. Wang, "Comparing public and private iaas cloud models," Proc. 3rd Annu. Conf. Res. Inf. Technol. - RIIT '14, pp. 69–70, 2014.

[34] D. W. Chadwick, M. Casenove, and K. Siu, "My private cloud – granting federated access to cloud resources," J. Cloud Comput. Adv. Syst. Appl., vol. 2, no. 1, p. 3, 2013.

[35] J. Dantas, R. Matos, J. Araujo, and P. Maciel, "Eucalyptus-based private clouds: availability modeling and comparison to the cost of a public cloud," Computing, vol. 97, no. 11, pp. 1121–1140, 2015.

[36] B. An, X. Shan, Z. Cui, C. Cao, and D. Cao, "Workspace as a Service: An Online Working Environment for Private Cloud," 2017 IEEE Symp. Serv. Syst. Eng., pp. 19–27, 2017.

[37] E. D. Canedo, R. T. de Sousa, R. R. de Carvalho, and R. de Oliveira Albuquerque, "Trust model for private cloud," Cyber Secur. Cyber Warf. Digit. Forensic (CyberSec), 2012 Int. Conf., no. c, pp. 128–132, 2012.

[38] V. Davidovic, D. Ilijevic, V. Luk, and I. Pogarcic, "Private Cloud Computing and delegation of control," in Energy Procedia, 2015.

[39] C. Baun, M. Kunze, T. Kurze, and V. Mauch, "Private Cloud-Infrastrukturen und Cloud-Plattformen," Informatik-Spektrum, vol. 34, no. 3, pp. 242–254, 2011.

[40] U. Sharma, P. Shenoy, and S. Sahu, "A flexible elastic control plane for private clouds," Proc. 2013 ACM Cloud Auton. Comput. Conf. - CAC '13, p. 1, 2013.

[41] Y. Li, J. Zhang, Q. Hu, and J. Pei, "Research and practice on the theory of private clouds migration," pp. 1813–1818, 2016.

[42] A. Marinos and G. Briscoe, "Community cloud computing," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5931 LNCS, no. December, pp. 472–484, 2009.

[43] M. Gall, A. Schneider, and N. Fallenbeck, "An Architecture for Community Clouds Using Concepts of the Intercloud," pp. 74–81, 2013.

[44] H. Nicanfar, Q. Liu, P. Talebifard, W. Cai, and V. C. M. Leung, "Community Cloud : Concept , Model , Attacks and Solution," 2013.

[45] B. K. Rani, B. P. Rani, and A. V. Babu, "Cloud Computing and Inter-Clouds – Types, Topologies and Research Issues," Procedia Comput. Sci., vol. 50, pp. 24–29, 2015.

[46] M. Industry and B. Robert, "Hybrid Public Private Cloud Computing for the," pp. 56–59.

[47] Y. H. Chang and J. Y. Chen, "A hybrid cloud for effective retrieval from public cloud services," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7802 LNAI, no. PART 1, pp. 61–69, 2013.

[48] D. Dobre and P. Viotti, "Hybris : Robust Hybrid Cloud Storage."

[49] M. Dayarathna, "Towards Scalable Distributed Graph Database Engine for Hybrid Clouds," 2014.

[50] G. S. Blair and B. Surajbali, "Experiences of Using a Hybrid Cloud to Construct an Environmental Virtual Observatory," pp. 13–18.

[51] W. J. Wang, Y. S. Chang, W. T. Lo, and Y. K. Lee, "Adaptive scheduling for parallel tasks with QoS satisfaction for hybrid cloud environments," J. Supercomput., vol. 66, no. 2, pp. 783–811, 2013.

[52] W.-W. Wu, "Developing an explorative model for SaaS adoption," Expert Syst. Appl., vol. 38, no. 12, pp. 15057–15064, Nov. 2011.

[53] J. Espadas, A. Molina, G. Jiménez, M. Molina, R. Ramírez, and D. Concha, "A tenant-based resource allocation model for scaling Software-as-a-Service applications over cloud computing infrastructures," Futur. Gener. Comput. Syst., vol. 29, no. 1, pp. 273–286, Jan. 2013.

[54] S. S. Kale and R. H. Borhade, "Development of Multitenant SaaS framework at single instance and with zero effort multitenancy," in 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013, pp. 834–839.

[55] A. F. Mohammad, J. Dargham, H. Mcheick, and A. T. Noor, "Software Evolution as SaaS: Evolution of Intelligent Design in Cloud," Procedia Comput. Sci., vol. 19, pp. 486–493, Jan. 2013.

[56] W. Tsai, X. Bai, and Y. Huang, "Software-as-a-service (SaaS): perspectives and challenges," Sci. China Inf. Sci., vol. 57, no. 5, pp. 1–15, May 2014.

[57] K. C. Huang and B. J. Shen, "Service deployment strategies for efficient execution of composite SaaS applications on cloud platform," J. Syst. Softw., vol. 107, pp. 127–141, 2015.

[58] Y. Cheng, Y. Chen, R. Wei, and H. Luo, "Development of a Construction Quality Supervision Collaboration System Based on a SaaS Private Cloud," J. Intell. Robot. Syst. Theory Appl., vol. 79, no. 3–4, pp. 613–627, 2015.

[59] C. Lv, Q. Li, Z. Lei, J. Peng, W. Zhang, and T. Wang, "PaaS: A revolution for information technology platforms," 2010, pp. 346–349.

[60] W. Zhang, X. Huang, N. Chen, W. Wang, and H. Zhong, "PaaS-Oriented Performance Modeling for Cloud Computing," in 2012 IEEE 36th Annual Computer Software and Applications Conference, 2012, pp. 395–404.

[61] L. Rodero-Merino, L. M. Vaquero, E. Caron, A. Muresan, and F. Desprez, "Building safe PaaS clouds: A survey on security in multitenant software platforms," Comput. Secur., vol. 31, no. 1, pp. 96–108, Feb. 2012.

[62] H. Sun, X. Wang, M. Yan, Y. Tang, and X. Liu, "Towards a Scalable PaaS for Service Oriented Software," in 2013 International Conference on Parallel and Distributed Systems, 2013, pp. 522–527.

[63] C. Teixeira, J. S. Pinto, R. Azevedo, T. Batista, and A. Monteiro, "The Building Blocks of a PaaS," J. Netw. Syst. Manag., vol. 22, no. 1, pp. 75–99, Jan. 2014.

[64] F. Paraiso, P. Merle, and L. Seinturier, "soCloud: a service-oriented component-based PaaS for managing portability, provisioning, elasticity, and high availability across multiple clouds," Computing, vol. 98, no. 5, pp. 539–565, May 2016.

[65] P. A. Dhuldhule, J. Lakshmi, and S. K. Nandy, "High Performance

Computing Cloud – A Platform-as-a-Service Perspective," in 2015 International Conference on Cloud Computing and Big Data (CCBD), 2015, pp. 21–28.

[66] D. Shin, H. Akkan, W. Claycomb, and K. Kim, "Toward role-based provisioning and access control for infrastructure as a service (IaaS)," J. Internet Serv. Appl., vol. 2, no. 3, pp. 243–255, Dec. 2011.

[67] A. Nathani, S. Chaudhary, and G. Somani, "Policy based resource allocation in IaaS cloud," Futur. Gener. Comput. Syst., vol. 28, no. 1, pp. 94–103, Jan. 2012.

[68] S. S. Manvi and G. Krishna Shyam, "Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey," J. Netw. Comput. Appl., vol. 41, pp. 424–440, May 2014.

[69] J. A. Wickboldt, R. P. Esteves, M. B. De Carvalho, and L. Z. Granville, "Resource management in IaaS cloud platforms made flexible through programmability," Comput. Networks, vol. 68, pp. 54–70, 2014.

[70] M. Carvalho, D. Menascé, and F. Brasileiro, "Prediction-Based Admission Control for IaaS Clouds with Multiple Service Classes," in 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom), 2015, pp. 82–90.

[71] U. A. Kashif, Z. A. Memon, A. R. Balouch, and J. A. Chandio, "Distributed trust protocol for IaaS Cloud Computing," in 2015 12th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2015, pp. 275–279.

[72] Y. Yamato, Tran., "Automatic verification technology of software patches for user virtual environments on IaaS cloud," J. Cloud Comput., vol. 4, no. 1, pp. 1–14, 2015.

[73] S. H. H. Madni, M. S. A. Latiff, Y. Coulibaly, and S. M. Abdulhamid, "Resource scheduling for infrastructure as a service (IaaS) in cloud computing: Challenges and opportunities," J. Netw. Comput. Appl., vol. 68, pp. 173–200, Jun. 2016.

[74] E. Duipmans and D. L. F. Pires, "Business process management in the cloud: business process as a service (BPaaS)," Univ. Twente, 2012.

[75] T. M. H. Le, L. A. Alfredo, H. R. Choi, M. J. Cho, and C. S. Kim, "A Study on BPaaS with TCO Model," in 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, 2014, pp. 249–256.

[76] R. Woitsch and W. Utz, "Business Processes as a Service (BPaaS): A model-based approach to align business with cloud offerings," 2015, pp. 1–8.

[77] R. Woitsch and W. Utz, "Business Process as a Service: Model Based Business and IT Cloud Alignment as a Cloud Offering," 2015, pp. 121–130.

[78] M. M. Alabbadi, "Cloud computing for education and learning: Education and learning as a service (ELaaS)," 2011, pp. 589–594.

[79] A. Ghazizadeh and M. Manouchehry, "Cloud Computing Based Technologies, Applications and Structure in U-learning," in Mobile and Ubiquitous Technology in Education 2012 IEEE Seventh International Conference on Wireless, 2012, pp. 196–198.

[80] Q. H. Vu, T. V. Pham, H. L. Truong, S. Dustdar, and R. Asal, "DEMODS: A Description Model for Data-as-a-Service," 2012, pp. 605–612.

[81] R. J. Thara, S. Shine, and C. Sanu, "Optimizing the performance of Database as a Service (DaaS) model #x2014; A distributed approach," 2013, pp. 1–5.

[82] A. Dukhanov, M. Karpova, and K. Bochenina, "Design Virtual Learning Labs for Courses in Computational Science with Use of Cloud Computing Technologies," Procedia Comput. Sci., vol. 29, pp. 2472–2482, 2014.

[83] V. Chang, "The Business Intelligence as a Service in the Cloud," Futur. Gener. Comput. Syst., vol. 37, pp. 512–534, Jul. 2014.

[84] H. Al-Aqrabi, L. Liu, R. Hill, and N. Antonopoulos, "Cloud BI: Future of business intelligence in the Cloud," J. Comput. Syst. Sci., vol. 81, no. 1, pp. 85–96, Feb. 2015.

[85] P. Mishra and N. Tripathi, "Testing as a Service," H. Mohanty, J. R. Mohanty, and A. Balakrishnan, Eds. Springer Singapore, 2017, pp. 149–176.

[86] C. Diana, E. Pacenti, and R. Tassi, "Communication tools for (service) design," p. 65.

[87] R. Abdullah, Z. D. Eri, and A. M. Talib, "A model of knowledge management system for facilitating knowledge as a service (KaaS) in cloud computing environment," 2011, pp. 1–4.

[88] A. Depeige and D. Doyencourt, "Actionable Knowledge As A Service (AKAAS): Leveraging big data analytics in cloud computing environments," J. Big Data, vol. 2, no. 1, Dec. 2015.

[89] A. M. Talib and R. Abdullah, "A Model of Information Technology as a Service (ITaaS) in Cloud Computing: A Case of Collaborative Knowledge Management System," 2012, pp. 83–86.

[90] M. Aazam and E. N. Huh, "Broker as a Service (BaaS) Pricing and Resource Estimation Model," 2014, pp. 463–468.

[91] S. Kothari, T. Peck, J. Zeng, F. Oblea, A. E. Votaw, and G. Dispoto, "Simulation as a cloud service for short-run high throughput industrial print production using a service broker architecture," Simul. Model. Pract. Theory, vol. 58, pp. 115–139, 2015.

[92] T. Kiss, H. Dagdeviren, S. J. E. Taylor, A. Anagnostou, and N. Fantini, "Business models for cloud computing: experiences from developing Modeling Simulation as a Service applications in industry," 2015, pp. 2656–2667.

[93] C. Atkinson, T. Schulze, and S. Klingert, "Modelling as a Service (MaaS): Minimizing the Environmental Impact of Computing Services," 2011, pp. 519–523.

[94] G. Zou, B. Zhang, J. Zheng, Y. Li, and J. Ma, "MaaS: Model as a Service in Cloud Computing and Cyber-I Space," 2012, pp. 1125–1130.

[95] J. Köhler, K. Jünemann, and H. Hartenstein, "Confidential database-as-a-service approaches: taxonomy and survey," J. Cloud Comput., vol. 4, no. 1, pp. 1–14, 2015.

[96] K. Munir, "Security model for cloud database as a service (DBaaS)," 2015, pp. 1–5.

[97] O. Agmon Ben-Yehuda, M. Ben-Yehuda, A. Schuster, and D. Tsafrir, "The Rise of RaaS: The Resource-as-a-service Cloud," Commun. ACM, vol. 57, no. 7, pp. 76–84, Jul. 2014.

[98] J. Mbale, K. Mufeti, and V. Hamutenya, "Examining Ubiquitous Security-Capital Issues in Implementing a Campus-System-as-a-Service(CSaaS) Model in the Cloud Computing Age: Case Study sub-Saharan Region," 2012 International Conference for Internet Technology and Secured Transactions. pp. 99–104, 2012.

[99] A. Furfaro, A. Garro, and A. Tundis, "Towards Security as a Service (SecaaS): On the modeling of Security Services for Cloud Computing," 2014, pp. 1–6.

[100] V. Varadharajan and U. Tupakula, "Security as a Service Model for Cloud Environment," IEEE Trans. Netw. Serv. Manag., vol. 11, no. 1, pp. 60–75, Mar. 2014.

[101] S. Qanbari, N. Rekabsaz, and S. Dustdar, "Open Government Data as a Service (GoDaaS): Big Data Platform for Mobile App Developers," in 2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud), 2015, pp. 398–403.

[102] Mike Roberts, "Serverless Architectures," 2016.

[103] M. Roberts and J. Chapin, What is Serverless ? 2017.

[104] S. Zawoad, A. K. Dutta, and R. Hasan, "SecLaaS: secure logging-as-a-service for cloud forensics," in Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, 2013, pp. 219–230.

[105] F. Faghri, S. Bazarbayev, M. Overholt, R. Farivar, R. H. Campbell, and W. H. Sanders, "Failure scenario as a service (FSaaS) for Hadoop clusters," in Proceedings of the Workshop on Secure and Dependable Middleware for Cloud Monitoring and Management, 2012, p. 5.

[106] [106]M. H. Radadiya and V. Rohokale, "Implementation of Costing Model for High Performance Computing as a Services on the Cloud Environment," 2016, pp. 1–6.

[107] J. LIANG et al., "Programmable virtual network as a service: towards a future Internet testbed based on IaaS cloud," J. China Univ. Posts Telecommun., vol. 20, pp. 126–130, Aug. 2013.

[108] Z. Bouzidi, L. Terrissa, A. Lahmadi, N. Zerhouni, and R. Gouriveau, "Neuro-fuzzy model for Prognostic as a Service in private cloud computing," in 2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech), 2016, pp. 360–367.

[109] J. F. Bravo-Torres, E. F. Ordóñez-Morales, M. López-Nores, Y. Blanco-Fernández, and J. J. Pazos-Arias, "Virtualization in VANETs to support the vehicular cloud #x2014; Experiments with the network as a service model," in Third International Conference on Future Generation Communication Technologies (FGCT 2014), 2014, pp. 1–6.

[110] A. M. Talib and F. O. Alomary, "Cloud Computing Based E-Commerce As a Service Model: Impacts and Recommendations," in Proceedings of the International Conference on Internet of Things and Cloud Computing, 2016, p. 27:1–27:7.

[111] C. S. Vorugunti and S. S. V. Indukuri, "A Secure and Efficient BiometricAuthentication As a Service for Cloud Computing," in Proceedings of the 6th IBM Collaborative Academia Research Exchange Conference (I-CARE) on I-CARE 2014, 2014, p. 7:1–7:4.

[112] M. I. Hassan and R. Bahsoon, "Green-as-a-service (GaaS) for Cloud Service Provision Operation," in Proceedings of the 29th Annual ACM Symposium on Applied Computing, 2014, pp. 1219–1220.

[113] B. K. Raju, B. Moharil, and G. Geethakumari, "FaaSeC: Enabling Forensics-as-a-service for Cloud Computing Systems," in Proceedings of the 9th International Conference on Utility and Cloud Computing, 2016, pp. 220–227.

[114] A. N. Moussa, N. Ithnin, and A. Zainal, "CFaaS: bilaterally agreed evidence collection," J. Cloud Comput., vol. 7, no. 1, pp. 1–19, 2018.

[115] M. Altamimi, R. Palit, K. Naik, and A. Nayak, "Energy-as-a-Service (EaaS): On the efficacy of multimedia cloud computing to save smartphone energy," Proc. - 2012 IEEE 5th Int. Conf. Cloud Comput. CLOUD 2012, pp. 764–771, 2012.

[116] A. G. Neiat, A. Bouguettaya, T. Sellis, and S. Mistry, "Crowdsourced Coverage as a Service: Two-Level Composition of Sensor Cloud Services," IEEE Trans. Knowl. Data Eng., vol. 29, no. 7, pp. 1384–1397, 2017.

[117] M. Elsayed and M. Zulkernine, "IFCaaS: Information Flow Control as a Service for Cloud Security," 2016 11th Int. Conf. Availability, Reliab. Secur., pp. 211–216, 2016.

[118] M. Tawfik et al., "Laboratory as a Service (LaaS): A model for developing and implementing remote laboratories as modular components," Proc. 2014 11th Int. Conf. Remote Eng. Virtual Instrumentation, REV 2014, no. February, pp. 11–20, 2014.

[119] D. W. McKee, S. J. Clement, X. Ouyang, J. Xu, R. Romanoy, and J. Davies, "The Internet of Simulation, a Specialisation of the Internet of Things with Simulation and Workflow as a Service (SIM/WFaaS)," Proc. - 11th IEEE Int. Symp. Serv. Syst. Eng. SOSE 2017, pp. 47–56, 2017.

[120] U. I. Vdedm et al., "A dististributed simulation platform for cloud computing," Soc. Model. Simul. Int., vol. 87, 2017.

[121] K. M. Abbasi, I. Haq, A. K. Malik, T. A. Khan, and A. C. C. Services, "Data Security in Cloud as a Service for Access Control Among Multilevel Users," Int. Conf. Commun. Technol., 2017.

[122] G. Sun, S. Member, T. Cui, J. Yong, S. Member, and J. Shen, "MLaaS : a Cloud - based System for Delivering Adaptive Micro Learning in Mobile MOOC Learning," vol. 1374, no. c, pp. 1–14, 2015.

[123] M. Ibrahim, M. A. Iqbal, M. Aleem, and M. A. Islam, "SIM-Cumulus : An Academic Cloud for the Provisioning of Network-Simulation- as-a-Service (NSaaS)," pp. 27313–27323, 2018.

[124] B. Du, R. Huang, Z. Xie, J. Ma, and W. Lv, "EDGE COMPUTING FOR THE INTERNET OF THINGS KID Model-Driven Things-Edge-Cloud Computing Paradigm for Traffic Data as a Service," IEEE Netw., vol. 32, no. February, pp. 34–41, 2018.

Applying Machine Learning Techniques for Classifying Cyclin-Dependent Kinase Inhibitors

Ibrahim Z. Abdelbaky[1]
Agricultural Research Center
Cairo,
Egypt

Ahmed F. Al-Sadek[2]
Agricultural Research Center, Cairo, Egypt, Computer Science Department, October University for Modern Sciences and Arts, MSA, October 6th, Egypt

Amr A. Badr[3]
Computer Science department, Faculty of Computers and Information, Cairo University
Cairo, Egypt

***Abstract*—The importance of protein kinases made them a target for many drug design studies. They play an essential role in cell cycle development and many other biological processes. Kinases are divided into different subfamilies according to the type and mode of their enzymatic activity. Computational studies targeting kinase inhibitors identification is widely considered for modelling kinase-inhibitor. This modelling is expected to help in solving the selectivity problem arising from the high similarity between kinases and their binding profiles. In this study, we explore the ability of two machine-learning techniques in classifying compounds as inhibitors or non-inhibitors for two members of the cyclin-dependent kinases as a subfamily of protein kinases. Random forest and genetic programming were used to classify CDK5 and CDK2 kinases inhibitors. This classification is based on calculated values of chemical descriptors. In addition, the response of the classifiers to adding prior information about compounds promiscuity was investigated. The results from each classifier for the datasets were analyzed by calculating different accuracy measures and metrics. Confusion matrices, accuracy, ROC curves, AUC values, F1 scores, and Matthews correlation, were obtained for the outputs. The analysis of these accuracy measures showed a better performance for the RF classifier in most of the cases. In addition, the results show that promiscuity information improves the classification accuracy, but its significant effect was notably clear with GP classifiers.**

Keywords*—CDK inhibitors; random forest classification; genetic programming classification

I. Introduction

Different important biological processes in the human body is related to the process of phosphorylation. In which, a phosphate group is added to proteins to activate their functionality. Protein kinases are enzymes that catalyze this process by adding the phosphate group to other proteins.

Due to the importance of the phosphorylation process in cellular processes and metabolism, protein kinases gained their importance and they are subject to many studies including drug design studies. Protein kinases are related to different diseases and cancer types when inappropriately regulated [1].

In humans, there are more than 500 kinases. They are divided into three types as they catalyze three types of phosphorylation [2].

Although kinases were targeted by several drug discovery studies that led to developing many inhibitors for them, only few of these inhibitors were approved. The reason for that is the undesired side effects caused by inhibitor reactivity against unintended targets. This is caused by the high degree of similarity between kinases, as there are few structural differences between them especially in their highly conserved binding domains. This similarity in binding domains led to the selectivity problem in many kinase inhibitors [2]. In most cases, this problem is caused by the high conservation is in the ATP binding site, which is the target for most of the inhibitors developed for kinases [3].

Among protein kinases, cyclin-dependent kinases (CDKs) are protein kinases, which have essential roles in cell divisions and transcription. CDKs are marked by being dependent on a protein subunit called cyclin to activate their enzymatic function. They belong to the serine/threonine kinase family [4].

Blocking the cell cycle by targeting kinases is proposed to kill cancer cells as in [5]. CDKs related to cell cycle are divided into three subfamilies, Cdk1, Cdk4, and Cdk5. The Cdk1 family consists of CDK1, CDK2, and CDK3 kinases. Although CDK1 is the most important kinase in this family because its major role for cell cycle, CDK2 is also essential as it participates in the cycle of cell division. In addition, CDK2 is investigated as being related to cancer and is targeted for cancer treatment as in [6].

CDK5 is an important enzyme that has different functions related to cell-cycle, gene expression, and others. CDK5 belongs to the cdk5 subfamily. In addition to its role in the cell-cycle progress, it is also known for controlling neuronal proteins [4]. CDK5 is also related to neurodegenerative diseases if was deregulated [7]. It is also linked to cancer and other diseases [8].

Computer-based approaches is being utilized in order to help profile the activity of different inhibitors against kinases and to explore and tackle the selectivity problem. Among these techniques is machine learning, which is widely utilized in biological and medical related problems. Different machine learning techniques were used in interaction modelling studies to predict protein-inhibitor interactions.

In [9] random forest was used to classify kinases variants in order to understand the relation with different diseases. The classification was based on protein kinases sequence features. The resulting accuracy was 88%.

In the area of kinase inhibitors, machine learning was used in [10] to a study the kinase inhibitory data of [11] in order to model the prediction of interactions between kinases and their inhibitors. The study aimed for building a computational-experimental framework by using Kernel-based regression methods on molecular descriptors and fingerprints of kinase inhibitors. The predicted results were found correlated to kinase assays experimental results by 0.77.

In [12] Machine learning for predicting the binding of kinases to inhibitors by modelling different sets of features. Features used for kinases are based on sequences, in addition to phylogenetic features and amino acid positions in the active site. For inhibitors, 2D structural features and chemical features were used. Their experiments showed the importance of different sets of features based on the decision tree and SVM modelling results. The highest prediction accuracy achieved was 86.1%.

Another application of machine learning to predict active or inactive confirmations of kinases was done in [13]. The study proved that classification based on the activation segment orientation is performing better than other methods.

Genetic Programming (GP) [14] is a machine learning technique that simulates biological evolution and is used for modelling by regression or classification. It starts by a random population, then it continues to produce generations and individuals by performing evolutionary operations such as mutations, crossover, and selection, aiming to improve a fitness function. The individuals of GP is trees representing mathematical models to relate the modelled features to a target variable [15].

Random Forest (RF)[16], is a machine learning technique based on a large number of decision trees. A bootstrap sample is drawn and a set of variables are selected randomly to decide the split of each node. The tree grows and splits using the variables at each node until a specified criteria is achieved [17].

In this study, we use genetic programming and random forest classification techniques for classifying inhibitors and non-inhibitors for two of the cyclin-dependent kinases, CDK5 and CDK2. Both techniques were used for modelling chemical descriptors information. In addition to classification, we investigated the response of the classifiers to adding information about kinase binding promiscuity of compounds.

The outputs of the classifiers were analyzed using different accuracy measures and metrics. Because there is no standard single evaluator of classification accuracy, we calculated and obtained a group of measures for a wide evaluation of the results. These measures are confusion matrices, accuracy, ROC curves, AUC values, F1 score, and Matthews correlation coefficient.

Additionally, the analysis shows how could the classifiers reflect compound promiscuity information. Compound *promiscuity* against kinases is the ratio of the kinases that could be inhibited by that compound at a specific concentration [11].

This document is structured as follows: In section 2 we describe the dataset we used and illustrate data processing and workflow steps. In section 3, different results are presented and discussed. In section 3, we present our conclusion on the results and expectations for future improvements.

II. Data and Methods

We describe in this section the data sources, tools, and the methodology we used in order to achieve our objectives in building and evaluating the classifiers.

A. Data Sources

The dataset we used was extracted from the data of [11]. The original dataset contains the measured interaction values for more than 3000 compounds against 172 kinases. The values represent the pK_I values, which are the negative values of base-10 log of the K_I interaction value. We extracted the values for the first 1497 compounds against two protein kinases belonging to the cyclin-dependent kinases subfamily, CDK2, and CDK5.

The original dataset contains five cyclin-dependent kinases. We decided to study the data for CDK2, and CDK5 only as they have higher number of measured inhibitor activities with 868 and 1038 values respectively.

A threshold pK_I value of (value >5.9) was used for classifying compounds as inhibitors or not. This threshold was determined based on what the original study in [11] mentioned about compound activity against kinases.

We used the molecular descriptor values for the 1497 compounds. These values were obtained previously using e-dragon online tool [18]. The number of descriptors extracted for each compound is 1666 descriptor values.

Promiscuity value for each compound is provided in the original dataset in [11]. Promiscuity_1uM of a compound represents the portion of kinases tested with a potency of 1uM achieved by that compound.

B. Data Preparation

For each of the two proteins, two files were created with all the information needed for modelling. Each file contained the descriptor values for the compounds that interacted with one protein after removing columns that contained only zeros for all compounds. In addition, the interaction values were added as the last column as the target value. For each protein, another version of the data file was created including the value of *promiscuity_1uM* for each compound as an additional feature. So, each of the two proteins had two data files, and building a classifier was done twice for each protein. One time with descriptor values only, and another time with promiscuity and descriptor values.

The interaction values were classified based on the threshold mentioned in [11], considering the value of 5.9 pK_I as the inhibition threshold. The data file for each protein was modified replacing the interaction value with the class number. Class 1 represents that the corresponding compound is a potential inhibitor (pK_I > 5.9) while class 2 represents a non-inhibitor compound (pK_I <= 5.9). Table I shows the counts of compounds as inhibitors or non-inhibitors according to the specified threshold in both protein datasets.

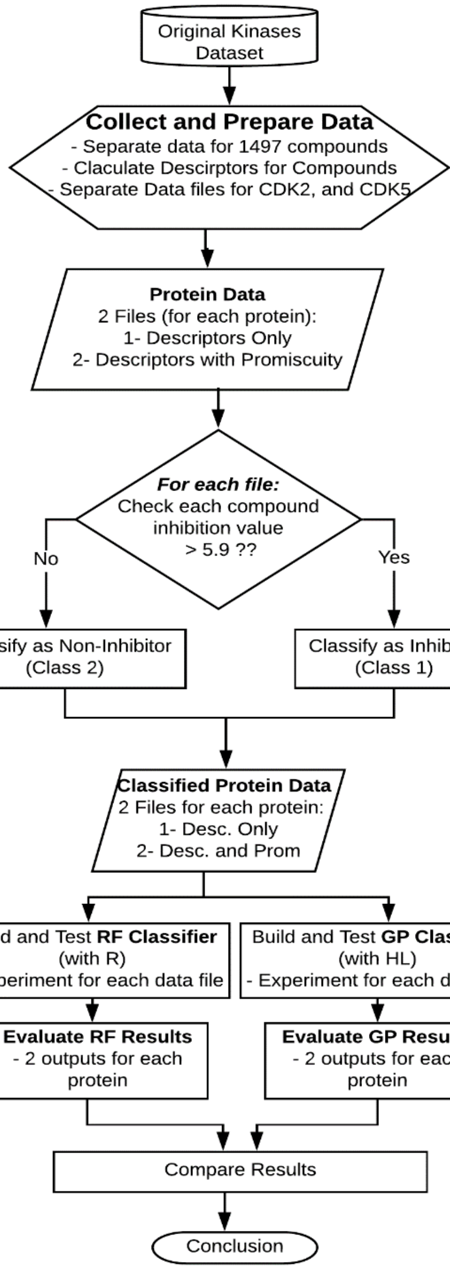


Fig. 1. Complete Workflow for Buidling and Eavaluating Classifiers.

Data separation, processing and cleaning were all done using python scripts that we wrote to read, manipulate and write csv files with the desired structure.

C. Methodology

In this study, we followed a multi-step methodology to explore the ability of two machine-learning techniques for predicting the inhibition activity of compounds against two cyclin dependent kinases. Random forest and genetic programming were applied on datasets of CDK5, and CDK2 kinases. The effect of adding promiscuity to the modelling features was also investigated, as it provides information about the ability of a compound to inhibit kinases.

First, we obtained, separated, and preprocessed the data sets for CDK5 and CDK2 kinases. After that, we obtained and prepared the required tools for testing random forest, and genetic programming classifications. Then, we performed different experiments, namely four for each protein, and collected the outputs. Finally, we evaluated the performance of the classifiers with different measures and compared the results. We concentrated more on the outputs of the RF classifier. The workflow of the complete steps for our work is shown in Fig. 1, which shows the steps followed to build both classifiers for each protein dataset.

In all experiments, descriptor values were considered as variables or features, and the class number was the target to be predicted. Each dataset was divided into a 70% training set, and a 30% test set.

D. Genetic Programming Classification

To perform GP classification we used a free desktop tool, HeuristicLab Optimizer 3.3.15 [19], in the mode (symbolic classification).

The input in each GP experiment was one of the files we created previously, in addition to setting GP parameters. We used different combination of parameters trying to achieve higher accuracy. The set of parameter values used with GP experiments are shown in table II.

TABLE I. Number of Compounds in ecah class

	Inhibitors	**Non-Inhibitors**
CDK5	234	804
CDK2	251	618

TABLE II. GP CLASSIFICATION PARAMETER VALUES

Population size	1000
No. of generations	1000 - 5000
Selection method	Tournament Selector
Crossover rate	0.90
Mutation rate	0.15
Objective	Min (MSE), Or Min (Penalty Score)
Model Depth	10
Model length	100

E. Random Forest Classification

Data files for each protein were loaded into R studio. Each dataset was divided by random sampling into a 70% training set, and a 30% testing set for validation.

The R package *(randomForest)* was used for the modelling. We set two basic RF parameters, the number of trees constructed *(ntree)*, and the number of randomly preselected features, or variables, in each tree *(mrty)*. We tried different values for these two parameters until we achieved a relatively low error value. Parameter values used for RF experiments are shown for different datasets in table III. The parameter values used with the datasets including promiscuity are also shown in the table. The clear difference when using promiscuity in the dataset was achieving higher accuracies with lower number of trees.

Finally, results of different experiments were collected and different accuracy metrics were calculated for enhanced analysis.

TABLE III. RF CLASSIFICATION PARAMETER VALUES

	Promiscuity	Trees *(ntree)*	Variables *(mrty)*
CDK5	No	600	70
	Yes	450	110
CDK2	No	600	85
	Yes	450	65

III. RESULTS AND DISCUSSION

Both machine-learning classification techniques, genetic programming and random forest, were tested on two datasets for two cyclin dependent kinases and their inhibitors. Results varied between datasets and techniques. We mention the results in this section showing different accuracy measures we used, along with a discussion of the variations in these accuracies.

A. Accuracy

RF classifier could classify all test data. Table IV shows the overall accuracy of the RF classifier in terms of all correctly classified items in training and testing sets for both proteins. The accuracy is also shown when promiscuity was used in the data set.

B. Confusion Matrix

A confusion matrix is a table with a specific layout that is usually used to describe and visualize the performance of a classification algorithm. We show here the confusion matrices for each experiment.

The confusion matrices resulted from RF experiments are shown in tables V to VIII. Table V shows the matrix for the RF Result Accuracies for both proteins with and without promiscuity.

TABLE IV. RF RESULT ACCURACIES FOR BOTH PROTEINS WITH AND WITHOUT PROMISCUITY

	Promiscuity	Accuracy	
		Training	*Test*
CDK5	No	83.47 %	83.97 %
	Yes	84.44 %	85.26 %
CDK2	No	83.22 %	80.08 %
	Yes	85.53 %	88.51 %

In all confusion tables, the columns show the number of predicted items in each class (Inhibitor, Non-Inhibitor), and the rows display the actual items in each class. The results are shown for training and testing sets.

Table V shows results from CDK5 dataset without promiscuity, while table VI shows the results for CDK5 dataset including promiscuity information. Similarly, for CDK2 dataset, tables VII shows the results without promiscuity information, while table VIII displays the confusion matrix of CDK2 dataset that includes promiscuity values.

It should be noticed that the test sets were selected by random sampling, so, number of items in each class will not remain the same among different experiments.

The confusion matrices in all experiments show a high ability of the RF classifier to identify non-inhibitors. On the other hand, the ability to identify inhibitors is not in the same level. The reason for that could be the imbalance in data provided for the classifier, as most of the compounds in the data sets are already non-inhibitors as shown in table I.

TABLE V. RF CLASSIFIER CONFUSION MATRIX FOR CDK5 INHIBITORS (DESCRIPTORS ONLY)

		Predicted			
		Training		***Testing***	
		Inhibitor	*Non-Inhibitor*	*Inhibitor*	*Non-Inhibitor*
Actual	*Inhibitor*	52	109	28	45
	Non-Inhibitor	11	554	5	234

TABLE VI. RF CLASSIFIER CONFUSION MATRIX FOR CDK5 INHIBITORS (DESCRIPTORS AND PROMISCUITY)

		Predicted			
		Training		***Testing***	
		Inhibitor	*Non-Inhibitor*	*Inhibitor*	*Non-Inhibitor*
Actual	*Inhibitor*	57	103	30	44
	Non-Inhibitor	10	556	2	236

TABLE VII. RF CLASSIFIER CONFUSION MATRIX FOR CDK2 INHIBITORS (DESCRIPTORS ONLY)

		Predicted			
		Training		***Testing***	
		Inhibitor	*Non-Inhibitor*	*Inhibitor*	*Non-Inhibitor*
Actual	*Inhibitor*	82	87	36	46
	Non-Inhibitor	15	424	6	173

TABLE VIII. RF CLASSIFIER CONFUSION MATRIX FOR CDK2 INHIBITORS (DESCRIPTORS AND PROMISCUITY)

		Predicted			
		Training		***Testing***	
		Inhibitor	*Non-Inhibitor*	*Inhibitor*	*Non-Inhibitor*
Actual	*Inhibitor*	100	75	49	27
	Non-Inhibitor	13	420	3	182

C. ROC Curves

The Receiver operating Characteristics curve (ROC Curve) was plotted for all outputs to understand the ability of each classifier in discriminating between inhibitors and non-inhibitors. RF ROC curves were plotted using ROCR package in R [20], and are shown in Fig. 2, while ROC curves for GP experiments were obtained from HeuristicLab, and are shown in Fig. 3.

The curves show a fairly high ability for the RF classifier to label and determine the class for test data. For additional better understanding of the ROC curves, we show the values of AUC (Area Under the ROC Curve) for these ROC curves in table IX.

From the AUC values and the ROC curves, we can see that RF outperforms GP with both protein datasets, especially when promiscuity information exists. The AUC values also show a remarkable improvement when promiscuity information exists in the dataset for both proteins and with the two techniques. However, the improvement ratio in the case of promiscuity information is notably higher with GP classifier.

D. F1 Score

F1score is calculated based on the precision and recall measures. F1 score measures the accuracy of a classification model based on the number of positives identified correctly and the total number of positives. Tables X and XI show the F1 scores for RF and GP classifiers on CDK5 and CDK2 dataset respectively.

For CDK2, F1 scores are almost within a close range to each other except for GP classifier without promiscuity. Also in this measure, we can see that GP could better reflect the promiscuity information by increasing the F1 score value with a higher ratio than RF, although RF values were better beforehand. Additional note here is that CDK5 dataset without promiscuity could not result in high accuracy predictions of positives, even after many experiments.

E. Matthews Correlation Coefficient

Matthews correlation coefficient (MCC), is a quality measure used to evaluate binary classifications. So it is applicable in our case. It takes into consideration true positives and negatives and hence it is considered as a balanced measure. Tables XII and XIII show the MCC values for RF and GP classifiers on CDK5 and CDK2 datasets respectively.

TABLE IX. AREA UNNDER ROC CURVE FOR RF AND GP CLASSIFIERS

	AUC			
	CDK5		***CDK2***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.82	0.94	0.87	0.94
GP	0.56	0.85	0.62	0.68

TABLE X. F1 SCORES FOR CDK5 DATASET RESULTS

	F1 Score			
	Training		***Testing***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.46	0.50	0.53	0.57
GP	0.62	0.70	0.23	0.51

TABLE XI. F1 SCORES FOR CDK2 DATASET RESULTS

	F1 Score			
	Training		***Testing***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.62	0.69	0.58	0.77
GP	0.66	0.88	0.39	0.43

TABLE XII. MCC SCORES FOR CDK5 DATASET RESULTS

	MCC			
	Training		***Testing***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.45	0.49	0.50	0.56
GP	0.53	0.62	0.11	0.42

TABLE XIII. MCC SCORES FOR CDK2 DATASET RESULTS

	MCC			
	Training		***Testing***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.55	0.63	0.51	0.71
GP	0.54	0.83	0.18	0.33

The values of MCC measure in general ranges between -1 (No prediction), and 1 (Perfect prediction). In this case, the values for MCC in both datasets almost near to 0.5 or higher, except in the cases where GP classifier predicts the test sets for both proteins. As a general note, GP is performing better than RF in training data, but it cannot predict test sets accurately. On the other hand, RF is more accurate in predicting the test sets classes.

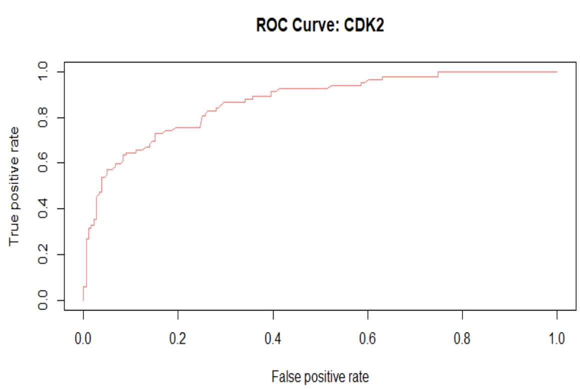
It is also clear from the tables that MCC value increases when promiscuity information is included in the datasets. Promiscuity information improved the accuracy of GP classifier more than its improvement for RF classifier on the training set level. This improvement is clearly noticeable in GP results for the test sets, although GP accuracy is still low on test sets compared to RF.

F. Important Vairables

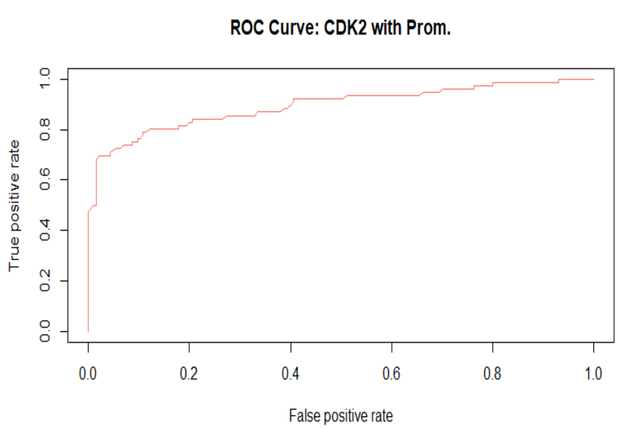
RF has the ability to rank different available considered while training. So, it can produce a list in each experiment with the most important variable affecting the prediction results. In table XIV we show a portion of the top important variables in the two experiments for each protein. We selected these important variables that had high rank in RF ranking for both mean decrease accuracy, and mean decrease Gini, and appeared with each protein in its corresponding two experiments. Variables names represent chemical descriptors produced by e-dragon.

TABLE XIV. IMPORTANT VARAIBLES AS SELECTED BY RF

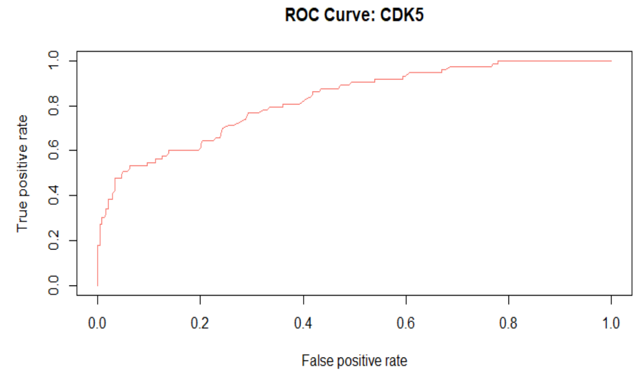
CDK5	**CDK2**
MATS7p	Mor32m
MAXDP	MATS1v
Cl-090	MATS1p
MATS7v	Mor18m



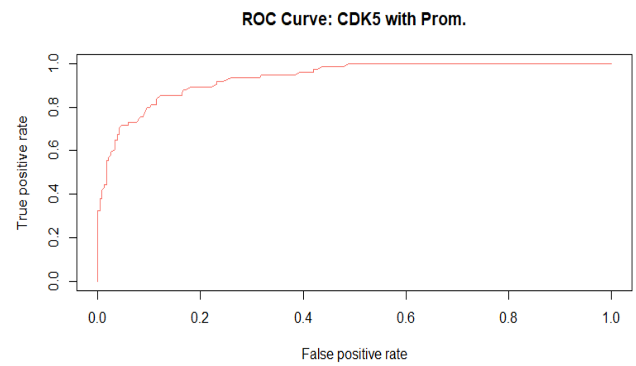
(a) CDK2 without Prmiscuity.



(b) CDK2 with Promiscuity.

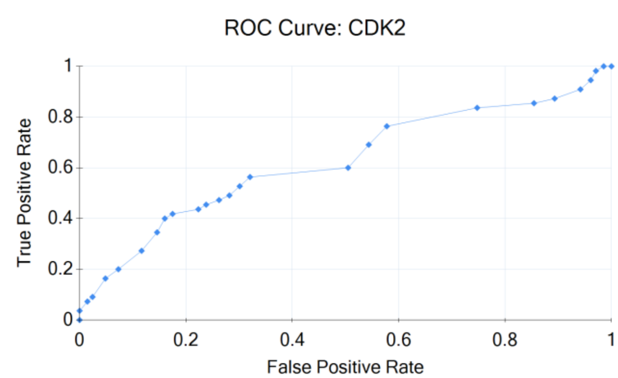


(c) CDK5 without Promiscuity.

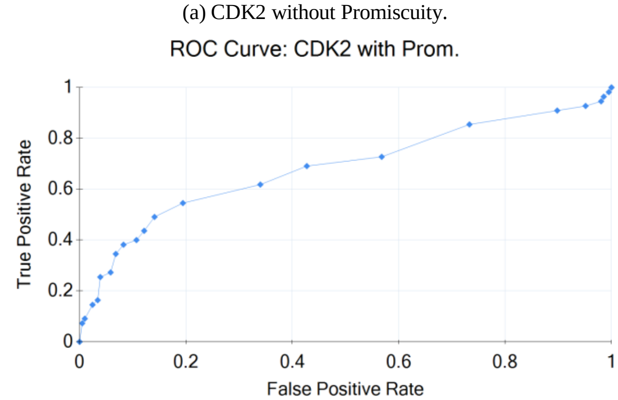


(d) CDK5 with Promiscuity.

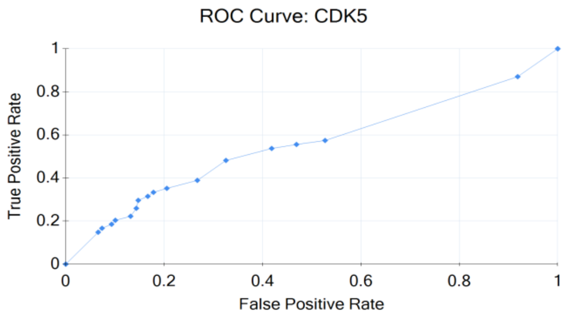
Fig. 2. RF Results ROC Curves for both proteins.



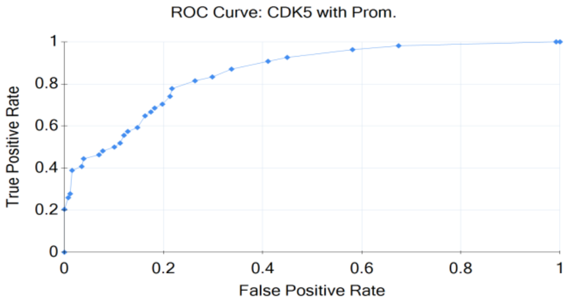
(a) CDK2 without Promiscuity.



(b) CDK2 with Promiscuity.



(c) CDK5 without Promiscuity.



(d) CDK5 with Promiscuity.

Fig. 3. GP Results ROC Curves for both Proteins.

IV. Conclusion

Machine learning techniques provides a useful means to model and understand kinase-inhibitor interaction data. Although the results were not usually of high accuracy for different accuracy measures, but still there are many measures showing promising values and representing good predictions. Machine learning classifiers produced good predictions for the class with more data in the dataset, non-inhibitor class. We suppose that this could be a result of imbalanced data distribution.

Another important conclusion is the ability of the classifiers to response effectively to one feature reflecting its importance. Kinase inhibitors are likely to bind to more than one kinase. The improvement in predictions when compound promiscuity is added to the features means that it was efficiently modeled. This suggests that adding more features such as protein binding site properties could highly improve the prediction accuracy.

Compared to previous work using different techniques mentioned in section 1, our results achieved promising values in terms of overall accuracy. The average overall accuracy from RF experiments was about 85%, which is comparable to the 88% in [9], and 86% in [12]. Most of previous work tried to predict kinase inhibitors for the whole family, while in our work here we concentrated on the CDK subfamily to be more specific and more responsive to any special binding features of CDKs. We expect that extending the data by adding more features and considering protein-related properties on different levels would improve the classification accuracy.

Finally, it is not necessarily that a good performing technique to be usually the most sensitive one for new features. Different approached should be tired with different datasets and features with a comprehensive and accurate evaluation of the results.

References

[1] S. Enna and D. Bylund, XPharm: The Comprehensive Pharmacology Reference. Amsterdam: Elsevier, 2008.

[2] H. Park, K. Kim, C. Kim, J. Shin and K. No, "Descriptor-Based Profile Analysis of Kinase Inhibitors to Predict Inhibitory Activity and to Grasp Kinase Selectivity", Bulletin of the Korean Chemical Society, vol. 34, no. 9, pp. 2680-2684, 2013.

[3] Y. Hu, N. Furtmann and J. Bajorath, "Current Compound Coverage of the Kinome", Journal of Medicinal Chemistry, vol. 58, no. 1, pp. 30-40, 2014.

[4] M. Malumbres, "Cyclin-dependent kinases", Genome Biology, vol. 15, no. 6, p. 122, 2014.

[5] W. Taylor and A. Grabovich, "Targeting the Cell Cycle to Kill Cancer Cells", Pharmacology, pp. 429-453, 2009.

[6] M. Sayour, I. Basheer, R. Salam, M. Yamany, A. Badr, A. Al-sadek and R. El-awady, "Potential mechanistic profiling of an otc analgesic as a cytotoxic agent in the treatment of hepatocellular carcinoma", ACTA Pharmaceutica Sciencia, vol. 56, no. 1, p. 95, 2018.

[7] Z. Cheung and N. Ip, "Cdk5: a multifaceted kinase in neurodegenerative diseases", Trends in Cell Biology, vol. 22, no. 3, pp. 169-175, 2012.

[8] A. Arif, "Extraneuronal activities and regulatory mechanisms of the atypical cyclin-dependent kinase Cdk5", Biochemical Pharmacology, vol. 84, no. 8, pp. 985-993, 2012.

[9] T. Pons, M. Vazquez, M. Matey-Hernandez, S. Brunak, A. Valencia and J. Izarzugaza, "KinMutRF: a random forest classifier of sequence variants in the human protein kinase superfamily", BMC Genomics, vol. 17, no. 2, 2016.

[10] A. Cichonska, B. Ravikumar, E. Parri, S. Timonen, T. Pahikkala, A. Airola, K. Wennerberg, J. Rousu and T. Aittokallio, "Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors", PLOS Computational Biology, vol. 13, no. 8, p. e1005678, 2017.

[11] Metz, E. Johnson, N. Soni, P. Merta, L. Kifle and P. Hajduk, "Navigating the kinome", Nature Chemical Biology, vol. 7, no. 4, pp. 200-202, 2011.

[12] F. Buchwald, L. Richter and S. Kramer, "Predicting a small molecule-kinase interaction map: A machine learning approach", Journal of Cheminformatics, vol. 3, no. 1, p. 22, 2011.

[13] D. McSkimming, K. Rasheed and N. Kannan, "Classifying kinase conformations using a machine learning approach", BMC Bioinformatics, vol. 18, no. 1, 2017.

[14] Koza, Genetic programming. Cambridge, Mass.: MIT Press, 1994.

[15] M. Tan, J. Tan, S. Chang, H. Yap, S. Abdul Kareem and R. Zain, "A genetic programming approach to oral cancer prognosis", PeerJ, vol. 4, p. e2482, 2016.

[16] L. Breiman, “Machine Learning,” Machine Learning, vol. 45, no. 3, pp. 261–277, 2001.

[17] A. Boulesteix, S. Janitza, J. Kruppa and I. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 6, pp. 493-507, 2012..

[18] Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. Palyulin, E. Radchenko, N. Zefirov, A. Makarenko, V. Tanchuk and V. Prokopenko, "Virtual Computational Chemistry Laboratory – Design and Description", Journal of Computer-Aided Molecular Design, vol. 19, no. 6, pp. 453-463, 2005.

[19] A. Elyasaf and M. Sipper, "Software review: the HeuristicLab framework", Genetic Programming and Evolvable Machines, vol. 15, no. 2, pp. 215-218, 2014.

[20] T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer, "ROCR: visualizing classifier performance in R", Bioinformatics, vol. 21, no. 20, pp. 3940-3941, 2005.

Bound Model of Clustering and Classification (BMCC) for Proficient Performance Prediction of Didactical Outcomes of Students

Anoopkumar M[1*]
Research and Development Centre,
Bharathiar University,
Coimbatore – 46, Tamilnadu, India

A. M. J. Md. Zubair Rahman[2]
Principal, Head of the Institution,
Al-Ameen Engineering College,
Erode, Tamilnadu, India

***Abstract*—In this era of High-Performance High computing systems, Large-scale Data Mining methodologies in the field of education have become a convenience to discover and extract knowledge from Databased of their respective educational archives. Typically, all educational institutions around the world maintain student data repositories. Attributes of students such as the name of the student, gender of student, age group (date of birth), religion, eligibility details, academic assessment details, etc. are kept in it. With this knowledge, in this paper, didactical data mining (DDM) is used to leverage the performance prediction of student and to analyse it proactively. As it is known, Classification and Clustering are the liveliest techniques in mining the required data. Hence, Bound Model of Clustering and Classification (BMCC) have been proposed in this research for most proficient educational data mining. Classification is one of the distinguished options in Data Mining to assign an object under some pre-defined classes according to their attributes, and hence it comes under a supervised learning problem. On the other side, clustering is considered as a non-supervised learning problem that involves in grouping up of objects with respect to some similarities. Moreover, this paper uses the dataset collected from Kerala Technological University-SNG College of Engineering (KTU_SNG) for performing the BMCC. An efficient J48 decision tree algorithm is used for classification and the k-means algorithm is incorporated for clustering here and is optimised with Bootstrap Aggregation (Bagging). The implementation has been done and analysed with a data mining tool called WEKA (Waikato Environment for Knowledge Analysis), and the results are compared with some most used classifications such as Bayes Classifier (NB), Neural Network (Multilayer Perceptron MLP) and J48. It is provable from the results that the model, proposed in this provides high Precision Rate (PR), accuracy and robustness with less computational time, though the sample data set includes some missing values.**

***Keywords*—Classification; clustering; precision rate; accuracy; j48 decision tree; bagging; educational data mining**

I. Introduction

In general, data mining is the process of examining the large databases for extracting the new or required information. It can also be stated that it is the procedure of effective classification of folders with respect to the specified data patterns that are acquired from the dataset. There are enormous algorithms have been developed for retrieving the valuable information and knowledge discovery patterns, which is functional for decision support. In another word, data mining can also be known as KDD, which is Knowledge Discovery in Database, handles with the non-trivial retrieval of inherent, completely novel and valuable information from the databases [4]. The Fig. 1 contains the typical steps involved in the data mining process of retrieving valuable information.

In the present decade, there are rising research scopes in utilizing data mining techniques for education, henceforth for Didactical Data Mining (DDM) a similar one to Educational Data Mining (EDM). This newly developing interdisciplinary field EDM involves knowledge extraction from the data obtained from the educational environments [11]. The main intention is to have a good understanding about the students learning standards and identifying the procedure in which they study to enhance the educational results, to process with the obtained outcomes based on the educational phenomena. Moreover, the educational information system can store a wide range of prospective data that would be collected from the historical and multiple sources exist in the databases of distinctive educational institutions. The collected data are from different sources, formats and also at variant granularity levels and that may contain the personal or academic details of the students. In another option, the huge data can also be collected from the e-learning systems that are already provided with a huge amount of data from various institutions. For handling those data effectively, there is a need for effective implementation of an EDM technique. The following Fig 2 depicts the basic attributes that have to be combined with the Didactical Data Mining (DDM) or EDM system in addition to the typical data mining functions.

The main goal of the educational institutes is to afford a good quality of education and to enhance the power of managerial strategies. In order to accomplish the highest level of educational domain, analysing the major attributes of student's performance and discovering knowledge is a significant part. This can provide useful and beneficial recommendations to the academicians and the management to improve their decision-making structure, the performance of the students and teaching abilities of the tutors or faculties [7].

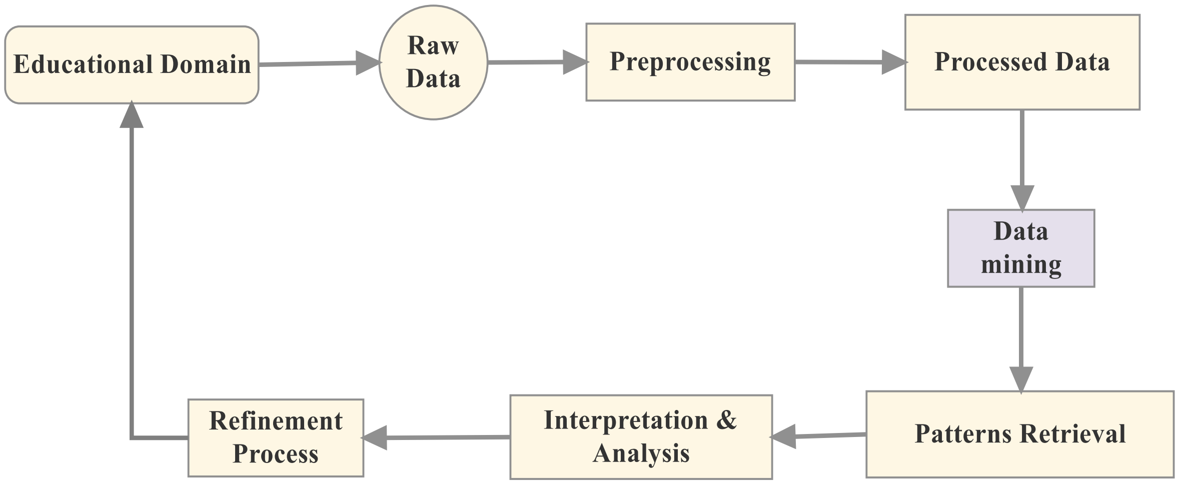


Fig. 1. Functions Involved in Typical Data Mining.

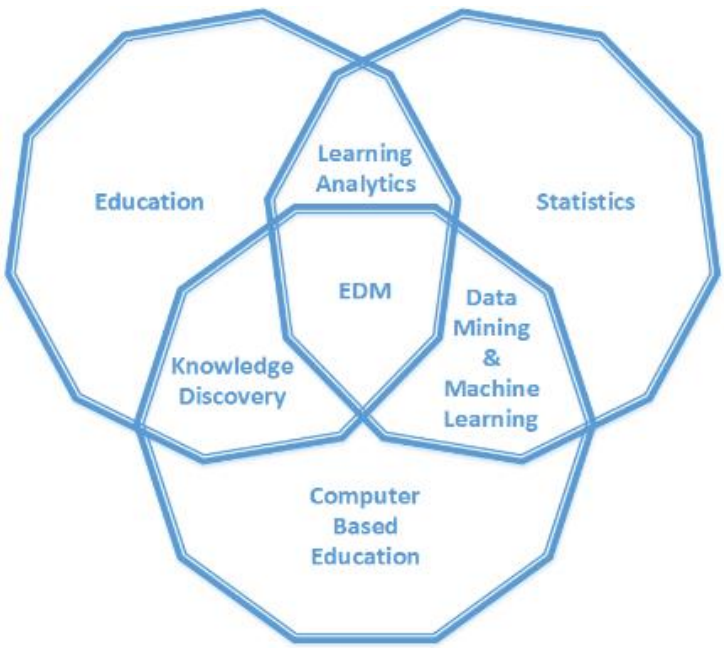


Fig. 2. Basic Factors of DDM/ EDM.

In general, EDM uses many data mining algorithms like decision tree algorithm, neural networks, rule induction, naive Bayesian and so on. On using these methodologies, the powerful knowledge can be retrieved using classification, association rules and clustering. In this paper, the main objective is about the combination of Classification and Clustering, the well-known model for enhancing the precision of the results. The framed work Bound Model of Clustering and Classification (BMCC), concentrates on the classification of yearly-growing data sets using J48 decision tree classifier since it is comparably faster and more exact than the other classification techniques. And also, a decision tree can be transformed into simple and understandable classification rules. Furthermore, clustering is the course of grouping up of similar objects and that comes under the unsupervised pattern classification and K-means algorithm is incorporated for clustering. In this paper, an efficient machine learning tool called WEKA (Waikato Environment for Knowledge Analysis) has been used for the implementation of classification and clustering algorithms. The dataset acquired here is the KTU_SNG student dataset that contains 232 samples with 60 attributes for each. The proposed (BMCC) model is implemented with this data set in WEKA environment for producing précised results to analyse and monitor the student’s performance and take organisational decision for the betterment of students.

The leftovers of this paper are systematized as follows: Section 2 describes the problem statement in short. Section 3 considers the related works based on Educational Data Mining. Section 4 presents the operations involved in the proposed (BMCC) model. The results and discussions are given in section 5 and finally, section 6 concludes the work with some key points for future enhancement.

II. Problem Statement

The problem is defined specifically designing a model for predicting the performance of KTU Students. The work [21] carried out in the field of EDM have identified the use of a Tuned J48. There were studied which used clustering [6] along with classification to enhance its effectiveness [23]. Here in this, it is enquiring the scope of these combinations in predicting student performance in KTU by combining the Tuned J48 classification algorithm with the K-Means. The bound model has been analysed with various factors referred in section IV and using the KTU_SNG student dataset containing 232 samples of with 60 attributes each (comprises both personal information and academic performance). This is a research enquiry about how best the model can serve the expectation of stakeholders since it is new University established in the year 2015 and the status of Datasets are also in an infant stage.

III. Related Works

Because of the potentials of data mining to educational domains, it has become the most efficient research area. There are so many works have been done based on this. In [1] a valuable survey work on data mining has been provided. The authors have discussed the EDM as a hopeful research [16] field and some particular advancement that are not provided by other fields. In another work [12], a case study is provided, in which educational data mining has been used to examine the learning behaviour of students. The main intention of the study was to describe the significance of the EDM in enhancing student performance in higher secondary education. The authors have used the data from the warehouse and the gathered data comprises both the personal and academic records of the students. Data and records gathered up from e-learning management systems (e-LMS) have also been incorporated. The described the study work on the source of applying some classification techniques using decision tree algorithms and association rules for obtaining different kinds of knowledge. All outliers had been detected for analysis and finally, the student's records are presented with the knowledge discovered to enhance the overall performance of the students. A valuable and efficient review work has been done by the authors in [16]. The paper has discussed various concepts and techniques involved in EDM. Moreover, the comprehensive analysis has been made for the methodologies used for faculty and the student performance evaluation, which helps the institution.

In [2], it is stated that the classification techniques are included to help in enhancing the standard of the educational system by analysing the student data to handle the main factors that may have great change on the student performances in specializations. The classification rules are derived on the basis of decision tree making, and then the derived rules are examined and evaluated. It permits the students to establish the final grade under specializations. Another work in [5], also discussed the classification in data mining for student's performance evaluation. The contribution of the work was to extract the information from the database that defines the status of the students at the end of all semesters. The collected student information includes data about class test performances, assignments, seminars and attendances. The analysis helps in the identification of students who require special care and need some counselling from the mentors. Likewise, the concept provided in [3] involved in the prediction of student's enrolment based on the academic data using classification rules. The derived rules are evaluated by various methods and permit the university management to organize the required resources for the newly enrolled students. The process aids in providing the earlier information about the category of the students going to be admitted and the areas to be concentrated in their specializations to maintain them in a better way. WEKA tool has been used for the implementation and examination of classification techniques such as Rule-based classification, Naive Bayes and Decision Tree [17]. The paper involved in obtaining efficient results, using classification in required data extraction from large databases. The process is limited to handle the missing values and incomplete data in the collection.

In [6], the authors explained the J48 decision tree classification algorithm clearly. Moreover, the decision tree has been used for classifying the diabetes person's data. They have also discussed the basic steps involved in decision tree algorithm and computations. In a variant form, the performance of the instructor has been evaluated in [20]. Initially, the analysis has been made with four classifying techniques such as Support Vector Machine, Decision Tree algorithms, Artificial Neural Networks and discriminant analysis for developing an efficient classifier. The results have been stated that the decision tree algorithm provided more précised classification. A Quality model has been developed in [18] to process better classification and student performance examination. The author has done a great work by implementing various data mining technique to develop a predictive model for categorizing students based on their study rate and personal information. CART (Classification and Regression Tree), ID3 algorithm, CHAID (Chi-Squared Automatic Interaction Detection), and C4.5 all these four decision tree methods have been used for classification and the detailed comparative analysis has been made with that.

In order to categorize the failure patterns of students, the authors of [8] used the mining process based on association rules. The major contribution of the work was to detect the background connection between the failed courses and the equivalent course suggestions for improving the performance of low-grade students. The association rules derived from the algorithm provide some hidden relationships about the failed courses which could make a base knowledge for academic designers in taking academic choices. This could pave a way for modification and re-construction of curriculum for the improvement of student's study rate and decrease the failure rates. K-means clustering algorithm has been utilized by the authors in [10], for determining the learning activities of the students that may include the class quizzes, internal examination results and assignments. There clustered data will be provided by their respective tutors prior to the commencement of final examination. This may help the tutors to decrease the fail ratio of the students by taking corrective actions at the right time to help the student's progression.

In [9], ID3 and C4.5 classification algorithm are used for the performance prediction of students. They have also discussed the missing data issues in classification. The causes for non-availability of data are given as follows:

- Malfunction of Equipment's
- Deletion due to a discrepancy with other stored information
- No-data enrolment because of some misconception
- No-data registration

Moreover, in [13], a valuable case study has been provided by analysing the classification based data mining techniques for the application-oriented analysis of education sectors. The academic data sets are examined with the classification algorithms such as Naive Bayes (NB), J48, Multilayer Perceptron (MLP), SMO (Sequential Minimal Optimization) and REPTree (Reduced Error Pruning Decision Tree) by using the WEKA tool. A research work has been discussed about the various constraints influenced in student performance evaluation and grade computation for alerting the students for final examinations [19]. Moreover, the developed methodology is mainly used for grade calculation that paves a way for developing the coaching techniques to the students.

In [15], the performances of undergraduate and PG students of two different institutions are utilized in decision tree classification and Bayesian Network algorithms. This aids to identify the students for scholarships and also for special care. The comparative results concluded that the decision tree algorithm provides results with more precision than the Bayesian Network algorithm. In another work [14], a case study has been performed with 300 student samples that are obtained from the Punjab University. The study results in finding a dignified correlation in some factors such as parent's educational factors, the income of the family and the performance of individual students.

As classification as a prominent of techniques identified in [16], in the [21] the work is taken J48 as the prerogative of research over educational data and to have an extended analysis on *Naïve Bayes, Bayes Net, Multilayer Perceptron, SVM, REPTree and Random Forest* classifications with a *Tuned J48.* And the proposed model has shown overall betterments among the other ones over the datasets used. The study generates a realisation that a varied and better implementation of J48 can make differences in the performance prediction in a realistic environment.

IV. Proposed Model

Through the valuable review of previous works and discussions on EDM for analysing student's performance, a number of attributes influenced in the process are being identified and defined effectively. Those attributes are characterized as input variables. For this concern, the recent real-time data called KTU_SNG student dataset is collected from the Engineering College. Some manual techniques are used for figure out the data and the data is transformed into a format that is feasible for processing in WEKA tool. Following that, the selection of parameters and features are taken into consideration. Moreover, it is to be stated that the data set contains 13920 records (i.e. 232 samples with 60 attributes). Table I shows the sample of the data.

The Fig 3 portrays the generic data mining methodology that comprises data acquisition, data pre-processing, data mining and pattern extraction processes. Initially, data pre-pre-processing has been done with the acquired data from the database before providing the data set for the mining process (i.e. removing irrelevant attributes). For an instance, some factors like caste or category, etc. are not required for analysing the intellectual performance of the students, since they come under the personal data. Following that, the missing values are removed from the data set in order to reduce the complexities on mining. The overall motive is to categorize Nature_of_Student under the categories such as Outstanding, Excellent, Good, Average, low and very low. The following Table I presents some sample attributes and their descriptions along with the domain values obtained from the source database.

A. *Bound Model of Clustering and Classification (BMCC Model)*

Classification is described as the process of identifying a set of models that determined and differentiate classes and concepts of data for using the mining model to detect the unknown classes. On the other end, clustering is examined with the similarities between the features present in the objects. The following Fig 4 depicts the overall framework for the combined function of clustering and classification. The Bond Model of Clustering and Classification, (BMCC) model works effectively even if the dataset has some missing values. Furthermore, the block diagram presents the steps involved in the evaluation and comparative analysis to identify the Nature_of_Student and Performance for enhancing the overall academic performance of the institution.

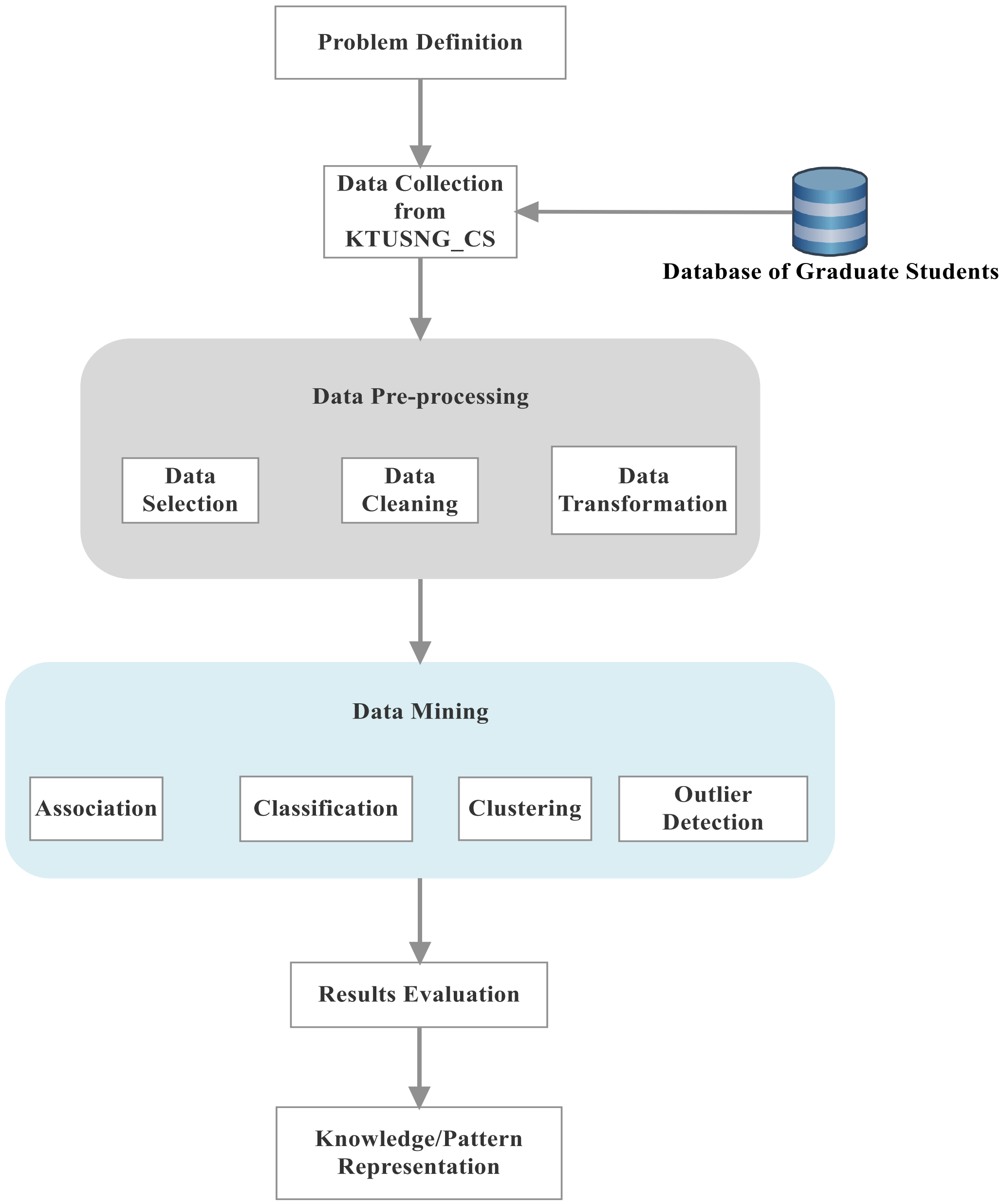


Fig. 3. The Flow of Generic Data Mining Methodology.

TABLE I. SAMPLE ATTRIBUTES OF THE ACQUIRED DATASET WITH DESCRIPTIONS AND DOMAIN VALUES

ATTRIBUTES	DESCRIPTION	DOMAIN VALUES
Gender	Student's sex	{M,F}
DOB	Date of Birth	Varying for samples
Mother Tongue	For language fluency	{Malayalam, Tamil}
Economically Backward	The financial status of the family	{True, False}
Handicapped	Health-Based analysis	{True, False}
Admission Type	Type of Admission	{Regular, Lateral}
Branch	Department of the student	{CS, CE, ME, ECE, EEE, NASB}
Attendance	Regularity	Based on the attendance count
Total Credits	Marks obtained	Given in percentage
Result	Calculated from marks	{PASS, FAIL}

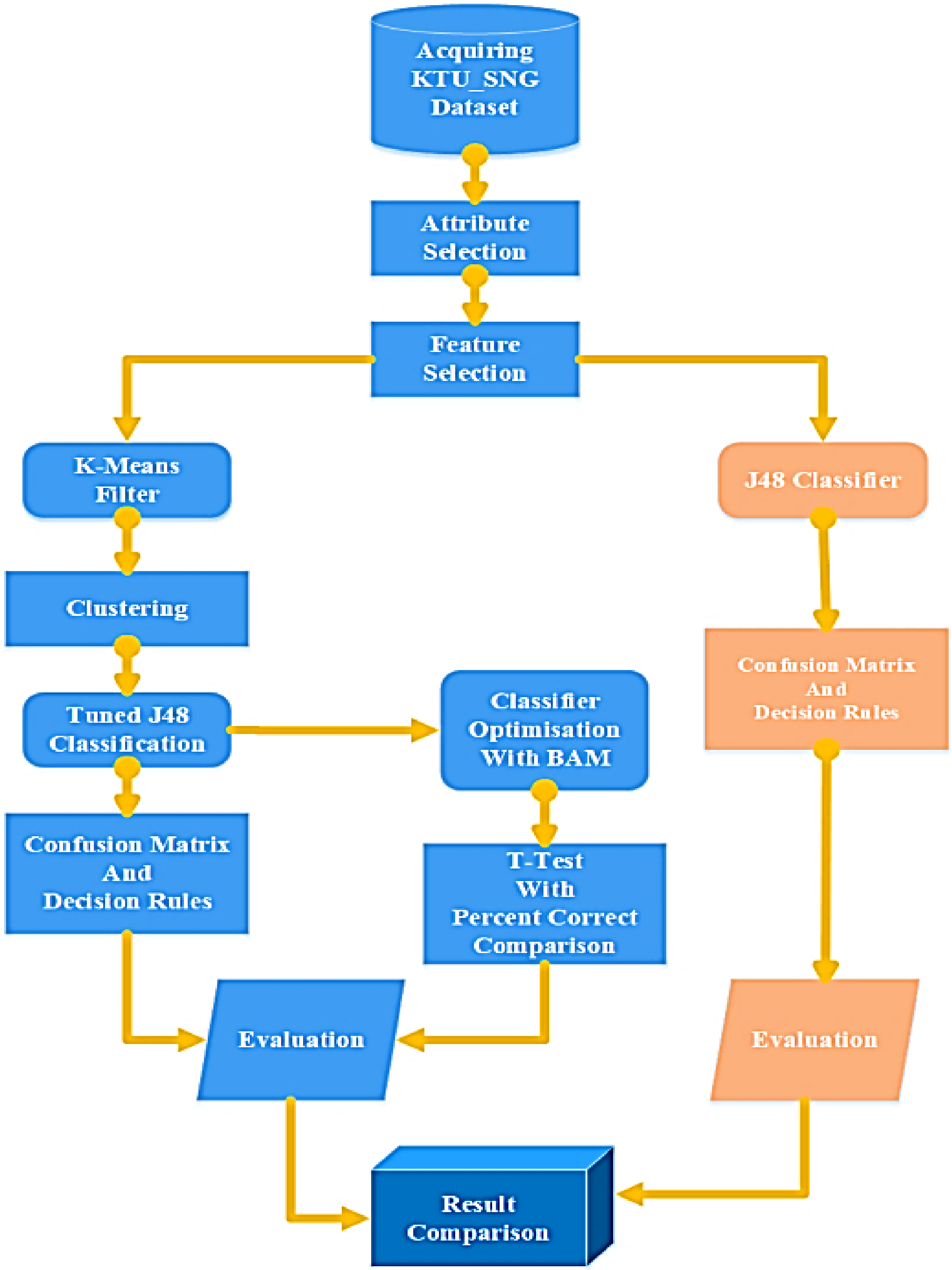


Fig. 4. The Framework of (BMCC) Model.

A tuned Classification technique [21] applied here is the J48 classifier using the WEKA tool. The tuning process may opt for a various combination of base J48, as demo choices. And, K-Means clustering is used here as a feature enhancement tool for natural clustering of classes, with a number of clusters equal to the number of classes [23]. incorporated here for the integration with classification. In this work, the finest classification rules are identified for classifying the students under Outstanding, Excellent, Good, Average, low and very low using the data mining tool.

The model is optimised using the Bootstrap Aggregation Model (BAM) for reducing the complexity in performance prediction. The result is compared with Base BMCC and experimental tests are also undertaken. Also, these results are referred to compare with j48 algorithms to check the efficiency of the proposed model against a base classifier. Other well-known base models in this domain, like Naïve Bayes and MLP, are also tested against the results

1) Developing classifiers (J48): Basically, J48 is the advancement of the ID3 algorithm as mentioned earlier. There are some additional features in J48 such as accounting missing values, pruning, rule derivation, etc. In the WEKA tool, the J48 classifier is processed with open source Java implementation. In this algorithm, the classification is performed constantly till it obtains the pure leaf; hence the results obtained must be as appropriate as possible.

Steps in the Algorithm:

- If some instances fit into the same class of the leaf in tree representation is provided with the label under the same Classification.
- The potential data is evaluated for every attributes provided in the sample and then the Gain value is calculated.
- The current selection criterion provides the best attribute and that could be selected for effective branching.

a) Gain Calculation

The gain computation is dependent on the entropy measure of the data disorders. The Entropy $\vec{y}$ is calculated as,

$$Entropy\ (\vec{y}) = \sum_{b=1}^{n} \frac{|y_a|}{|\vec{y}|} \log \frac{|y_a|}{|\vec{y}|} \quad (1)$$

$$Entropy\ (b/\vec{y}) = \frac{|y_a|}{|\vec{y}|} \log \frac{|y_a|}{|\vec{y}|} \quad (2)$$

And the gain value is calculated from the entropy computations,

$$Gain(\vec{y}, b) = Entropy(\vec{y}) - Entropy\ (b/\vec{y}) \quad (3)$$

The gain value can be maximized by dividing the entropy values by using the split function $\vec{y}$ through the value of b.

b) Pruning

This is the significant process in classification. Some attributes are there in all data sets which may differ from other neighbourhood attribute and may not be well-defined. The classification operation has to be performed with these instances and the decision tree is to be formed. Here, pruning in incorporated to reduce the classification errors, which can be occurred due to the categorization in the training set. Specifically, pruning is performed for the simplification of the tree.

TABLE II. SAMPLES OF KTU_SNG DATASET UNDER CLASSIFICATION

Attendance Percentage	**Total Credit**	**Result**	**Nature_of_Student**
More than 90%	90 and above	Pass	**Outstanding**
More than 80%	Between 80 and 90	Pass	**Excellent**
More than 75%	Between 70 and 80	Pass	**Good**
More than 75%	Between 60 and 70	Pass	**Average**
More than 70%	Between 45 and 60	Pass	**Low**
More than 70%	Below 45	Fail	**Very low**

Here, J48 involves in developing decision trees from a training data set as in ID3 algorithm, based on the information entropy theory. The training data set is considered as TS={ts1, ts2, ..., tsn} of some earlier classified instances. Each sample tsi= a1, a2... is a vector, where a1, a2... denotes the attributes of the samples. The significance of the decision tree is to project the data with a good precision rate. The algorithm chooses an effective feature of the data that presents at each tree node, capable of splitting the sample data set into two subclasses or sets to be developed with one class or another. The samples under classification are provided in Table II, based on the classification of the Nature_of_Student under classes such as Outstanding, Excellent, Good, Average, Low and Very Low.

2) K-Means clustering: K-means clustering is a traditional clustering methodology that combined similar items in large data sets into groups. It involves selecting the initial centroids and determine the 'K' number of clusters by conveying the given instances to the near most centroids detected. Computation of Euclidean distance is the distance measure used in k-means clustering for centroids detection. Further, the computation is as follows,

$$F = \sum_{i=1}^{K} \sum_{j=1}^{n} \left\| x_j^{(i)} - CF_i \right\|^2 \quad (4)$$

Where 'F' is the objective function, 'K' is the number of clusters and 'n' is the samples. The centroid function is denoted as 'CF' and 'x' be the specific case of the instances.

B. *Parameters and Performance Evaluation*

The two most used parameters such as specificity and sensitivity are used to measure the performance of the proposed methodology [25]. The specificity can also be termed as True Positive Rate (TPR). Hence; the Specificity and sensitivity are computed as follows,

$$Specificity = \frac{True\ Negative\ (TN)}{False\ Positive\ (FP) + True\ Negative\ (TN)} = TPR = Recall \quad (5)$$

$$Senitivity = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \quad (6)$$

$$Error\ Rate = \frac{FP}{FP+TN} \quad (7)$$

1) Confusion matrix: The following Fig 5 depicts the confusion matrix for three class case. When a set of objects are evaluated, the results are effectively counted and the confusion matrix has been prepared. It can also be denoted as the contingency table. In the figure, the table has shown the correct decisions of classifier presents on the major diagonal, whereas the errors are at the rest.

The column denotes the Actual class and the rows denote the predictions. An edge is given as TP, whether it is positive or negative as same as the predictions respectively. FP denotes the results that are not predicted accurately. TN is representing the correctly predicted results, whereas, the FN values denote the falsely predicted with the preferred network.

Actual				
		Positive	Negative	Non-Existent
Predicted	Positive	TP	FP	FN
	Negative	FP	TP	FN
	Non-Existent	FP	FP	TN

Fig. 5. Confusion Matrix.

2) Precision rate (PR): The retrieval of positive predictions is called precision. In particular, it is computed as the number of appropriate classification samples belong to 'A' divided by a number of samples categorized as under the class 'A'. Hence, it is defined as the ratio of the predicted true positives out of all actually positive results. The formula is given as follows:

$$Precision\ Rate\ (PR) = \frac{True\ Positive\ (TP)}{True\ Positive + False\ Positive} \quad (8)$$

3) Accuracy: Accuracy of computation is plainly defined as the relationship of a total number of correctly classified samples into the total sum of adopted samples. Mathematically, it can be defined as,

$$Accuracy = \frac{TP+TN}{FN\ ((TP+1)+(1+TN))} * 100 \quad (9)$$

4) F-Measure: F-measure is a significant parameter for evaluating the proficiency of the proposed model. It combines the TPR and the Precision Rates (PR) into an instant measure of performance. The equation is given as,

$$F - Measure = \frac{2*TPR*PR}{TPR+PR} \quad (10)$$

5) T-Test: The final evaluation of this experimental work for classifier evaluation is done using Paired T-Test for classifiers. Weka workbench experiment options are used to test its effectiveness. Weak learning model and strong learning model can easily be identified from the outputs. Percent Correct is set as the comparison filed all through the test. This t-tester assumes the samples are independent.

V. Experimental Results and Discussions

In this section of the experimental analysis, the results obtained for the proposed (BMCC) model corresponding to the parameters given in the (section 4.2) are compared with the traditionalistic classification techniques such as Naive Bayes, Multilayer Perceptron (MLP) and J48. Moreover, for the evaluation, KTU_SNG student dataset which contains 13920 records (includes 232 samples with 60 attributes) has been taken. For implementation, the training data set is classified through the WEKA tool with J48 decision-making classifier algorithm. The Bound Model of Clustering and Classification of data mining has initially employed with the K-Means algorithm for clustering as feature addition. It helps to remove the class attributes through the unsupervised learning process with the implemented training sample, following that Tuned J48 classifier has been implemented on the instances for classifying the student's performance.

The model use clustering as filter feature addition for the subsequent Tuned J48 classifier in this model. The K-Means cluster divides the sample into K clusters and the provision to specify the number of clusters help it used in classification. It is going to be one another strong feature during the classification.

Since similar studies are found in these areas, a famous model of clustering and classification have been opted with an optimisation to improve its performance. There were feature selection and ensemble models [24] of approaches. The optimisation is implemented using Bootstrap Aggregation Model. During the optimisation of BMCC, there are used choices of J48 as the demo options of the combination. How improvements have produced in BMCC are analysed and presented hereafter.

The experimental results illustrate that the proposed (BMCC) model provides potential results with utmost Precision Rate (PR), accuracy and the adaptability among the individual utilization of classification and clustering techniques in Table III, Table IV. The Fig. 6 illustrates the performance of BMCC and Improved BMCC. The Fig. 7 depicts the distribution of students according to their classes provided by the data set. The Precision Rate and accuracy are evaluated on the basis of True Positive, True Negative, False Positive and False Negative as per the (8) and (9), given in the previous sections B.2 and B.3.

The Table III depicts the comparative analysis based on the results of the parameters evaluated. Form the observation, it is apparent that the error rate for (BMCC) model is **0.095** which is comparably low than others. And, the accuracy rate is **94.83%** which is most desirable than other models. It shows that the (BMCC) model provides better-précised results than others and helpful for efficient student classification and improvement of student's performance.

For computation purpose, it is to be considered that, the false positive rate or the value of sensitivity would be **0**. As per the results provided in Table III and Table IV, the sensitivity rate obtained for the combined model is the lowest (closer to 0), when compared to the other techniques. Another assumption is also to be made for Precision Rate as 1. And from the results provided in Table III, it is noticeable that the proposed model comprises a higher rate of precision than the others.

The experiment is also conducted for identifying Nature_of_Student, the Class of students, have also done, however, the results were not competing with performance prediction of students. And the details are slightly suspended from this paper as it focuses on a binary classification problem. The result of the study regarding nature had shown a significant rise in FN rate in many experiments, and it is assumed to be due to the perceived quality of the internal assessments and its recording. Students attitude and their sentiments towards activities are not considered to grade students in KTU. It may be a work proposed to analyse the impact of sentiment analysis in the field of student classification problems, and thereby a solution may be devised permanently.

TABLE III. PARAMETERS BASED PERFORMANCE EVALUATION

Parameters	**Naive Bayes**	**MLP**	**J48**	**BMCC**
TP	161	176	170	177
FP	5	10	8	5
TN	43	38	40	43
FN	23	8	14	7
Error Rate	0.121	0.078	0.095	0.052
Accuracy	87.93%	92.24%	90.52%	94.83%
Precision	0.97	0.947	0.956	0.973
Recall	0.875	0.957	0.924	0.962
F1 Measure	0.921	0.952	0.94	0.968
Sensitivity	0.88	0.96	0.92	0.96
Specificity	0.90	0.79	0.83	0.90

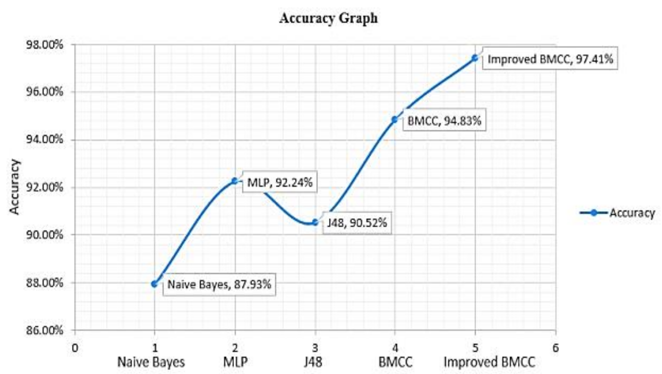


Fig. 6. Performance Accuracy of BMCC and Improved BMCC with Others.

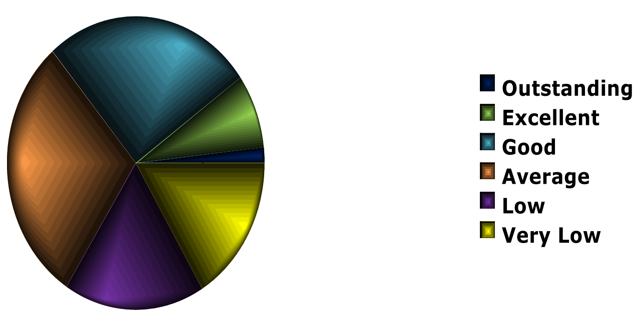


Fig. 7. Distribution of Students According to their Classes.

The following Table IV depicts the values obtained for the proposed (BMCC) models with overall classification results under "PASS", "FAIL". It is portrayed from the table that out of 232 instances, 220 are classified accurately under the suitable class and the rest are not accurately classified as given in the Base BMCC. Whereas the improved one has 226 and 6 as correctly classified and misclassified correspondingly to its credits. The optimum estimated results of the evaluation parameters are given in Table IV. Further, the Fig.9 shows the results obtained with Classification using (BMCC) Model and implemented with and without optimisation in the Weka Workbench environment.

TABLE IV. RESULTS FOR EVALUATION PARAMETERS FOR CLASSIFYING "PASS" OR "FAIL"

Total Number of Instances= 232			**Values In - Dataset Base BMCC**		**Values In - Dataset Improved BMCC**	
Correctly Classified			220	**94.80%**	226	**97.4%**
Incorrectly Classified			12	**5.20%**	6	**2.60%**
TP	**FP**	**PR**	**RC**	**F-M**		**CLASS**
0.962	0.962	0.973	0.962	0.967		PASS
0.896	0.038	0.86	0.896	0.878		FAIL
Weighted Average:						
0.948	**0.09**	**0.949**	**0.948**	**0.949**		
0.984	0.063	0.984	0.984	0.984		PASS
0.938	0.016	0.938	0.938	0.938		FAIL
Weighted Average:						
0.974	**0.053**	**0.974**	**0.974**	**0.974**		

The decision tree based classification sample and rules are given in Fig 8.

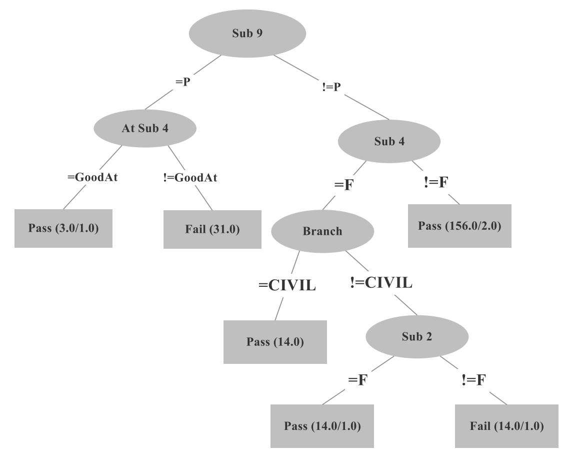
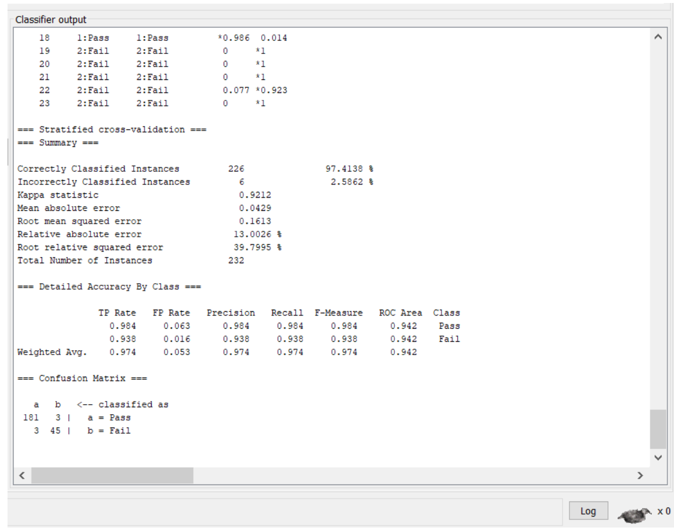


Fig. 8. Decision-Tree-based Classification Sample.



```
Classifier output
    18     1:Pass     1:Pass        *0.986  0.014
    19     2:Fail     2:Fail         0     *1
    20     2:Fail     2:Fail         0     *1
    21     2:Fail     2:Fail         0     *1
    22     2:Fail     2:Fail         0.077 *0.923
    23     2:Fail     2:Fail         0     *1

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         226               97.4138 %
Incorrectly Classified Instances         6                2.5862 %
Kappa statistic                          0.9212
Mean absolute error                      0.0429
Root mean squared error                  0.1613
Relative absolute error                 13.0026 %
Root relative squared error             39.7995 %
Total Number of Instances              232

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.984     0.063      0.984     0.984     0.984      0.942    Pass
                 0.938     0.016      0.938     0.938     0.938      0.942    Fail
Weighted Avg.    0.974     0.053      0.974     0.974     0.974      0.942

=== Confusion Matrix ===

   a   b   <-- classified as
 181   3 |   a = Pass
   3  45 |   b = Fail

Log x 0
```

Fig. 9. A sample Result of (BMCC) Model Implementation in WEKA Tool.

TABLE V. AVERAGE ERROR RATE OF CLASSES COMPARED TO (BMCC) CLASSIFICATION MODEL

Method	**Naive Bayes**	**MLP**	**J48**	**Base BMCC**	**Improved BMCC**
Error Rate	0.108	0.174	0.148	0.09	0.053

TABLE VI. RESULTS OF RELATED EXPERIMENTS

Method	**J48**	**Base BMCC**	**Improved BMCC**
Accuracy	90.52%	94.83%	97.41%

From the outset of this work, it is clear that the utilisation of classification techniques is considered. Especially the accuracy and precision of each of them considered more. From Table V, it is clear that the proposed bound model of clustering and classification has produced the most competent outputs. A random improvisation is also tried with the BMCC and a marginal improvement has shown in it. However, a detailed observation may be done later with more option to improve the proposed one.

Whatever being implemented had shown an appealing progress in the experiments, even though few variations in results based on the filtering used at the pre-processing had also shown. It may be considered and experimented with more kinds of data from various environments collected. However, it is intuitively suggested to try-out for an optimum feature extraction and pre-processing for these kinds of researches. The results of variations of models used in this and related models are depicted in Table VI.

Based on the experimental analysis given in this paper, it is explicit that the Bound Model of Clustering and Classification provides better results in classifying instances with a better rate of accuracy.

During the experiment in this paper, subsamples of KTU_SNG Data are prepared using different filters such as Gain Ratio, Info Gain, Correlation Attribute filtering...etc. Also, the actual data along with the data prepared using Weka for testing and training are used during the bootstrap optimisation of BMCC. It shows an incremental uplifting of accuracy during bootstrapping (Bagging-Tuned J48). Similar to every bagging its learning curve started from a minimum of Base BMCC rate and progressed to achieve a score of 99.569% accuracy. The score initially felt a little exaggerating as the database size and the stability of it is taken into account. However, the Max-Rate (99.569%) may be reached at the year goes by and the features of database tend to become normal form. The comparison of the last few observations of optimised BMCC is given in Table VII.

TABLE VII. OPTIMISED PERFORMANCE OUTCOME OF BMCC

Model	**Experiment 1**	**Experiment 2**	**Experiment 3**	**Experiment 4**
BMCC	94.3966%	97.8448%	99.569%	99.569%

TABLE VIII. OPTIMISED PERFORMANCE OUTCOME OF BMCC TEST

Model	**Exp. T-Tester1**	**Exp. T-Tester 2**	**Exp. T-Tester 3**	**Exp. T-Tester 4**
BMCC	93.09%	93.44%	95.08%	98.71%

During the performance test, the same had produced comparatively lover figures with the experimental models are given in Table VIII.

The model could Minimise the Cost-Benefit at 75.72 gain threshold with 99.569% accuracy as given in Fig 10.

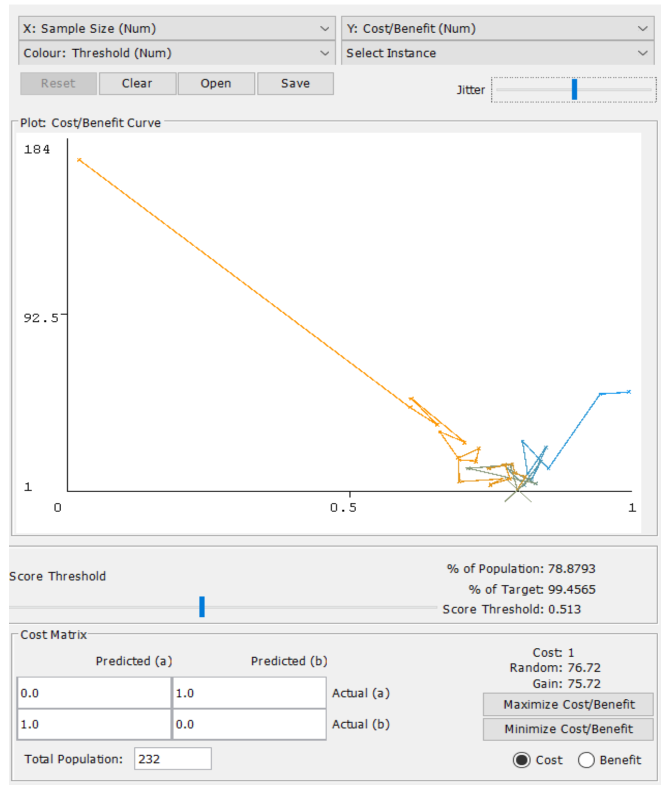


Fig. 10. Minimised Cost-Benefit of (BMCC) at Optimum Gain.

Interesting properties of Reinforcement Learning [22] in the form of a reutilising model, for information extraction, might ameliorate BMCC. Hence, non-rationally believed that it provides an overriding Educational Data Mining technique or DDM technique for the managing performance of the students.

VI. CONCLUSION AND FUTURE ENHANCEMENT

The proposed (BMCC) model is an effective technique for the Proficient Performance Prediction (PPP) of educational datasets and mining classification that helps in successfully identifying the huge data sets of educational domains. Decision tree J48, proposed BMCC and its improved model came up with **90.52%**, **94.83%** and **97.41%** accuracy respectively in predicting the performance of students in the KTU_SNG-Dataset. The evaluation results are evidence for the Bound Model of Clustering and Classification technique provides results with more precision than the traditional classification methodologies such as Naïve Bayes, MLP and base J48, for categorizing the given instances with their specific classes and attributes. A substantial increase of **7.0%** is noted here. As is well known, clustering operations belong to the unsupervised learning process; hence, the classes are developed on the basis of formed clusters to which the samples belong to. And then, the unknown datasets are classified on the basis of the decision rules, which are acquired by employing a classification technique on these clusters. The classification results obtained from this combined methodology brings utmost precision rate and accuracy with minimal error rate and cost benefit. The **99.569%** accuracy at the experiment and the **98.71%** accuracy at test obtained by the proposed model serve the best result compared to all other. The result shows that the Optimised BMCC outperformed all other methods in this research experiment. And hence it may be a good choice for the future also to enhance the model with optimisation and booting methodologies to enhance classification accuracy in Educational Data Mining. And, the potential results are valuable for the academicians for analysing the student's performances and making decisions or arrangements according to that to enhance their quality along with the institution.

As it is discussed performance improvisation methods and its applications to BMCC and a dependable, refined and standardised information extraction technique for it are going to be a good choice of future work. Also, in future, the combined work can be extended and integration can also be done with the association or ensemble techniques and the results can be analysed with some other efficient models. Furthermore, the methodology can be applied and examined with some other interesting areas like the sentiment of students towards their activities, sports, medicine and so on, which have to deal with huge datasets.

REFERENCES

[1] Romero, C. and Ventura, S. "Educational Data Mining: A Survey from 1995 to 2005", Exp.t Sys. with App. (33), (2007), pp. 135-146.

[2] Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. "Mining Student Data Using Decision Trees", The 2006 Int. Arab Conf. on Info. Technol. (ACIT'2006) – Conference Proceedings, (2006).

[3] Shannaq, B., Rafael, Y. and Alexandro, V. "Student Relationship in Higher Education Using Data Mining Techniques", Glob. J. of Comput. Sci. and Technol., (2010), vol. 10, no. 11, pp. 54-59.

[4] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[5] Baradwaj, B. and Pal, S. "Mining Educational Data to Analyse Student s' Performance", Int. J. of Adv. Comput. Sci. and Appl., (2011), vol. 2, no. 6, pp. 63-69.

[6] Gaganjot Kaur and Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", Int. J. of Comput. Appl. (0975 – 8887), July 2014, Volume 98 – No.22, pp. 13-17.

[7] Kumar, V. and Chadha, A. "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", Int. J. of Adv. Comput. Sci. and Appl., (2011), vol. 2, no. 3, pp. 80-84

[8] Chandra, E. and Nandhini, K. "Knowledge Mining from Student Data", Europ. J. of Sci. Res., (2010), vol. 47, no. 1, pp. 156-163.

[9] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, "predicting students' performance using id3 and c4.5 classification algorithms", Int. J. of Data Mining & Knowl. Manag. Proc. (IJDKP) Vol.3, No.5, September 2013, pp. 39-52.

[10] Ayesha, S., Mustafa, T., Sattar, A. and Khan, I. "Data Mining Model for Higher Education System", Europ. J. of Scienti. Res., vol. 43, no. 1, (2010), pp. 24-29.

[11] Romero, C., Ventura, S. and Garcia, E. "Data mining in course management systems: Moodle case study and tutorial", Comput. & Educ., vol. 51, no. 1, (2008), pp. 368-384.

[12] El-Halees, A. "Mining Students Data to Analyse Learning Behavior: A Case Study", The 2008 Int. Arab Conf. of Info. Technol. (ACIT2008) – Conf. Proc.s, University of Sfax, Tunisia, Dec 15- 18. (2008).

[13] Parneet Kaura, Manpreet Singh, Gurpreet Singh Josanc, "Classification and prediction based data mining algorithms to predict slow learners in the education sector", 3rd Int. Conf. on Rec. Tren. in Comput. 2015(ICRTC-2015), Science Direct, Procedia Comput. Sci. 57, (2015), pp. 500 – 508.

[14] Syed Tahir Hijazi, and S. M. M. Raza Naqvi, "Factors affecting students' performance: A Case of Private Colleges", Bangladesh e-J. of Sociol., Volume3.Number1, January 2006.

[15] Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance", In Proc.s of the 37th ASEE/IEEE Front. in Educ. Conf., 2007 Pp. 7- 12.

[16] Anoopkumar M and A. M. J. Md. Zubair Rahman, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration," IEEE Int. Conf. on Data Min.g and Adv. Comput., Ernakulam, 2016, pp. 122-133.

[17] Fadhilah Ahmad*, Nur Hafieza Ismail and Azwa Abdul Aziz, "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques" Appli. Math. Sci.s, Vol. 9, 2015, no. 129, 6415 – 6426.

[18] Amjad Abu Saa, "Educational Data Mining & Students' Performance Prediction" (IJACSA) Int. J. of Adv. Comput. Sci. and Appl., 2016Vol. 7, No. 5.

[19] V. Ramesh, P. Parkavi, K.Ramar "Predicting Student Performance: A Statistical and Data Mining Approach," Int. J. of Comput. Appl. (0975 – 8887), February 2013, Volume 63– No.8.

[20] MUSTAFA AGAOGLU, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education", IEEE Access, May 13, 2016, Digital Object Identifier 10.1109/ACCESS.2016.2568756.

[21] Anoopkumar M and A. M. J. Md. Zubair Rahman, "Model of Tuned J48 Classification and Analysis of Performance Prediction in Educational Data Mining", Int. J. of Appli. Engg. Res. ISSN 0973-4562 Volume 13, Number 20 (2018) pp. 14717-14727

[22] Narasimhan, K., Yala, A., & Barzilay, R. (2016). "Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning". Proceed. of the 2016 Conf. on Empir. Meth. in Nat. Lang. Proc. pages 2355–2365, Austin, Texas, November 1-5, 2016 https://doi.org/10.18653/v1/D16-1261

[23] Kyriakopoulou, A., Kalamboukis, T.: Using clustering to enhance text classification. In SIGIR 2007, 30th Annual International ACM SIGIR Conf. on Res. and Devel.t in Info. Ret., 2007. pp805–806.

[24] Pushpalata Pujari, Jyoti Bala Gupta, "Improving Classification Accuracy by Using Feature Selection and Ensemble Model", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2, May 2012.

[25] D. L. Gupta, A. K. Malviya, Satyendra Singh, "Performance Analysis of Classification Tree Learning Algorithms", Int. J. of Comp. Appli. (0975 – 8887), Volume 55– No.6, October 2012.

Brain Signal Classification using Genetic Algorithm for Right-Left Motion Pattern

Cahya Rahmad[1], Rudy Ariyanto[2], Dika Rizky Yunianto[3]
Information Technology, State Polytechnic of Malang
Indonesia

***Abstract*—Brain signals or EEG are non-stationary signals and are difficult to analyze visually. The brain signal has five waves alpha, beta, delta, gamma, and theta. The five waves have their frequency to describe the level of attention, alertness, character and external stimuli. The five waves can be used to analyze stimulation patterns when turning left and right. Giving weight to the five brain waves utilizes genetic algorithms to get one signal. Genetic algorithms can be used to find the best signal for classification. In this paper, the EEG signal will be classified to determine the right or left movement pattern. After combining the five brain waves with genetic algorithms is then classified using the Logistic Regression, Linear Discriminant Analysis, K-Neighbors Classifier, Decision Tree, Naïve Bayes Gaussian, and Support Vector Machine. From the six methods above that have the highest accuracy is 56% and SVM is a method that has better accuracy than others on this problem.**

***Keywords*—Brain wave; EEG; genetic algorithm; classification; left right movement**

I. Introduction

When the body does a job or movement, it basically coordinates with the mind. These conditions, focus on one object without being affected by other things and focus on doing the movement. Knowing someone's focus condition is not easy, one way to find out the condition of one's focus is through information on brain signals or often called an Electroensephalogram (EEG) signal.

EEG is an instrument used to record static electricity activity resulting from stimuli received by the brain. Research on the classification of EEG signals has been carried out, including classification of fatigue levels, identification of EEG signals for sound stimulation, identification of alertness, emotional conditions, attention classification, identification of epilepsy, and other studies classifying EEG signals against imagination of body movements, classification visual stimulation, classification of EEG signals with two mental conditions with an introduction of up to 83%, identification of epilepsy waves, and to recognize movements of artifacts [1][2][3].

Transforming EEG signals into a model is an effective way of analysis to classify EEG signals. An EEG signal in a person generally consists of wave components which are differentiated based on their frequency region, delta, theta, alpha, beta, and gamma [4][5][6].

TABLE I. FREQUENCY EEG

EEG	Frequency
Delta	0 - 4 Hz
Theta	4 – 8 Hz
Alpha	8 – 13 Hz
Beta	13 – 30 Hz
Gamma	30 – 100 Hz

In Table 1 it can be seen that Delta waves have a frequency of 0-4 Hz. Delta waves appear when someone is sleeping soundly. Theta waves have a frequency of 4 - 8 Hz. Theta waves appear when a person sleeps lightly and is in a happy state. Some recent research links these waves such as rapid eye movements during sleep, and hypnosis. Alfa waves have a frequency of 8-13 Hz. Alpha waves appear when a person relaxes, and eyes are closed. These waves are often used to see normal or abnormal brain functions. Beta waves have a frequency of 13-30 Hz. Beta waves appear when someone is doing activities concerning remembering such conditions as thinking. Gamma waves have a frequency of 30-100 Hz. Gamma waves are related to brain activity to integrate various stimuli [7]–[9].

The EEG signal analysis in this paper will be used to analyze for control an object for right-left movement. EEG analysis uses a genetic algorithm to combine five brain signals. the genetic algorithm is one of the heuristic methods which is a branch of an evolutionary algorithm, which is a technique for solving complex optimization problems by imitating the evolutionary process of living things. genetic algorithm proved to be suitable to be used to solve multi-objective problems. genetic algorithm develops along with the rapid development of information technology. This algorithm is widely used in the fields of physics, biology, economics, sociology, and others who often face optimization problems with complex or even difficult mathematical models. by using the genetic algorithm, one brain signal from the five brain signals is obtained. one signal will be used for classification in the next process.

II. Methods

In this paper, several stages are carried out. These stages can be seen in figure 1.

In Figure 2 it is explained that the method proposed. Where each of the five brain waves signals will be weighted by the genetic algorithm method. The result a signal that will be used for classification.

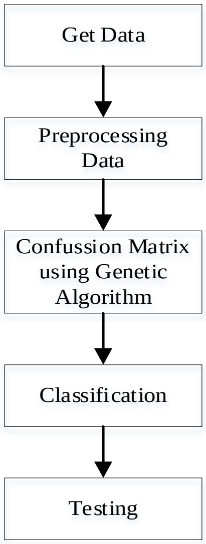


Fig. 1. Process Diagram Flow.

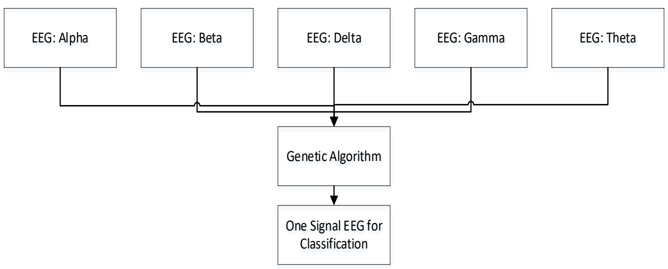


Fig. 2. Proposed Method.

A. Data

The data is taken using a brain signal reader. The tool is MindWave Mobile Brainwave Starter Kit from NeuroSky. The device can directly read the five brain waves. The use of this tool utilizes Bluetooth technology, and this tool is compatible with operating systems such as Windows, Mac, Android or Linux. The device can be seen in figure 3.

Brainwave data is taken from 10 people where each person is recorded 10 times. 5 times the brainwave recording for the condition of the right moves and 5 times the brainwave recording for the condition of the left movement. So the total data captured is 100 data.

Fig. 3. Brainwave Reader.

Recording of brain wave is done when the subject controls the right or left of the remote control car toy and recording is done for one minute. The data took 15 seconds in the middle for the next analysis. Example and visualization of data can be seen at table 2 and figure 4.

TABLE II. EXAMPLE RAW DATA

ms	1	2	…	15	CLASS
ALPHA	69987,5	49024,5	…	335993	LEFT MOVEMENTS
BETA	6527,5	9185	…	8822	
GAMMA	1409	1220	…	7123	
THETA	1532876	856975	…	716820	
DELTA	2	2	…	2	

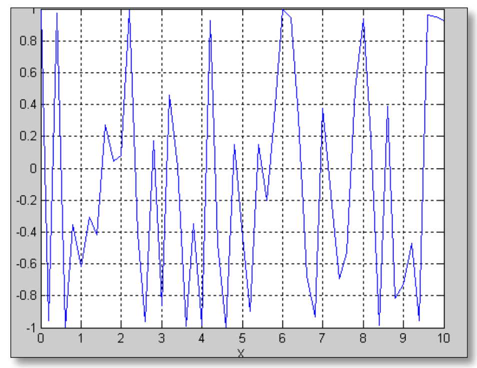


Fig. 4. Visualization Data.

B. Preprocessing

At this stage, the data will be normalized using the min-max method. This normalization is done to adjust the signal magnitude. All waves will be mapped with a range between 0 and 1. Min-max normalization calculations are carried out following equation 1.

$$y = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

C. Genetic Alghorithm

A genetic algorithm is a search heuristic that is inspired by Charles Darwin's theory of natural evolution. This algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction to produce offspring of the next generation.

The process of natural selection starts with the selection of fittest individuals from a population. They produce offspring which inherit the characteristics of the parents and will be added to the next generation. If parents have better fitness, their offspring will be better than parents and have a better chance of surviving. This process keeps on iterating, and at the end, a generation with the fittest individuals will be found[10].

This notion can be applied to a search problem. We consider a set of solutions for a problem and select the set of best ones out of them. Five phases are considered in a genetic algorithm is the initial population, fitness function, selection, crossover, and mutation.

The initial population is processed begins with a set of individuals which is called a Population. Each is a solution to the problem you want to solve. On this EEG problem at equation 2, it is known that y is the new value of combining five signals and x is the signal captured by the reader. while w is the weight sought for each type of wave where the sum of w must be 1 like equation 3. In the process of the genetic algorithm we will find the best w weights to be multiplied by each of the types of waves that have been taken on average to determine that weight[10], [11].

$$y = w1.alpha(x) + w2.beta(x) + w3.delta(x) + w4.gamma(x) + w5.theta(x) \quad (2)$$

$$w1 + w2 + w3 + w4 + w5 = 1 \quad (3)$$

An individual is characterized by a set of parameters (variables) known as Genes. Genes are joined into a string to form a Chromosome (solution). In a genetic algorithm, the set of genes of an individual is represented using a string, regarding an alphabet. Usually, binary values are used (string of 1s and 0s). We say that we encode the genes in a chromosome [10].

The fitness function determines how to fit an individual is (the ability of an individual to compete with other individuals). It gives a fitness score to each. The probability that an individual will be selected for reproduction is based on its fitness score [12].

The idea of the selection phase is to select the fittest individuals and let them pass their genes to the next generation. Two pairs of individuals (parents) are selected based on their fitness scores. Individuals with high fitness have more chance to be selected for reproduction [12].

Crossover is the most significant phase in a genetic algorithm. For each pair of parents to be mated, a crossover point is chosen at random from within the genes. Offspring are created by exchanging the genes of parents among themselves until the crossover point is reached. The new offspring are added to the population [11].

In particular new offspring formed, some of their genes can be subjected to a mutation with a low random probability. This implies that some of the bits in the bit string can be flipped. The mutation occurs to maintain diversity within the population and prevent premature convergence [10].

The algorithm terminates if the population has converged (does not produce offspring which are significantly different from the previous generation). Then it is said that the genetic algorithm has provided a set of solutions to our problem [10].

D. Classification

From the process of genetic algorithm each data produces a signal that can be classified. In this paper uses six classification methods to get the best classification.

- Logistic Regression (LR)

In the statistical model with two categories, with response variables contain elements of "success" or "failure". This binary data is the simplest form of data category. The most frequently used model for data two the category is binary logistic regression [13].

Logistic Regression (LR) is used to measure the functional relationship between one dependent variable from a qualitative type of dichotomous with independent variables of type quantitative and qualitative. It is somewhat similar to multiple linear regression. It is usually appropriate for models where dependent variables are of the qualitative type of dichotomous. Model parameters are estimated using the maximum-likelihood method [13]. Form a logistic regression model with variables i predictors are as follows equation (4)

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (4)$$

By using a logit transformation from π (x), then the logit function regression model can be defined as following equation (5).

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots . + + \beta_i x_i \quad (5)$$

The logit form *g(x)* is a logit model, a linear function in its parameters, and is within the distance between -∞ to + ∞ depends on variable *X*.

- Linear Discriminant Analysis (LDA)

LDA performs linear analysis has its representation (vectors base) from dimensionless EEG vector space high, depending on the statistical point of view. By projecting an EEG vector into its base vector, representation will be obtained the feature of the wave per unit time [14].

Similarity measurements will then be made between EEG representations with data testing. Representations in this method are considered as a linear transformation from the original EEG vector to in a projection space (base vectors)[14].

- K-Neighbours Classifier (KNN)

K-nearest neighbor algorithm is a classification technique a very popular one introduced by Fix and Hodges (1951), which have been proven to be a good simple algorithm. KNN is one method used in classification by using a supervised algorithm (Chan et al. 2010). The purpose of this algorithm is to classify new objects based on the distance of an object to be classified to sample data. Classifier only uses the distance function from new data to training data.

K-Nearest Neighbor is to find the closest distance between data will be evaluated with neighbor K closest neighbors in training data. Training data is projected into space many dimensions, where each dimension represent features of data. This space is divided into parts based on the classification of training data. A point in this space is marked class c, if class c is a classification that is most commonly found in neighboring fruit closest to that point [15]. Near or far neighbors are usually calculated based on distance Euclidean with the following equation (6).

$$d_i = \sqrt{\sum_{i=1}^{p} (x2_i - x1_i)^2} \quad (6)$$

When x1 is data sample, x2 is data testing, and i is variable data. d is a symbol of distance and p is a symbol of data dimension.

In the learning phase, this algorithm only does vector-vector storage features and classification of learning data. In the classification phase, the features are the same calculated for test data (which is the classification not known). The distance from the new vector this is against the entire vector of learning data counted, and a number of the most fruits close taken. The new point is classification predicted included in the classification most of these points. The best k value for this algorithm depends on the data [16].

Generally, the value of k is height will reduce the effect of noise on classification, but make boundaries between each classification becomes more blurred. Value k good can be selected by optimization parameters, for example by using cross-validation. Where special cases classification is predicted based on data the closest learning (with words another, k = 1) is called the nearest neighbor algorithm [16].

The accuracy of the KNN algorithm is very influenced by the presence or absence of features irrelevant, or if feature weight it is not equivalent to its relevance against classification. When is the amount of data approaching infinity, this algorithm guarantees error rate of no more than twice Bayes error rate (minimum level for distribution certain data)[15].

- Decision Tree Classifier (CART)

Decision tree is one of the most popular data mining classification techniques. The appropriate decision tree is used for cases that have the following characteristics [17]:

1) Data or examples are expressed by pairs of attributes and values.

2) Labels or output data are usually discrete values.

3) Data has a missing value.

Entropy theory is adopted to choose the right attribute breakdown for the C4.5 algorithm, stating the average number information needed to classify samples.

- Naïve Byaes Gaussian (NB)

Naïve Bayes is an algorithm in data mining technique that applies the Bayes theory in classification. Bayes' decision theorem is a fundamental statistical approach within pattern recognition (pattern recognition). Naive Bayes based on the assumption of simplifying that value attributes conditionally free if given output value. In other words, given the output value, the probability of observing together is a product of individual probabilities. In equation 7 P(c|x) is the posterior probability of class (target) given predictor (attribute) and P(c) is the prior probability of class. P(x|c) is the likelihood which is the probability of predictor given class and P(x) is the prior probability of predictor [18][19].

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{7}$$

- Support Vector Machine (SVM)

The support vector machine (SVM) was first introduced by Vapnik in 1992 when presented at Annual Workshop on Computational Learning Theory. The basic principle of SVM is a linear classifier, namely case classifications that are linearly separated [20].

Support Vector Machine (SVM) is a method classification that works by searching hyperplane with the biggest margin. The hyperplane is the inter-class data dividing line. The margin is the distance between the hyperplane and the closest data in each class. The closest data with a hyperplane in each of these classes which are called support vector. SVM is a method used for the classification of two classes (binary classification). In its development, several a method proposed for SVM to be used for multi-class classification by combining some binary classifier. Method ever proposed is the One-against-one method. As for the One-against-one method, it will be constructed by a number of k (k-1) / 2 SVM classification models with each model being trained to use data from two different classes [21][12].

Therefore, for data on classes i and j, SVM completes binary classification problems for The decision function for the function above is taken through voting, if the result is $sign$ states that data x is in class i, then vote value for class i plus one. Furthermore, class predictions from data x are classes with values highest vote. Otherwise, the vote value for class j plus one. In general, the stages of document classification with multi-class SVM this research can describe as the following stages. First of all, from the results of the pre-process and the decision, then generated k-groups of data terms in the representation vector space model. Then each group was trained to use the SVM Multi-class with the One-against-one method. Therefore the document classification model will be obtained on each group [20], [21].

E. Testing

At this stage, testing of the method will be done by dividing the data into training and testing. Each type of wave on the EEG signal will be classification after weighting of the genetic algorithm process done. The testing process is done by allocating data testing at 10%, 15%, 20%, 25%, and 30%.

III. Results

The implementation of the proposed method is run in the Python programming language. Python uses various types of libraries, sklearn, matplotlib and pandas.

The results of implementation and testing can be seen in Table 3. From the table it can be seen that the method has the highest average accuracy is SVM with a total average of 48%. And the method which has the smallest average is KNN method which is not at odds with the decision tree method with the average accuracy of 38% and 39%.

In figure 5 it can be seen that the SVM method has an average above other methods. of all methods the smaller the training data, the accuracy will decrease.

TABLE III. TESTING RESULTS

Data Testing	Acuration (%)					
	LR	LDA	KNN	CART	NB	SVM
10%	57	52	30	27	40	40
15%	50	53	32	35	50	40
20%	40	45	50	50	45	55
25%	50	35	45	50	50	55
30%	35	25	35	35	40	50
AVERAGE	**46**	**42**	**38**	**39**	**45**	**48**

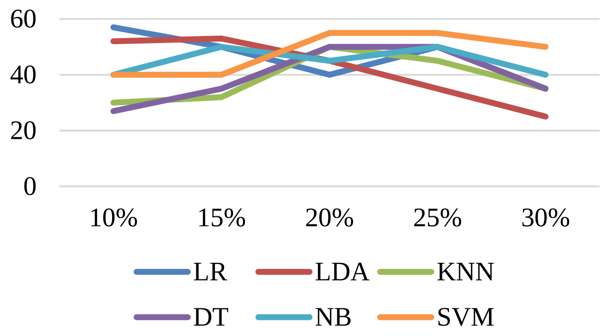


Fig. 5. Diagram Result.

IV. CONCLUSION

The brain signal has five different frequency waves. The five signals will be combined using Genetic Algorithm into one signal that can be used for classification. The signal classification is intended to determine the pattern of right-left movement thoughts. The results of the classification with the highest accuracy were 56% using the Logistic Regression method. However, the highest average accuracy is owned by SVM with 48%. This research can be further developed by further shortening the time unit in retrieving signals and changing the weight values for each type of wave to have better accuracy.

ACKNOWLEDGMENT

We are pleased to acknowledge the contribution of our colleagues (Afif, Yusron, Rendy, Yudi, Yesaya, Marcelina, Fikri, Mr. Erfan and Mr. Indrazno) and our institute Polytechnic State of Malang.

REFERENCES

[1] G. Townsend, B. Graimann, and G. Pfurtscheller, “Continuous EEG classification during motor imagery - Simulation of an asynchronous BCI,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 12, no. 2, pp. 258–265, 2004.

[2] M. Witkowski, M. Cortese, M. Cempini, J. Mellinger, N. Vitiello, and S. R. Soekadar, “Enhancing brain-machine interface (BMI) control of a hand exoskeleton using electrooculography (EOG),” J. Neuroeng. Rehabil., vol. 11, no. 1, pp. 1–6, 2014.

[3] V. J. Shetty, M. Laxman, M. S. Magdum, S. B. Sidnal, and S. Paraddi, “Head Motion Controlled Wheel Chair using,” vol. 2, no. 6, pp. 2393–2395, 2015.

[4] E. Haselsteiner and G. Pfurtscheller, “Using time-dependent neural networks for EEG classification,” IEEE Trans. Rehabil. Eng., vol. 8, no. 4, pp. 457–463, 2000.

[5] D. Wang, D. Miao, and C. Xie, “Best basis-based wavelet packet entropy feature extraction and hierarchical EEG classification for epileptic detection,” Expert Syst. Appl., vol. 38, no. 11, pp. 14314–14320, 2011.

[6] I. Sturm, S. Lapuschkin, W. Samek, and K. R. Müller, “Interpretable deep neural networks for single-trial EEG classification,” J. Neurosci. Methods, vol. 274, pp. 141–145, 2016.

[7] Y. Zhang, G. Zhou, J. Jin, Q. Zhao, X. Wang, and A. Cichocki, “Sparse Bayesian Classification of EEG for Brain-Computer Interface,” IEEE Trans. Neural Networks Learn. Syst., vol. 27, no. 11, pp. 2256–2267, 2016.

[8] L. Hellstro m-Westas, I. Rosen, L. S. de Vries, and G. Greisen, “Amplitude-integrated EEG Classification and Interpretation in Preterm and Term Infants,” Neoreviews, vol. 7, no. 2, pp. e76–e87, 2006.

[9] A. Arif and E. C. Djamal, “Klasifikasi Tingkat Kelelahan Berdasarkan Sinyal Electroencephalogram (EEG) Menggunakan Jaringan Syaraf Tiruan Backpropagation,” in SNIJA Unjani, 2014, p. 2014.

[10] R. A. Welikala et al., “2015 - Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy,” 2015.

[11] V. Jain, A. Kumar, S. Kumar, and C. Chandra, “Weight restrictions in Data Envelopment Analysis: A comprehensive Genetic Algorithm based approach for incorporating value judgments,” Expert Syst. Appl., vol. 42, no. 3, pp. 1503–1512, 2015.

[12] P. Ghamisi and J. A. Benediktsson, “Feature selection based on hybridization of genetic algorithm and particle swarm optimization,” IEEE Geosci. Remote Sens. Lett., vol. 12, no. 2, pp. 309–313, 2015.

[13] D. Liu, T. Li, and D. Liang, “Incorporating logistic regression to decision-theoretic rough sets for classifications,” Int. J. Approx. Reason., vol. 55, no. 1 PART 2, pp. 197–210, 2014.

[14] E. Neto, F. Biessmann, H. Aurlien, H. Nordby, and T. Eichele, “Regularized linear discriminant analysis of EEG features in dementia patients,” Front. Aging Neurosci., vol. 8, no. NOV, pp. 1–10, 2016.

[15] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for kNN Classification,” ACM Trans. Intell. Syst. Technol., vol. 8, no. 3, pp. 1–19, 2017.

[16] M. Havaei, P. M. Jodoin, and H. Larochelle, “Efficient interactive brain tumor segmentation as within-brain kNN classification,” Proc. - Int. Conf. Pattern Recognit., pp. 556–561, 2014.

[17] E. Gokgoz, A. Subasi, and B. S. Processing, “E. Gokgoz, A. Subasi, Comparison of decision tree algorithms for EMG signal classification,” Biomed. Signal Process. Control, vol. 18, pp. 138–144, 2015.

[18] A. Choi, N. Tavabi, and A. Darwiche, “Structured Features in Naive Bayes Classification,” Proc. 30th Conf. Artif. Intell. (AAAI 2016), pp. 3233–3240, 2016.

[19] J. Wu, S. Pan, X. Zhu, Z. Cai, P. Zhang, and C. Zhang, “Self Adaptive Attribute Weighting for Naive Bayes Classification,” Expert Syst. Appl., vol. 42, no. 3, pp. 1487–1502, 2015.

[20] J. Cervantes, F. García Lamont, A. López-Chau, L. Rodríguez Mazahua, and J. Sergio Ruíz, “Data selection based on decision tree for SVM classification on large data sets,” Appl. Soft Comput. J., vol. 37, no. December, pp. 787–798, 2015.

[21] F. Riaz, A. Hassan, S. Rehman, I. K. Niazi, and K. Dremstrup, “EMD-based temporal and spectral features for the classification of EEG signals using supervised learning,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 24, no. 1, pp. 28–35, 2016.

Amharic based Knowledge-Based System for Diagnosis and Treatment of Chronic Kidney Disease using Machine Learning

Siraj Mohammed
Department of information Technology
Jigjiga University (JJU)
Jigjiga, Ethiopia

Tibebe Beshah (PhD)
Information Science Department
Addis Ababa University
Addis Ababa, Ethiopia

***Abstract*—Chronic kidney disease is an important challenge for health systems around the world and consuming a huge proportion of health care finances. Around 85% of the world populations live in developing country, where chronic kidney disease prevention programs are undeveloped. Treatment options for chronic kidney disease are not readily available for most countries in sub-Saharan Africa including Ethiopia. Many rural and urban communities in Ethiopia have extremely limited access to medical advice where medical experts are not readily available. To address such a problem, a medical knowledge-based system can play a significant role. Therefore, the aim of this research was developing a self- learning knowledge based system for diagnosis and treatment of on the first three stages of kidney disease that can update the knowledge without the involvement of knowledge engineer. In the development of this system, the following procedures are followed: Knowledge Engineering research design was used to developed prototype system. Purposive sampling strategies were utilized to choose specialists. The information was acquired using both structured and unstructured interviews and all knowledge's are represented using production rule. The represented production rule was modeled by using decision tree modeling approach. Implementation was employed using pro-log tools. Testing and evolution was performed through test case and user acceptance methods. Furthermore, we extensively evaluate the prototype system through visual interactions and test cases. Finally, the results show that our approach is better than the current ones.**

***Keywords*—Knowledge-based system; kidney diseases; machine learning; knowledge engineering; knowledge representation**

I. Introduction

Chronic kidney disease is an important challenge for health systems around the world and consuming a huge proportion of health care finances. It is even more significant for developing countries which now face the double burden of infectious diseases and growing problems of non-communicable diseases such as cardiovascular, diabetes and hypertension [1]. Around 85% of the world populations live in developing country of the world [2]. Treatment options for chronic kidney disease are not readily available for most countries in sub-Saharan Africa including Ethiopia. For many years the treatment and diagnoses chronic kidney disease in Ethiopia has not been studied and there is no national strategy for prevention and care of patients with chronic kidney disease. Among the people of developing countries like Ethiopia, chronic kidney disease are the one the increase causes of death, because in Ethiopia there is no well-established chronic kidney disease prevention programs and a shortage of specialists and health facilities are there. Therefore, it is difficult to diagnosis and treatment of chronic kidney disease in such like conditions.

Artificial intelligence has emerged after the introduction of the first computers. Currently, the concept of the artificial intelligence is understood as a branch of computer science, in which the computer programs have capability to simulate human behaviors. Computer programs are used both for experimental and practical purposes, such as formulation of medical diagnoses. KBS is one of braches of the AI groups. A knowledge-based system is a software system that contains a significant amount of knowledge in an explicit, declarative form. This has now been replaced by methodological approaches that have many similarities with general software engineering practice. KBS development is best seen as software engineering for a particular class of application problems. These applications problems typically require some form of reasoning to produce the required results. In current business practice there is an increasing need for such systems, due to progression of information technology in our daily work. Some typical applications are systems for diagnosing problems in medical sections [3]. Knowledge based system has become as platform to solve real problem in a fashion similar to the expert in medical areas. Knowledge based system is a computer program which captures the knowledge of experts on a given problem to solve problems in a fashion similar to the expert.

The system can assist the expert during problem-solving, or act like expert in situations where the expertise is lacking**.** The commonly used knowledge acquisition techniques are interviews. Interviews can be classified into structure, semi-structure and unstructured, observation and document analysis. Learning technologies to develop KBS are neural networks, intelligent agents, fuzzy logic, and genetic learning and self-learning [4]. Just a few research papers are accessible that have the capacity of self-learning with KBS. In broad aspect, the self-learning field is called as machine learning in which the computer programs learn from their own experience upon utilization. Self-learning is one of the elements of KBS which tries to imitate the learning capability of human beings. It is possible to update the knowledge base of the KBS either manually or automatically using machine learning algorithms [5, 6]. According to Akerkar, Sajja and Solomon G. [5, 7], "Self-learning is a scientific task that enables the knowledge-based system to learn automatically from the inference process". The aim of this research was to develop a self-learning Amharic based knowledge-based system to help the patients and physicians for diagnosis and treatment through Amharic language for kidney diseases. The prototype system is provided to the end users via computers. The key contributions of this paper are as follows.

- Acquired Real facts data form physicians and modeling all facts to integration with knowledge rules using a self-learning approach.
- Knowledge analysis: the process of making sense of the knowledge was performed.
- Knowledge structuring: The processes of expressing the analyzed knowledge in an understandable and usable form are presented.
- Knowledge modeling: the process of connecting the decision flow of the acquired knowledge and relationship between concepts was presented.
- Expressing knowledge in a format of rule representation techniques.
- Design and development of a flexible self-learning knowledge-based system.

The rest of this paper is organized as follows. In Section 2, we provide an overview of the Statement of the Problem. Sections 3 describe the objective of this paper. In section 5 describe the review of related works. In Section 6, we describe the methodology of each of the key components of the prototype system in detail and the implementation of the prototype system. Section 6 considers a case study and user visual interaction approaches to evaluate the proposed system. The evaluation results are analyzed and presented for better understanding. Section 7, 8 concludes the work with implications and possible future research directions.

II. Statement of the Problem

Chronic kidney disease is an important challenge for health systems around the world and, consuming a huge proportion of health care finances. It is even more significant for developing countries which now face the double burden of infectious diseases and growing problems of non-communicable diseases such as cardiovascular, diabetes and hypertension [8]. Treatment options for chronic kidney disease are not readily available for most countries in sub-Saharan Africa including Ethiopia. For many years the treatment and diagnoses chronic kidney disease in Ethiopia has not been studied and there is no national strategy for prevention and care of patients with chronic kidney disease. Among the people of developing countries like Ethiopia, chronic kidney disease is the one the increase causes of death, because in Ethiopia there is no well-established chronic kidney disease prevention program. Unfortunately, the nature of chronic kidney disease is can progress to end stage, which is renal disease. In our country, there are no sufficient numbers of specialists and medical doctors [9]. Thus, all patients do not get enough diagnoses and treatment on time. In addition, a lot of people in Ethiopia were only known that they had this kidney disease when their kidney become chronic and affected their life. They think that all symptoms that they have were just normal fever and can be fine when they take some pills. That is a bad attitude and there are some factors that encourage to this bad attitude such as the traveling distance to the clinic or hospital, takes a lot of time in waiting especially in governmental hospital. Usually, most of the medical data Collected from patients are just saved in files or kept in folders, but those huge amounts of messy medical records have not meaning for users. Using learning KBS technique will solves the existing problem by analyzing and reusing a previous knowledge to turn that knowledge into useful knowledge base system for the diagnosis of kidney disease. Therefore, this system could help physicians, health professionals and patient to determine the diagnosis and treatment of the patient using their own language, which is Amharic language. The significance of conducting this study is to support physicians and patients in the diagnosis and treatment of kidney patients, particularly in our county where more than 85% of the Ethiopian population lives in rural areas [10]. Therefore, as stressed in the problem statement, the general issue that needs to be addressed is the possibility of integrating "a self- learning approach" with "rule-based approach" to develop a self-learning KBS for diagnosis and treatment of CKD. At the end, this study will answer the following research questions:

- What type of knowledge is required to design a self-learning knowledge-based system?
- What are the suitable approaches to acquired knowledge from experts?
- What are the suitable models, representation techniques and implementing tools for a self- learning Amharic based knowledge-based system?
- How to model and represent the acquired knowledge to developing a self- learning Amharic based knowledge-based system?
- How to design and implement learning knowledge-based system techniques that are automatically updates its knowledge from experience?
- How to measure the performance of the prototype system.

III. Objective

A. General Objectives:

The general objective of this study is to design and develop a self-learning Amharic based knowledge based system that can provide advice through Amharic user interface for physicians and patients in order to facilitate the diagnosis and treatment of kidney disease.

B. Specific Objectives:

To achieve the general objective of this study, the general objective should be described as the following sub specific tasks:

- To review literatures on the concept of knowledge based system.
- To acquire knowledge in this area.
- To identify a set of learning processes.
- To model and represent acquired knowledge.
- To develop learning knowledge-based system for diagnosis and treatment of kidney diseases.
- To evaluate and validate the performance of the prototype system.
- To draw conclusions based on the findings and forward appropriate recommendations as future research directions.

IV. Significance

This research is significant in several sides. In user side, the beneficiaries of the system are physicians and patients. The prototype system is very important as a training tool in the areas where shortages of skilled experts are available. It is also advantageous for a rural area that has a computer system and scarcity of medical professionals and medication facilities. In system developer side, the result of this study will be used as an input for the development of a full knowledge based system and it could be one approach for knowledge acquisition techniques to develop case-based reasoning applications.

V. Review of Related Works

Chronic kidney disease (CKD) is a condition in which the kidneys are damaged and cannot filter blood as well as possible. This damage can cause wastes to build up in the body and lead to other health problems, including cardiovascular disease (CVD), anemia, and bone disease. People with early CKD tend not to feel any symptoms [11]. The only ways to detect CKD are through a blood test to estimate kidney function, and a urine test to assess kidney damage. CKD is usually an irreversible and progressive disease and can lead to kidney failure, also called End Stage Renal Disease (ESRD), over time if it is not treated The stages of chronic kidney disease are determined by the glomerular filtration rate. Glomerular filtration is the process by which the kidneys filter the blood, removing excess wastes and fluids. Glomerular filtration rate (GFR) is a calculation that determines how well the blood is filtered by the kidneys. It is one way to measure kidney function. Glomerular filtration rate is usually calculated using a formula that includes a person's age, gender, race, and serum creatinine levels. A knowledge-based system (KBS) is a software system that contains a significant amount of knowledge in an explicit, declarative form. The area of KBS development has matured over the past two decades. It started with first-generation expert systems with a single flat knowledge base and a general reasoning engine, typically built in a rapid-prototyping fashion. This has now been replaced by methodological approaches that have many similarities with general software engineering practice. KBS development is best seen as software engineering for a particular class of application problems. These applications problems typically require some form of reasoning to produce the required results. In current business practice there is an increasing need for such systems, due to progression of information technology in our daily work. Some typical applications are systems for assessing loans in a bank, for job-shop scheduling in a factory, for configuring an elevator, and for diagnosing problems in a production line [12].

Moreover, the related works conducted by local and global researchers in the medical domain have been reviewed, such as E.K.juuso and k.Leiviskä [13] proposed an “Adaptive Expert Systems for Metallurgical Processes” to integrate rule-based system with fuzzy models for adaptive system, prototype KBS in antiretroviral therapy [14], designing a KBS for blood transfusion [15], prototype KBS for anxiety mental disorder diagnosis [16], the potential for applying KBS for diagnosis of acute respiratory tract infections [17], human disease diagnosis using a fuzzy expert system [18] and development of online children skin diseases diagnosis system [19] in Ethiopia. Furthermore, to the knowledge of the researchers there is no a self-learning KBS for diagnosis and treatment of CKD diseases. Therefore, the objective of the study is to develop a self-learning KBS for diagnosis and treatment of CKD.

A. Learning Technologies

Learning system may incorporate neural networks, fuzzy logic, genetic learning, self-learning KBS, or a combination of any or all of these technologies. Four state-of-the-art learning technologies are discussed as follows. These are neural networks, fuzzy logic, genetic learning, and self-learning knowledge-based system.

VI. Methodology

The details of the research methodology appeared in the following sections, but an outline is being provided in this section. Methodological steps followed for this research are:

- Formulation of the problem and its justifications.
- Knowledge was acquired: Using structural, unstructured interviews with experts and from existing document.
- Develop a conceptual design for concept formation using decision tree and productive rule for rule representations techniques.
- Implement a self-learning prototype system: Architecture for prototype system and implementation part of the system was developed based on the conceptual design.
- Testing and evaluations of the prototype system through visual interactions and test case were carried out.

- Refinement: Refine the prototype system based on the outcomes of the testing and evaluation result and repeat all steps if necessary.

A. Study Area, Population and Sampling Technique

The study area for this research is Gondar University specialization hospital. Gondar University hospital is located in Amhara regional state from Ethiopia, in Gondar town at Gondar University medical campus. The size population is six. We used purposive sampling technique to select domain experts for knowledge acquisitions. The reasons to select six domain experts are to share their expertise and experience.

B. Source of Data

We used primary and secondary data as source of data. The primary data was collected by using interviews with domain experts and the secondary data was collected from published, public or private documents, books, journals articles, different past researches, reports, manuals and online materials.

C. Data Collection Methods and Implantation Tool

In this study, the necessary knowledge was acquired and elicited using interviews with the medical experts, particularly the physicians in the area of kidney disease in which six domain experts were selected from Gondar hospital based their professions, educational qualification level, and years of experience on kidney diseases. The interview with experts covered issues like, how the expert interacts with patient, what are the basic symptoms are, what techniques used to identify the patient stages, the procedures of diagnosing and what are the possible treatments recommended to the patient. A secondary data from documents such as medical books, training manuals, public or private documents, reports, online materials and journal medical articles was also assessed. Furthermore, demonstration and direct observation were considered to acquire the necessary knowledge.

Prolog language is chosen as implementation tool because it is suitable for rule based programming, backward chaining execution and pattern matching.

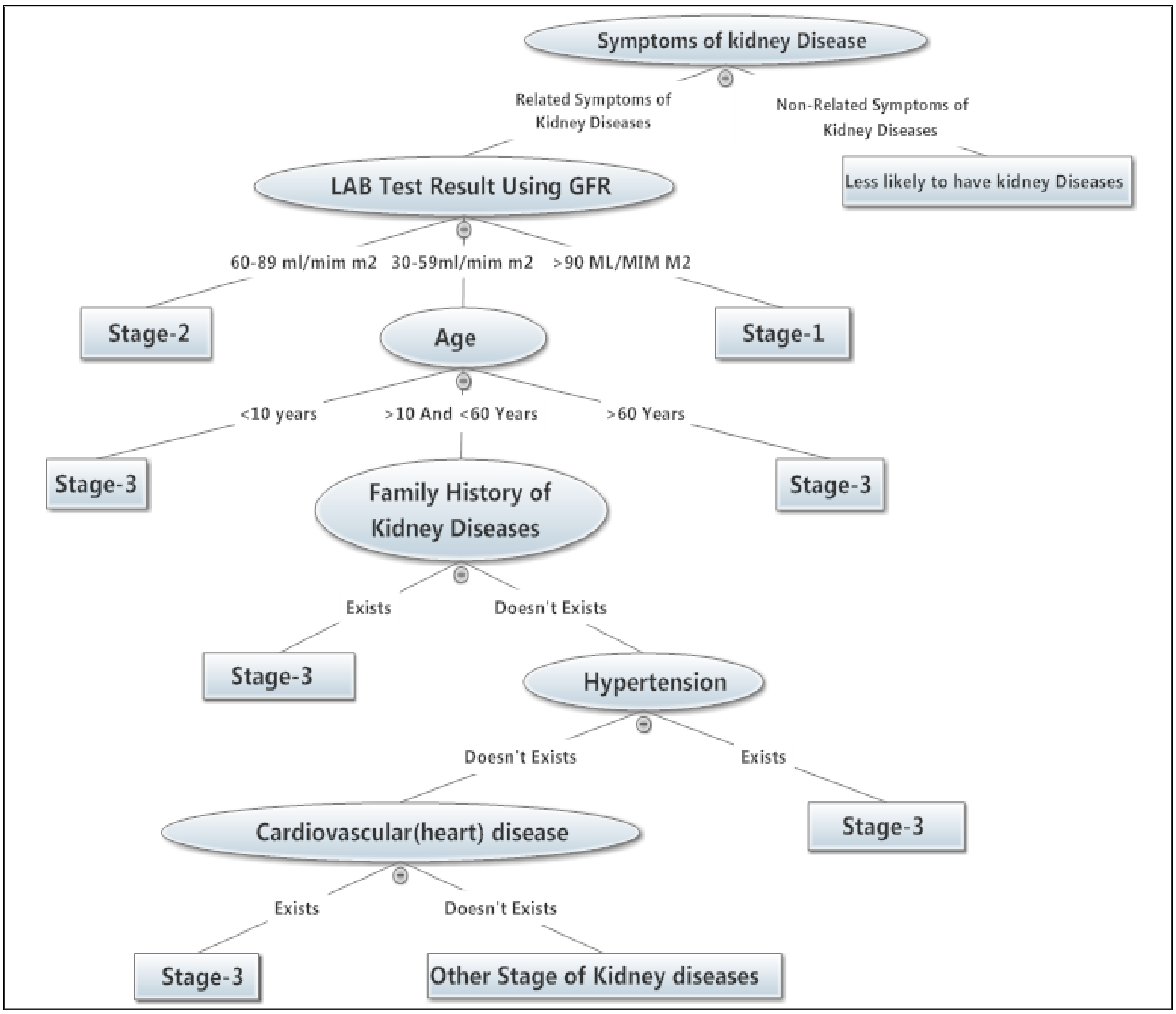


Fig. 1. Decision Trees for Diagnosis and Treatment of Kidney Diseases.

D. Model

Knowledge is to be modeled at a conceptual level. Typically, a knowledge model provides formats for writing down both static domain knowledge (rules, classes, relations) and reasoning strategies [20]. According to Richard et al. [21], one of the most extensively applied methods for conceptual modeling is decision tree. Decision tree commonly acts a key role in a knowledge modeling process. Figure I, show the decision tree structure that has the flow of knowledge in the diagnosis and treatment of CKD.

E. Knowledge Representation

In the Previous section knowledge has been acquired and modeled; the next step is knowledge representation by using appropriate format that is understandable by inference engine. Knowledge representation is a means of encoding the human expert knowledge in an appropriate way. We used rule-based knowledge representation techniques to represent the acquired knowledge. All rules that are used in this study are developed based on the concepts of knowledge modeling approaches. To setup all rules, production rule Knowledge representation techniques are used in which knowledge is represented in the form of condition-action pairs: In the same way, the rules that contain the stages of kidney diseases, major symptoms of kidney diseases, GFR laboratory values, family history of kidney, diabetes, and cardiovascular, and high blood pressure diseases are formulated in the following ways:

Rule 1:
IF
Symptoms of kidney disease = "Non-related symptoms of kidney disease",
THEN
Kidney Stage = "Kidney Stage Disease Free".
Rule 2:
IF
Symptoms of kidney disease = "Related symptoms of kidney disease"
AND** Lab test result using GFR = "≥90* **mL/min/1.73 m²",*
THEN
Kidney Stage = "Kidney Stage 1".
Rule 3:
IF
Symptoms of kidney disease = "Related symptoms of kidney disease"
***AND** Lab test result using GFR = "60-89* **mL/min/1.73 m²",**
THEN
Kidney Stage = "Kidney Stage 2"
Rule 4:
IF
Symptoms of kidney disease = "Related symptoms of kidney disease"
AND** Lab test result using GFR = 30-59* **mL/min/1.73 m²",*
THEN
Kidney Stage = "Kidney Stage 3"
Rule 5:
IF
Symptoms of kidney disease = "Related symptoms of kidney disease"
***AND** Lab test result using GFR = "30-59 mL/min/1.73 m²",*
***AND** Age = "≥60 years",*
THEN
Kidney Stage = "Kidney Stage 3"
Rule 6:
IF
Symptoms of kidney disease = "Related symptoms of kidney disease"
***AND** Lab test result using GFR = "30-59 mL/min/1.73 m²"*
***AND** Age = ">10 and <=60 years"*
***AND** Family history of kidney disease = "Exists",*
THEN
Kidney Stage = "Kidney Stage 3"
Rule 7:
IF
Symptoms of kidney diseases = "Related symptoms of kidney diseases"
AND** Lab test result using GFR = "30-59* **mL/min/1.73 m²",*
***AND** Age = "≥10 and <=60 years",*
***AND** Family history of kidney disease = "Doesn't exist",*
***AND** Hypertension = "Exists",*
THEN
Kidney Stage = "Kidney Stage 3"
Rule 8:
IF
Symptoms of kidney disease = "related symptoms of kidney disease"
AND** Lab test result using GFR = "30-59* **mL/min/1.73 m²",*
***AND** Age = "≥10 and <=60 years",*
***AND** Family history of kidney disease = "Doesn't exist"*
***AND** Hypertension = "Doesn't exist"*
***AND** Cardiovascular (heart) diseases = "Exists",*
THEN
Kidney Stage = "Kidney Stage 3"
Rule 9:
IF
Symptoms of kidney disease = "Related symptoms of kidney disease"
AND** Lab test result using GFR = "30-59* **mL/min/1.73 m²",*
***AND** Age = ">10 AND ≤60 years"*
***AND** Family history of kidney disease = "Doesn't exist"*
***AND** Hypertension = "Doesn't exist"*
***AND** Cardiovascular (heart) diseases = "Doesn't exist",*
THEN
Kidney Stage = "other Kidney Stage"

F. Implementation and Experimentations

In this section we presented the implementation of the prototype system. Knowledge was acquired using both structured and unstructured interviews with domain experts and a relevant document was assessed by using documents analysis method to get concepts and facts about kidney diseases. Then, those extracted facts and concepts were modeled by using decision tree approach and the model has been converted into productive rules and coded by using SWI-Prolog tool to form the self-learning knowledge-based system.

G. Architecture of the Prototype System

System architecture is a conceptual model that defines the structure and guidelines of the system. The overall architecture of our system is shown in Figure II.

Figure II, represents the structure of the prototype knowledge-based system with its main components. The architecture of such a self-learning knowledge-based system consists of some major components that are common for all knowledge-based system such us: human expert, knowledge-base, knowledge engineer, users, but in our system, two important component are added to enhance the knowledge based system such as, Self-learning with dynamic module and explanation facility components.

H. Self-Learning Components of the System

The prototype system has a capability of adding new facts/rules, signs and symptoms, patient history at run time and automatically update the old facts by replacing the new one without knowledge engineer involvements. Subsequently, self-learning module provides dynamic knowledge acquisition techniques to hold fact like, Stages of kidney disease, new symptoms, and the history of patient (name, age, address). Figure III show how to learn patient history and remember all facts after diagnosis and treatments.

The above sample window dialogue has ability to learn new sign, symptoms, and patient history and remember all facts after diagnosis and treatments process. To do this, we developed a dynamic database in order to store patient name, his/her kidney stage disease, age, adders, new sign and symptoms.

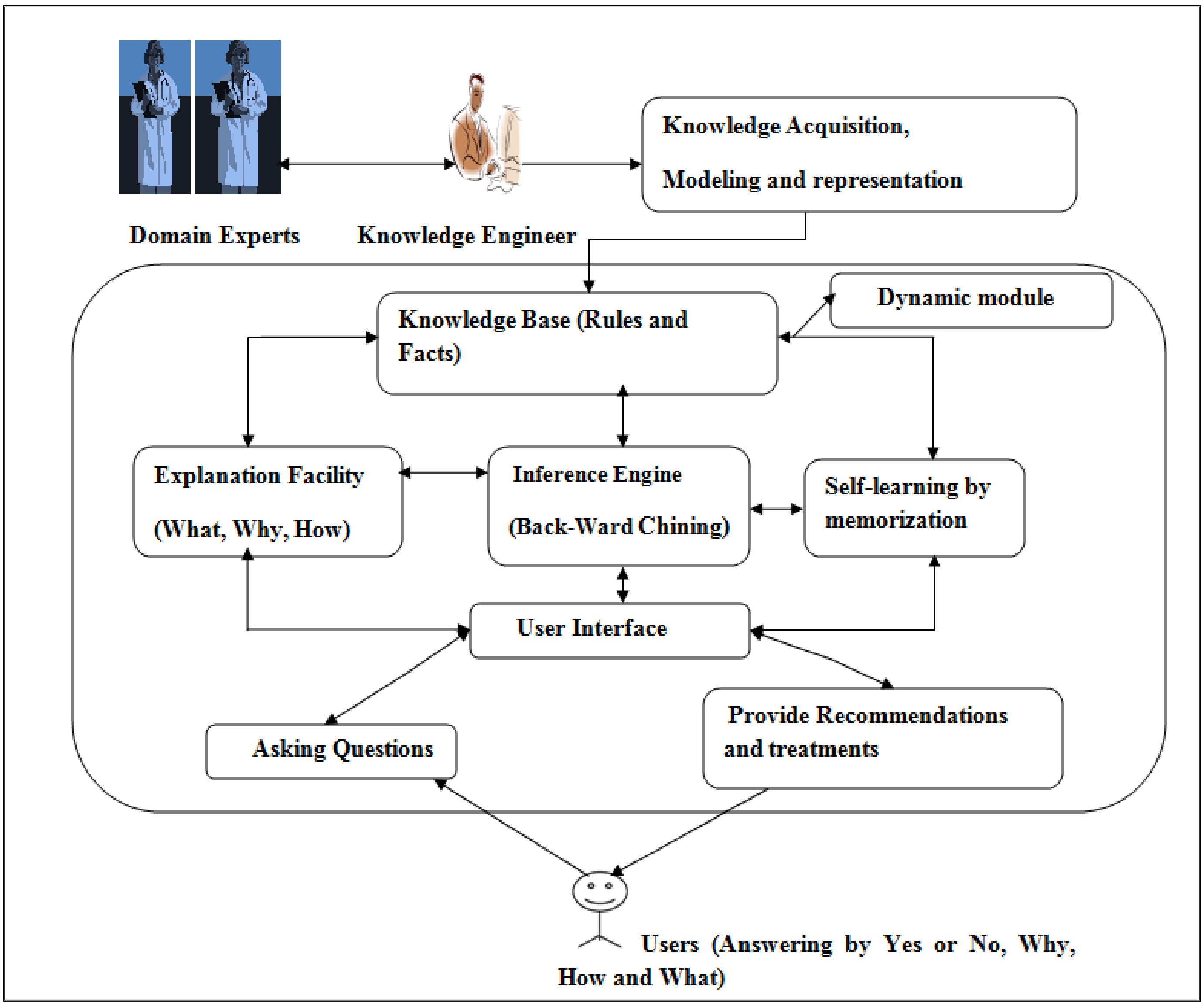


Fig. 2. The Prototype System Architecture.

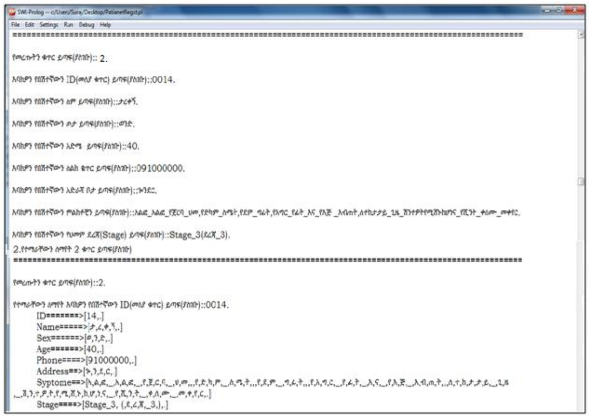


Fig. 3. Illustrates How to Learn Patient History and Remember all Facts after Diagnosis and Treatments.

I. The User Interface

The user interface of the prototype system is allows a non-expert and physicians users to query (question) the KBS, and to see delivery of advices. In this prototype, users can interact with the system through "yes" አዎ", "no" "አይ_የለብኝም", "**what**" "ምን", **"why"** "ለምን" and **"how"** "እንዴት" responses. Based on the user response, the system draws a conclusion and displayed in the user interface window. The picture in figure IV displays the home window that has different options for users to interact with the system.

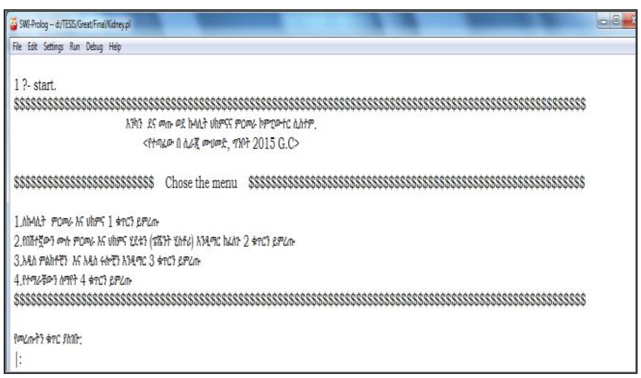


Fig. 4. Welcoming Window for Prototype System.

VII. Testing and Evaluations

Developing a knowledge-based system is an incremental process. Therefore, in developing the prototype system, the knowledge engineer at finishing step should test the performance of the system. In this section, we discussed the testing and evaluation of the prototype system. In our research, we used test cases and visual interactions approaches. The results of this evaluation and test show where the system is weak. Also, the correct and incorrect outcomes are identified by comparing decisions made by domain experts with decisions made by the prototype system.

A. Test Cases

The test cases are used to measure the accuracy of the system. To measure the accuracy of the system, the researcher was selected 60 test cases as a representative of the domain. We categorized those cases into three stages based on the GFR laboratory result values and their symptoms. Naturally, evaluation of the knowledge-based system using test case needs experts as evaluators.

The knowledge-based system testing procedure carried out by system evaluator to classify the test cases into correct or incorrect classes. The evaluation was done by comparing the system test result with the physician answers (as the human expert did). Therefore, System evaluators and knowledge engineer made decisions by comparing the system test result with the physician answers. The result of the comparison shows that our approach has made a close decision as the human expert did. The following table I shows the amount of test case and their exact class classifier.

Table I shows that from total of 60 test cases, in the first row show that, out of 20 diagnosed patients test cases in Stage-1, 20 test cases are classified as correct and 0 as incorrect. On the second row shows that, the total 20 diagnosed patients' test cases, 16 diagnosed patients taste cases are classified as correct and 4 test cases as incorrect. Also, the third rows show, that of the total 20 diagnosed patients' test cases, 19 diagnosed patients' test cases are classified as correctly classified and 1 as incorrectly classified. In general, from 60 diagnosed patients test cases 55 diagnosed patients' test cases are classified correctly and 5 diagnosed patients cases are classified incorrectly. The test case result that provided by system evaluators showed that the prototype system are about 91 % correct and 9% incorrect.

TABLE I. Total Taste Case with Accuracy Values

No.	**Selected Cases**	**Total number of cases**	**Correctly classified**	**Incorrectly classified**	**Accuracy in %**
1	Stage-1	20	20	0	100%
2	Stage-2	20	16	4	80%
3	Stage-3	20	19	1	95%

B. User Acceptance Evaluation

The goal of user acceptance testing is to assess if the system can support day-to-day business, user scenarios and to ensure the system is sufficient and correct for business usage [22]. Visual interaction evaluation method are used when the domain expert to directly interact with the system. The evaluators assess the accuracy of the prototype system by using the following standards, these are:

- Simplicity of use and interact with the prototype system,
- Attractiveness of the prototype system,
- efficiency in time, the accuracy of the prototype system in reaching a decision to identify the stage of kidney diseases,
- The ability of the prototype system in making the right conclusions and recommendations, the ability of the prototype system to remember the patient history, and
- The importance of the prototype system in the domain area.

TABLE II. ILLUSTRATE THE PERFORMANCE EVALUATION RATING THROUGH VISUAL INTERACTION

N0.	**Criteria of evaluation**	**Poor(1)**	**Fair(2)**	**Good(3)**	**Very good(4)**	**Excellent(5)**	**Average**	**Prestige %**
1.	Is the prototype is easy to use and interact with it?	0	0	0	2	3	4.6	92
2.	How do you think the attractiveness of the prototype system?	1	0	0	2	2	3.8	76
3.	Is the system is more efficient in time?	0	0	1	0	4	4.6	92
4.	How accuracy of the prototype system in reaching a decision to identify the sages of kidney disease?	0	0	1	2	2	4.2	84
5.	Does the system incorporate sufficient and practical knowledge?	0	0	0	3	2	4.4	88
6.	Does the system making right conclusions and treatments?	0	0	0	2	3	4.6	92
7.	How do you rate the significance of the system in the domain area?	0	0	0	1	4	4.8	96
8.	How do you rate the ability of a self-learning prototype system to remember the patient's history?	0	0	0	2	3	4.6	92
		Total average					4.45	89

We organized both closed-ended and open-ended questionnaires concerning on acceptance, attractiveness, time efficiency, accuracy of the prototype system, opinions, suggestions and feedback of experts. After evaluation the system the evaluators just put (ticked) their respond based on the researcher assigned values in numbers for each scale as excellent=5, very good=4, good=3, fair=2, and poor =1 for each closed-ended questions. We used the following formulas to get the average values of the first close-ended user acceptance testing questions.

$$\mathbf{1 \times \frac{0}{5} + 2 \times \frac{0}{5} + 3 \times \frac{0}{5} + 4 \times \frac{2}{5} + 5 \times \frac{3}{5} = 4.6}$$

As can be seen in table II, for the simplicity and understandability of the information provided by the system, 40% of the respondents evaluated as a Very Good and 60% of them evaluated as an excellent system. For the attractiveness of the prototype system, 20% of them rated the system as poor, 40% of them ticked as the system very good and the remaining 40% rated as excellent. In the same way, for question "efficient in time" of the prototype system, 20% of them rated the system as good and 80% of them rated as excellent. Similarly, for question "accuracy of the prototype system" of the prototype system, 20% of them rated the system as good, 40% of them ticked as the system very good and the remaining 40% rated as excellent. In the same sheet, the evaluators marked 60% as a very good system for "incorporate sufficient and practical knowledge", and 40% ticked as the system excellent. In the same way for criteria of the prototype system provide "making the right conclusions and treatments"40% of them marked as very good and 60% of them marked as the system excellent. Similarly, for questions "rate the significance of the system in the domain area", 20% of the evaluator rated as very good and the remaining 80% of the evaluated marked as very good. Finally, for the question "rate the ability of a self-learning prototype system to remember the patient's history", 40% of them marked as very good and 60% of them evaluated as excellent.

VIII. DISCUSSIONS AND RECOMMENDATION

The accuracy of the prototype system using test cases are calculated as 91% and the user acceptance testing using close-ended and open-ended questions average evaluation result is 89%, respectively. Therefore, the overall accuracy performance of the prototype system is 90%, because the system has Amharic language and self-learning features.

As stressed in the review of related work, some of the related works are conducted by Belay [23], Anteneh [24], Redeit [25], Solomon [26], Seblewongel [27], Zewditu [28], Guesh [12], and Solomon [17] in the area of medical in Ethiopia. Those local researchers recommend localizations concepts in terms of language and self-learning approaches. Therefore, conferring to their finding and suggestions this study has advancement, because the prototype system has Amharic language and self-learning ability to learn from experience which is enhance usability of the system. The recommendation section provides some clues to the interested researcher to investigate further approaches that are not covered in this research. Most of the time, a knowledge engineers have got some challenges when acquiring knowledge from experts, to solve that challenges the researcher suggested that for those who are interested to do research on the area knowledge acquisitions as future study, it is better using a self-learning KBS technology as knowledge acquisition approaches, because a self-learning KBS has the ability to hold all patient history. Correspondingly, this could be one approach for knowledge acquisition techniques to develop case-based reasoning applications. The prototype knowledge-based system has important modules for adding new rules, symptoms and learns from patient history but, rules that are found in knowledge base should be updated at run time without knowledge engineer involvement. Therefore, additional research needs to update the rules of the system automatically at knowledge base. The scope of this knowledge based system should be comprehensive to integrate others stages of kidney diseases, such as Stage-4 and Stage-5 in order to be converted into a complete system. Furthermore, extract the necessary knowledge from experts was difficult for the researcher due to the personal nature of tacit knowledge. Therefore, it is important to apply data mining techniques to extract the hidden knowledge.

IX. Conclusion

The knowledge engineer to develop a self-learning Amharic based KBS for diagnosis and treatments of kidney disease performed basic tasks: first the knowledge engineer assessed the problem area and the problem is considered feasible, then a statement of requirements is created and the development process continued as follows, knowledge was acquired using both structured and unstructured interviews with domain experts and from relevant documents by using documents analysis method to get concepts and facts of kidney diseases.

The acquired knowledge focuses on concepts and facts of kidney diseases. Then, those extracted facts and concepts were modeled by using decision tree approach and the model has been converted into productive rules and coded by using SWI-Prolog tool to form the knowledge-based system. Finally, the evaluating and testing result shows that, the overall performance of the prototype system registered 91% accurate result.

References

[1] Collins AJ, Li S, Gilbertson DT, et al. "Chronic kidney disease and cardiovascular disease in the Medicare population". Kidney Inter 2003; 64(suppl 87):S120-S27.

[2] E.D. Kock, "Decentralizing the Codification of Rules in a Decision Support Expert Knowledge Base", MSc Thesis, University of Pretoria, Faculty of Engineering, Built Environment and Information Technology, South Africa, 2003.

[3] P-H. Speel, A. Th. Schreiber, W. van Joolingen, G.van Heijst, G.J.Beijer "Conceptual Modeling for Knowledge-Based", Systems Unilever Research Vlaardingen, University of Amsterdam, CIBIT, Bolesian.

[4] Ting-Peng Liang, Special section: Research in integrating learning capabilities into information system, Guest editor, USA, 2010.

[5] Solomom G., "a self-learning KBS for diagnosis and treatment of diabetes" MSc Thesis, Addis Ababa University School of Information Science, Ethiopia 2010.

[6] Ajith Abraham ,"Rule-based Expert Systems" Oklahoma State University, Stillwater, OK, USA,2010.

[7] R.A. Akerkar and P.S. Sajja, "Knowledge-Based Systems", Jones and Bartlett Publishers, USA, 2010.

[8] E.D. Kock, "Decentralizing the Codification of Rules in a Decision Support Expert Knowledge Base", MSc Thesis, University of Pretoria, Faculty of Engineering, Built Environment and Information Technology, South Africa, 2003.

[9] P.J. Tuso, "SERVE Ethiopia", Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2471814/#bib10, Accessed date: June 05, 2011.

[10] World Health Organization, WHO Humanitarian Appeal 2009 Country Profile, June 15, 2011.

[11] Mustafa Arici ,Clinical Assessment of a Patient with Chronic Kidney Disease Department of Nephrology, Faculty of Medicine , Hacettepe University , Ankara 06100 , Turkey.

[12] P-H. Speel, A. Th. Schreiber, W. van Joolingen, G.van Heijst, G.J.Beijer Conceptual Modeling for Knowledge-Based , Systems Unilever Research Vlaardingen, University of Amsterdam, CIBIT, Bolesian.

[13] E.K.juuso and k.Leiviskä , Adaptive Expert Systems for Metallurgical Processes. fac workshop on expert systems in mineral and metal processing, ESPOO, Finland, 26-28 august 1992.

[14] Anteneh W., Prototype Knowledge Based System in Antiretroviral Therapy, MSc Thesis, Addis Ababa University School of Information Science, Ethiopia, 2004.

[15] Guesh, D. Designing knowledge based system for blood transfusion, MSc Thesis, Addis Ababa University School of Information Science, Ethiopia, 2013.

[16] Seblewongel, E. Prototype knowledge based system for anxiety mental disorder, MSc Thesis, Addis Ababa University School of Information Science, Ethiopia, 2012.

[17] Solomon. Develop self-learning knowledge based system for diagnosis and treatment of diabetes, MSc Thesis, Addis Ababa University School of Information Science, Ethiopia, 2013.

[18] Mir Anamul Hasan, Khaja Md. Sher-E-Alam, Ahsan Raja Chowdhury. Human Disease Diagnosis Using a Fuzzy Expert System, Journal of Computing, Vol. 2, No. 6, June 2010, NY, USA, ISSN 2151-9617, 2010.

[19] Munirah, M., Yusof, R., Aziz,J., Chew, S. Development of Online Children Skin Diseases Diagnosis System, 2013.

[20] Buchanan, g.b. and shortliffe, h.e. 1984. Rule-based expert systems: the mycin experiments of the stanford heuristic programming project. In g.b. buchanan and h.e. shortliffe, eds., 2-67.

[21] B.F. Richard, G. Violetta, and O.L. Mieczyslaw, On Features of Decision Trees as a Technique of Knowledge Modelling , Proceedings of the Workshop on Computer Science and Information Technologies CSIT"99, Moscow, Russia, 1999.

[22] Vince Bordo, CBAP Overview of User Acceptance Testing (UAT) for Business Analysts (bas) Gondar University College of Natural Science, Ethiopia, 2010.

[23] Belay, Amharic based KBS for diagnosis and treatment of dental disease, MSc Thesis, Gondar University College of Natural Science, Ethiopia, 2006.

[24] Anteneh W., Prototype Knowledge Based System in Antiretroviral Therapy. MSc Thesis, Addis Ababa University School of Information Science, Ethiopia, 2004.

[25] Redeit A., Design and Development of a prototype Knowledge-Based System for HIV PreTest Counseling. MSc Thesis, Addis Ababa University School of Information Science, Ethiopia, 2006.

[26] Solomon D., the Potential for Applying Knowledge-Based System for Diagnosis of Acute Respiratory Tract Infections. MSc Thesis, Addis Ababa University School of Information Science, Ethiopia, 2010.

[27] Seblewongel E. Prototype Knowledge Based System for Anxiety Mental Disorder Diagnosis. MSc Thesis, Addis Ababa University School of Information Science, Ethiopia, 2011. 83.

[28] Guesh D., Designing a Knowledge-Based System for Blood Transfusion. MSc Thesis, Addis Ababa University School of Information Science, Ethiopia, 2012.

Noble Method for Data Hiding using Steganography Discrete Wavelet Transformation and Cryptography Triple Data Encryption Standard: DES

Cahya Rahmad[1], Arief Prasetyo[3], Novriza Arizki[4]
Department of Information Technology
State Polytechnic of Malang
Malang, Indonesia

Kohei Arai[2]
Department of Information Science
Saga University
Saga City, Japan

***Abstract*—Noble method for data hiding using steganography Discrete Wavelet Transformation: DWT and cryptography triple Data Encryption Standard: DES is proposed. In the current era, information technology has become inseparable from human life, especially regarding the processing and dissemination of information. In line with advances in information technology, there are also parties who want to abuse such information by changing information or even damage it. To avoid that happening, then the data needs to be secured first into other media using the DWT method. The choosing of this method is because the image of the data insertion almost resembles the original image. Triple DES methods are also required to encode data and provide additional security so that hidden data will be difficult to solve. The choosing of this method is because it is resistant against brute force, chosen plaintext, and known plaintext attack. Based on the test, image insertion results in 100% immune to the image manipulation of brightness and contrast, but not so resistant to cropped, resized, and rotated image manipulation. Other tests also indicate that the data which is in the picture can be extracted again and will not undergo any changes.**

***Keywords*—Data hiding; steganography; DWT; cryptography; 3DES**

I. Introduction

As the development of the era, the development of technology is also overgrowing, especially information technology. Information technology has become inseparable from human life and helps people in many ways, such as processing and disseminating information. In line with advances in information technology, there are also parties who want to abuse such information by changing information or even damage it.

Data in 2016 showed 1061 reports of cyber-attacks with cybercrime category of 77.4% and 950 reports in 2017 with cybercrime category of 72.1% [1]. By looking at the data, we can know that cybercrime is a serious problem that we must handle and every year the number of cases of cybercrime will remain substantial. It indeed can be a nightmare for companies or even government agencies who want to send data containing confidential information to clients or other companies. The data needs to be secured first using cryptographic and steganographic methods before being sent to avoid that happening.

Data Hiding Based on Wavelet Multi-resolution Analysis is proposed[2] together with Data Hiding Based on Multiresolution Analysis Utilizing Information Content Concentrations using Eigen Value decomposition[3]. Other than these, Information Hiding Method Based on Coordinate Conversion is proposed [4]. Meanwhile, Data hiding based on Multi-Resolution Analysis: MRA taking into account scanning of the embedded image for improvement of invisibility is proposed[5].

Furthermore, Improvement of secret image invisibility in circulation image with Dyadic wavelet-based data hiding with run-length coding is also suggested [6] together with Method for data hiding based on Legal 5/2 (Cohen-DaubechiesFeauveau: CDF 5/3) wavelet with data compression and random scanning of secret imagery data [7] Moreover, Data Hiding Method Replacing LSB of Hidden Portion for Secrete Image with Run-Length Coded Image is proposed[8].

Steganography is a technique to hide messages or information into other media. Therefore, besides the intended person, they will not be aware of any messages or data in that other media and prevent the occurrence of suspicion.

One of the methods in steganography is the DWT method. DWT has advantages over other steganographic methods that is steganographic images almost resemble the original picture[9]. The message needs to be encoded in advance into another form that cannot be understood using a technique called cryptography to provide additional security for the method.

Cryptography is a technique to avoid information being known by unwanted parties and to convert it into an incomprehensible form. The primary purpose of cryptography is to protect data from unauthorized people[10].

In cryptography, there are many methods, and one of them is the Triple DES method (it is referred to 3DES hereafter)1 operates on 64-bit blocks and uses three keys, each of which is 56-bit in size [10]. Those keys are the strengths of the 3DES method. This method has a reasonably fast processing time and has resistance to attacks such as brute force, chosen-plaintext, and known plaintext[11].

In this research, data security application is made using DWT of steganography method and 3DES cryptography method to secure data by inserting data into an image before sending the data.

In the next section, the proposed method is described followed by some experiments. Then, the conclusion is described together with some discussions and future works.

II. Proposed Data Hiding Method

The proposed data hiding method is based on DWT based steganography and 3DES based cryptography.

A. 3DES

The DES algorithm has proved that highly competent algorithms can be considered uncomfortable and unreliable. Therefore, there is a search for a method for using it again by making it stronger and more secure than creating a new algorithm starting from the scratch. Two significant improvements result in double DES and triple DES algorithms (3DES). Double DES repeats the DES process twice using two keys. If the experiment to crack a key in a DES is 256, then the research to break into two different keys of n-bit is 22n. However, all of this is not entirely true since the introduction of the concept of a meet-in-the-middle attack [12].

Given the idea that double DES may not be strong enough to prevent meet-in-the-middle attacks has led to the development of 3DES algorithms developed by IBM in 1999 by a team headed by Walter Tuchman [13]. This kind of attack is one of the main reasons why double DES is replaced by 3DES that DES operation which is repeated three times using three different keys. It's important to avoid using the same key for encryption, as it will only result in a DES process with slower processing time. 3DES has two shapes, the first form uses three completely different keys and the second one uses two completely different keys [12].

According to [14], 3DES has advantages that are fast processing time and a reasonably reliable level of security. Also, 3DES has resistance to several attacks such as brute force, chosen-plaintext, and known plaintext. This method requires three keys that have a 56-bit length per core. The time needed to check all possible keys using 50 million keys per second for each 3DES key is 400 days [15].

The steps in the 3DES process according to [15] shown in Fig. 1 are as follows:

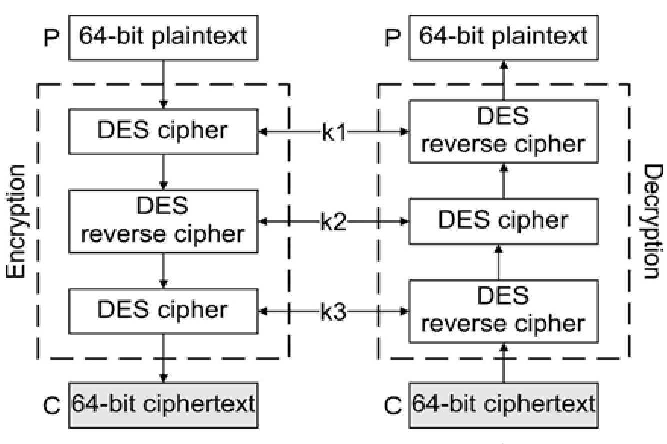


Fig. 1. Encryption and Decryption Process of 3DES.

- Encrypt a message block (plaintext) using a single DES with a K1 key.
- Decrypt the results obtained from step 1 using the K2 key.
- Encrypt the output from step 2 using the K3 key so that it produces ciphertext.
- The decryption process of ciphertext is the opposite of the 3DES Encryption process, which is decryption using K3 keys, encryption using K2 keys, and decryption using K1 keys.

By looking at the advantages of the 3DES algorithm and considering its weakness as well, in this paper, 3DES method is used by using different keys to provide stronger security on the document.

B. DWT

DWT is a method that can divide information from an image into the approach and signal detail. LL bands include low pass coefficients and procedures to a copy as well as more information of other sub-signals indicating vertical, horizontal, or diagonal information or changes in an image [16]. The general equation for DWT can be seen in the equation below [17].

$$DWT\{f\{t\}\} = W_{\emptyset}(j_0, k) + W_{\psi}(j, k) \quad (1)$$

where,.

$$W_{\phi}(j_0, k) \ \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] \phi_{jo,k}[n] \quad (2)$$

$$W_{\psi}(j, k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] \psi_{j,k}[n], \quad j \geq j_0 \quad (3)$$

In the DWT method, there are several techniques for representing images to approach and signal details, one of them is wavelet Haar. Wavelet Haar can be used to describe a picture with a wavelet counting process. The equation of wavelet Haar transform is in the equation shown below [18]:

$$X[2k] = \frac{1}{\sqrt{2}} \left(x[2k] + x[2k+1]\right] \quad (4)$$

$$X[2k+1] = \frac{1}{\sqrt{2}} \left(x[2k] - x[2k-1]\right] \quad (5)$$

The DWT method can represent images into approach and signal details. This method also has advantages compared to other steganography methods, namely the model of steganography results almost resemble the original image [19]. Therefore, in this paper, the DWT method is used to hide the message into the approach and details of the signal, so that the changes that occur in the image will not be too visible in the human vision system.

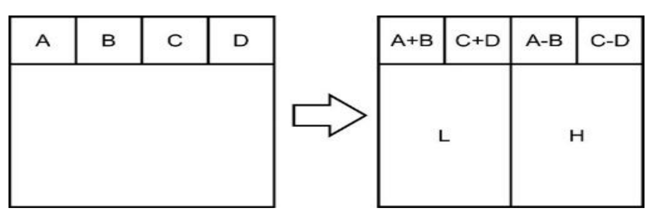


Fig. 2. Horizontal Operation on the First Row.

Haar-DWT is one of the most basic and straightforward transformations in the DWT family. This method reduces the calculation work. Haar-DWT decomposes each signal into two components. The first component is called the average and the second component is called the difference [10]. This process is used to reduce memory requirements and the amount of inefficient Haar coefficient movements. Disadvantages in the sum and subtraction operations can be balanced by decreasing the number of division operations; especially when used at low bit rates, it introduces compression artifacts [19]. A detailed procedure of Haar-DWT 2 dimensions according to [14] described as follows:

Step 1: First, scan the pixels from left to right with the horizontal direction. Then, perform addition and subtraction operations on neighboring pixels. Save the amount on the left and the difference on the right as illustrated in Figure 2. Repeat this operation until the end of the rows. The total pixel represents the low-frequency part (denoted by the symbol L) while the pixel difference represents the high-frequency portion of the original image (indicated by the symbol H).

Step 2: scan the pixels from top to bottom in a vertical direction. Perform the addition and subtraction operations on the neighboring pixels and then store the sums at the top and the difference at the bottom as illustrated in Figure 3. Repeat this operation until the end of the columns. Finally, we will get four sub-bands each denoted as LL, LH, HL, and HH. Sub-band LL is a low-frequency part, so it looks very similar to the original image.

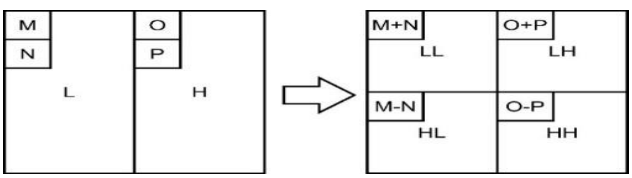


Fig. 3. Vertical Operation.

Haar-DWT is a technique that can be readily applied to get four sub-bands LL, LH, HL, and HH. The processes in Haar DWT are not too complicated to do, but still, represent the frequency of the image. Therefore, in this paper, the Haar-DWT technique is used as an application of the DWT method to get sub-bands of the image and insert the document into the subband.

C. Process Flow of the Proposed Method

The flow of the embedding data and extraction process using the DWT and 3DES methods is explained as follows, Fig. 4:

- Embedding Algorithm

In the insertion process requires a cover image, document (.docx, .xlsx, .pdf, or .txt), and keys (K1, K2, and K3). The following are the steps in the insertion process :

1) Encrypt the document using 3DES with the key, so that it will obtain the ciphertext.

2) Separate ciphertext into three parts that is ciphertext 1, ciphertext 2, and ciphertext 3.

3) Transform the cover image using DWT to get four subbands on the R, G, and B layers.

4) Insert each ciphertext into the HH sub-band at each layer R, G, and B.

5) Image reconstruction uses the IDWT process, resulting in a stego image.

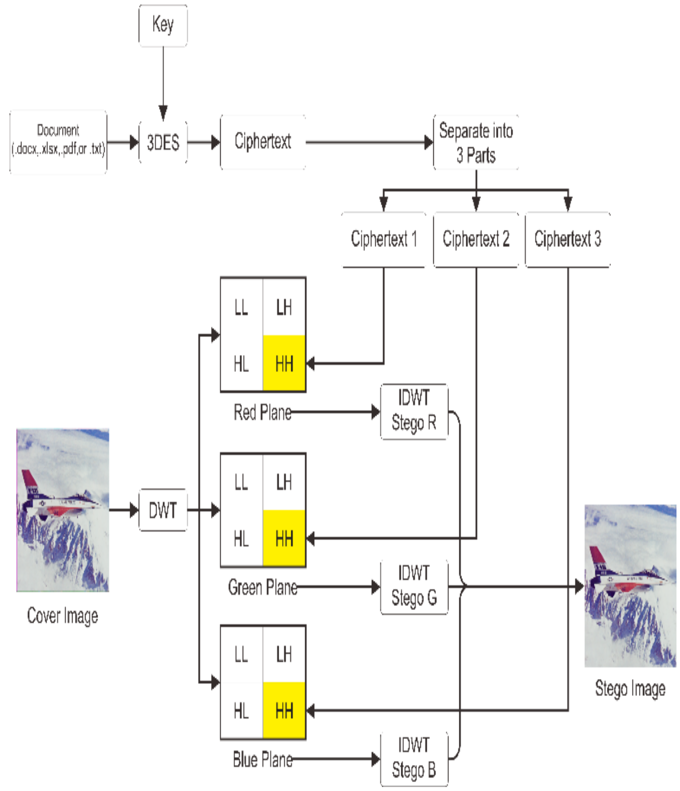


Fig. 4. Embedding Algorithm.

- Embedding Algorithm

In the extraction process requires a stego image and key (K1, K2, and K3). The following are the steps in the extraction process, Fig. 5:

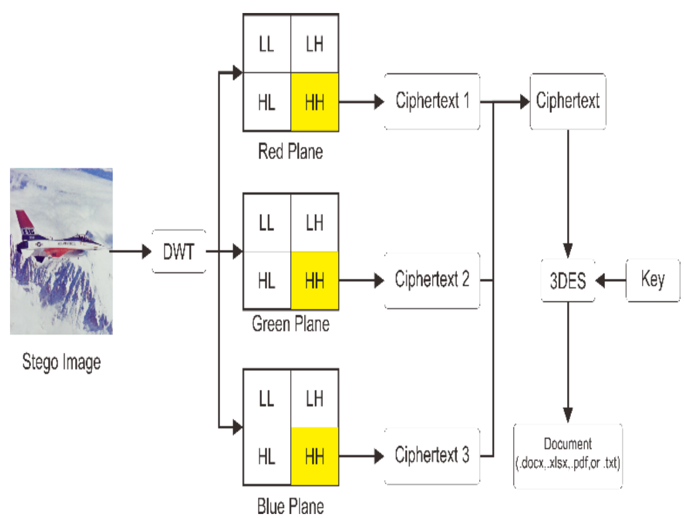


Fig. 5. Extraction Algorithm.

1) Transform the stego image using DWT to get four sub-bands on the R, G, and B layers.

2) Extract information in the HH sub-band on each sheet.

3) Combine the information that has been extracted to get a complete ciphertext.

4) Decrypt ciphertext using 3DES with the key to obtaining the document back.

III. Experiments

In this research, the embedding process uses a few cover images shown in Fig. 6.

Fig. 6. (a) Airplane.png (b) Arctichare.png (c) Baboon.png (d) Boat.png (e) Boy.bmp (f) Cat.png (g) Fruits.png (h) Frymire.png (i) Lena.png (j) Peppers.png [20].

Those images are the images that will be inserted by the document. The types of support materials and cover images shown in Table 1. Before embedding the document into images, the calculation needs to be done to know the maximum integrated document size. The equation to calculate it shown in the equation below.

$$d = (x * y * 3\ 44000/)$$

TABLE I. Types of Supported Documents

No	Dokumen	Cover Image
1	Docx	Png
2	Xlsx	Bmp
3	Pdf	Jpeg
4	Txt	Tiff

where,

d = document size (KB), x = image width, y = image height

The first test is the insertion test of the document. Any documents with different extensions shown in Table I will be inserted in each image to compare the time of insertion of each material in each image. The type of paper used in this test is a 12 KB docx, xlsx with a size of 9.67 KB, a pdf with a volume of 22.3 KB, and a txt with a capacity of 161 bytes. The results of this test shown in Fig. 7. Furthermore, the test results of document extraction from the image which contains the document in Fig. 8.

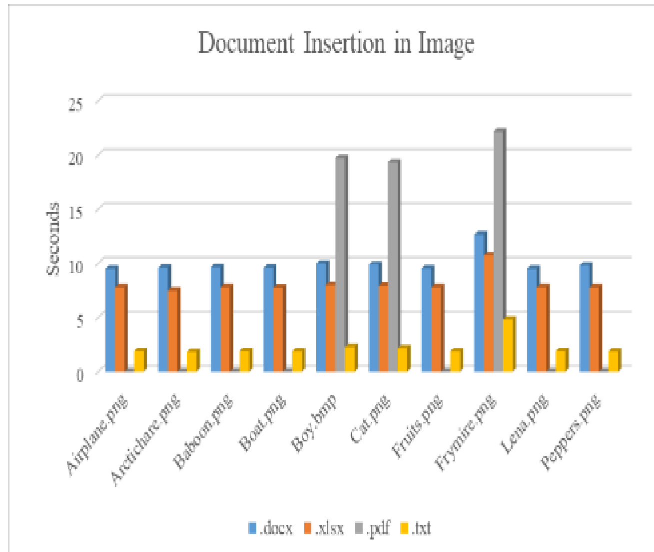


Fig. 7. Graph of Document Insertion in Image.

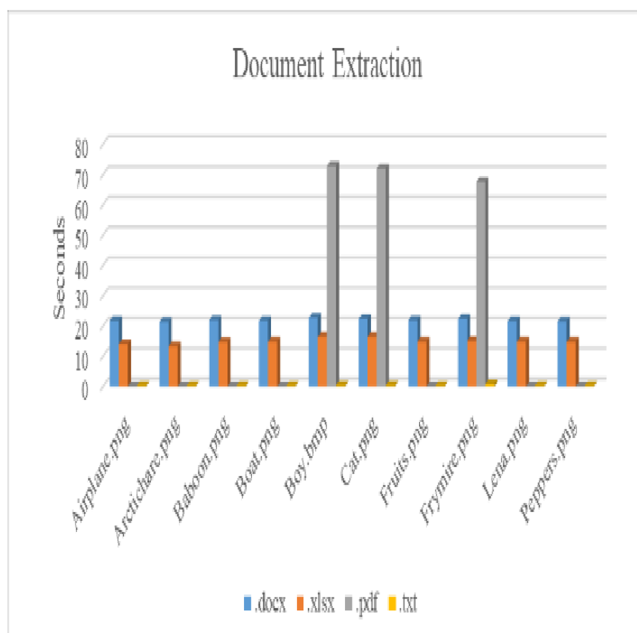


Fig. 8. Graph of Document Extraction.

The next test is testing the image quality on each image of the insertion shown in Table II.

TABLE II. RESULTS OF IMAGE QUALITY TESTING

Images	Image Assesment	Document Types			
		docx	xlsx	pdf	txt
Airplane 512x512	MSE	18	15	-	2
	PSNR	23	25	-	43
Arctichare 512x512	MSE	18	14	-	2
	PSNR	23	25	-	44
Baboon 512x512	MSE	18	16	-	2
	PSNR	23	24	-	42
Boat 512x512	MSE	18	15	-	2
	PSNR	23	25	-	43
Boy 768x512	MSE	14	11	22	1
	PSNR	26	27	21	46
Cat 490x733	MSE	14	12	23	2
	PSNR	25	27	21	45
Fruits 512x512	MSE	17	15	-	2
	PSNR	24	25	-	43
Frymire 1118x1106	MSE	5	5	7	1
	PSNR	34	35	31	52
Lena 512x512	MSE	17	15	-	2
	PSNR	23	25	-	43
Peppers 512x512	MSE	17	15	-	2
	PSNR	23	25	-	43

After testing the image quality has been done, then the next step is insertion testing on the different frequencies shown in Table III.

TABLE III. INSERTION TESTING ON DIFFERENT FREQUENCY

Original Image	Insertion Frequency	Encryption Result	Decryption Result	MSE	PSNR
	LL		Blank Document	11%	27 db
	LH		Original Document	2%	42 db
	HL		Original Document	2%	43 db
	HH		Original Document	2%	43 db

The next test is insertion testing using different insertion values as shown in Table IV.

After that the test of the stego image endurance against some image manipulation attacks shown in Table V.

TABLE IV. INSERTION TESTING ON DIFFERENT INSERTION VALUE

Original Image	Insertion Frequency	Insertion Value	Encryption Result	Decryption Result	MSE	PSNR
Airplane.png	HH	10	airplane.enkripsi.nilai.10.png	*Error*	1%	52 db
Airplane.png	HH	20	airplane.enkripsi.nilai.20.png	*Error*	1%	49 db
Airplane.png	HH	30	airplane.enkripsi.nilai.30.png	*Error*	1%	47 db
Airplane.png	HH	40	airplane.enkripsi.nilai.40.png	*Error*	2%	45 db
Airplane.png	HH	50	airplane.enkripsi.nilai.50.png	Original Document	2%	43 db
Airplane.png	HH	60	airplane.enkripsi.nilai.60.png	Original Document	2%	41 db
Airplane.png	HH	70	airplane.enkripsi.nilai.70.png	Original Document	3%	40 db
Airplane.png	HH	80	airplane.enkripsi.nilai.80.png	Original Document	3%	39 db
Airplane.png	HH	90	airplane.enkripsi.nilai.90.png	Original Document	3%	38 db
Airplane.png	HH	100	airplane.enkripsi.nilai.100.png	Original Document	4%	37 db
Airplane.png	HH	150	airplane.enkripsi.nilai.150.png	Original Document	5%	34 db
Airplane.png	HH	200	airplane.enkripsi.nilai.200.png	Original Document	6%	32 db
Airplane.png	HH	255	airplane.enkripsi.nilai.255.png	Original Document	7%	31 db

TABLE V. TESTING OF STEGO IMAGE RESISTANCE

Original Image	Manipulation	Degree / Area	Encryption Result	Decryption Result
Airplane.txt.png	Brightness	10%	Brightness.10%.png	Original Document
Airplane.txt.png	Brightness	20%	Brightness.20%.png	Original Document
Airplane.txt.png	Brightness	30%	Brightness.30%.png	Original Document
Airplane.txt.png	Brightness	50%	Brightness.50%.png	Original Document
Airplane.txt.png	Brightness	100%	Brightness.100%.png	Original Document
Airplane.txt.png	Brightness	150%	Brightness.150%.png	Original Document
Airplane.txt.png	Brightness	-150%	Brightness.minus150%.png	Original Document
Airplane.txt.png	Contrast	100%	Contrast.100%.png	Original Document
Airplane.txt.png	Contrast	-50%	Contrast.minus50%.png	Original Document
Airplane.txt.png	Crop	Up	UpCrop.png	Original Document
Airplane.txt.png	Crop	Down	Down Crop.png	Blank Document
Airplane.txt.png	Crop	Right	Right Crop.png	“Can’t be Extracted”
Airplane.txt.png	Crop	Left	Left Crop.png	"The Key That You Entered Is Incorrect"
Airplane.txt.png	Resize	400x400	Resize.400x400.png	“Can’t be Extracted”
Airplane.txt.png	Resize	500x500	Resize.500x500.png	“Can’t be Extracted”
Airplane.txt.png	Resize	510x510	Resize.510x510.png	“Can’t be Extracted”
Airplane.txt.png	Resize	514x514	Resize.514x514.png	“Can’t be Extracted”
Airplane.txt.png	Rotate	90^o CW	CW rotation.png	Blank Document
Airplane.txt.png	Rotate	90^o CCW	CCW rotation.png	“Can’t be Extracted”
Airplane.txt.png	Rotate	180^o	Full rotation.png	“Can’t be Extracted”

IV. Conclusions

Based on the results of the previous tests, can be drawn some conclusions that are :

- Encryption and decryption process using pdf documents takes longer than using other materials because it has a bigger size.
- Extraction documents are documents that match precisely the original content.
- The larger the images or, the smaller the document size, the better the quality.
- The insertion of materials in the HL and HH sub-bands produces images of better quality than in the other subbands because they have larger MSE and PSNR values.
- The best insertion value is 50 because it shows minimal MSE values and large PSNR values, as well as documents, can still be extracted.
- The stego image is resistant to brightness and contrast manipulation attacks. As for crop manipulation attacks, stego images cannot always survive. Meanwhile, the stego image cannot withstand the manipulation of resizing and rotate attacks.

Further experimental approaches are required for validation of the proposed data hiding method.

References

[1] P. Passeri, "2017 Cyber Attacks Statistics," 2017. [Online]. Available: https://www.hackmageddon.com/2018/01/17/2017-cyber-attacksstatistics/. [Accessed: 30-May-2018].

[2] Kohei Arai, Kaname Seto, Data Hiding Based on Wavelet Multiresolution Analysis, Journal of Visualization Society of Japan, Vol.22, Suppl.No.1, 229-232, 2002.

[3] Kohei Arai, Kaname Seto, Data Hiding Based on Multi-resolution Analysis Utilizing Information Content Concentrations by Means of Eigen Value decomposition, Journal of Visualization Society of Japan, Vol.23, No.8, pp.72-79,2003.

[4] Kohei Arai, Kaname Seto, Information Hiding Method Based on Coordinate Conversion, Journal of Visualization Society of Japan, 25, Suppl.No.1, 55-58,(2005)

[5] Kohei Arai, Kaname Seto, Data hiding based on Multi-Resolution Analysis taking into account scanning of the embedded image for improvement of invisibility, Journal of Visualization Society of Japan, 29, Suppl.1, 167-170, 2009.

[6] Kohei Arai and Yuji Yamada, Improvement of secret image invisibility in circulation image with Dyadic wavelet based data hiding with runlength coding, International Journal of Advanced Computer Science and Applications,2,7,33-40, 2011.

[7] Kohei Arai, Method for data hiding based on Legal 5/2 (CohenDaubechies-Feauveau: CDF 5/3) wavelet with data compression and random scanning of secret imagery data, International Journal of Wavelets Multi Solution and Information Processing, 11, 4, 1-18, B60006 World Scientific Publishing Company, DOI: I01142/SO219691313600060, 1360006-1, 2013.

[8] Kohei Arai, Data Hiding Method Replacing LSB of Hidden Portion for Secrete Image with Run-Length Coded Image, International Journal of Advanced Research on Artificial Intelligence, 5, 12, 8-16, 2016.

[9] P. Batarius and M. Maslim, "Perbandingan metode dalam teknik steganografi," vol. 2012, no. Semantik, pp. 307–313, 2012.

[10] R. Tripathi and S. Agrawal, "Comparative Study of Symmetric and Asymmetric Cryptography," Int. J. Adv. Found. Res. Comput., vol. 1, no. 6, pp. 2348–4853, 2014.

[11] T. Sobh, K. Elleithy, and A. Mahmood, Novel Algorithms and Techniques in Telecommunications and Networking. 2010.

[12] R. Debnath, P. Agrawal, and G. Vaishnav, "Des , Aes and Triple Des : Symmetric Key Cryptography Algorithm," Int. J. Sci. Eng. Technol. Res., vol. 3, no. 3, pp. 652–654, 2014.

[13] Y. Dinesh and A. P. Ramesh, "Efficient Capacity Image Steganography by Using Wavelets Yedla dinesh A ddanki purna ramesh," Int. J. Eng. Res. Appl., vol. 2, no. 1, pp. 251–259, 2012.

[14] A. Noura, "A comparison of the 3DES and AES encryption standards," Int. J. Secur. its Appl., vol. 9, no. 7, pp. 241–246, 2015.

[15] K. S. Thyagarajan, Still image and video compression with MATLAB. 2011.

[16] N. Kaur and P. Singh, "A New Method of Image Compression Using Improved SPIHT and MFHWT," Int. J. Latest Res. Sci. Technol., vol. 1,8no. 2, pp. 124–126, 2012.

[17] K. Kaur, "Image Compression using HAAR Wavelet Transform and Discrete Cosine Transform," vol. 125, no. 11, pp. 28–31, 2015.

[18] P.-Y. Chen and H.-J. Lin, "A DWT Based Approach for Image Steganography," Int. J. Appl. Sci. Eng. Int. J. Appl. Sci. Eng, vol. 4, no. 4, pp. 275–290, 2006.

[19] D. Coppersmith, "A proposed mode for triple-DES encryption," IBM J. Res. …, vol. 40, no. March, pp. 253–262, 1996.

[20] https://homepages.cae.wisc.edu/~ece533/images/)

Author's Profile

Cahya Rahmad: He received BS degrees from Brawijaya University Indonesia in 1998 and MS degrees from Informatics engineering at Sepuluh Nopember Institute of Technology Surabaya Indonesia in 2005. He is a lecturer in The State Polytechnic of Malang Since 2005. He received Doctoral degrees at Saga University japan in 2013, His interest researches are image processing, data mining and patterns recognition.

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 counselor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 37 books and published 570 journal papers. He received 30 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.ht

An Effective Lightweight Cryptographic Algorithm to Secure Resource-Constrained Devices

Sohel Rana[1]
Department of CSE
Bangladesh University of Business & Technology,
Mirpur-2, Dhaka,
Bangladesh

Saddam Hossain[2], Hasan Imam Shoun[3], Dr. Mohammod Abul Kashem[4]
Department of CSE
Dhaka University of Engineering & Technology, Gazipur
Gazipur, Bangladesh

Abstract—In recent years, small computing devices like embedded devices, wireless sensors, RFID tags (Radio Frequency Identification), Internet of Things (IoT) devices are increasing rapidly. They are expected to generate massive amount of sensitive data for controlling and monitoring purposes. But their resources and capabilities are limited. Those also work with valuable private data thus making security of those devices of paramount importance. Therefore, a secure encryption algorithm should be there to protect those vulnerable devices. Conventional encryption ciphers like RSA or AES are computationally expensive; require large memory but hinder performances of those devices. Simple encryption techniques, on the other hand are easy to crack, compromising security. In this paper a secure and efficient lightweight cryptographic algorithm for small computing devices has been proposed. It is a symmetric key block cipher, employing custom substitution-permutation (SP) network and a modified Feistel architecture. Two basic concepts from Genetic algorithm are used. A Linux based benchmark tool, FELICS is used for the measurement and MATLAB for the purpose of encryption quality testing. An improvement over the existing algorithm, the proposed algorithm reduces the use of processing cycles but at the same time provides sufficient security.

Keywords—Lightweight cryptography; IoT; RFID tags; genetic algorithm; feistel architecture; SP network; FELICS; MATLAB

I. INTRODUCTION

Lightweight cryptography [1] is a sub-category in the field of cryptography that intends to provide security solutions for resource-constrained devices. Cryptography means “secret writing” [2]. In computer communication all want to encrypt information so that no unwanted entity but the expected one can decipher the information. At the core of lightweight cryptography there is a trade-off between security and lightweightness: that is how anyone can achieve a good level of security in small computing devices? Recently, academic communities have been doing a significant amount of work related to lightweight cryptography; to implement conventional cryptography standards efficiently, to design and analyze new lightweight algorithms and protocol. The widespread utilization of small computing devices such as sensors nodes, Radio-Frequency Identification (RFID) tags, industrial controllers and smart cards indicates there have been massive changes in people’s lives. New security and privacy considerations arise as one shift from desktop computer to small devices. It is challenging to implement heavyweight cryptographic standards to small devices [3]. Many conventional cryptographic algorithms, was optimized for desktop and server environments. Optimization in terms of security, performance and resource requirements makes those algorithms difficult or impossible to implement in resource-constrained devices. Even if they can be implemented, they hinder the performance on the small devices. Lightweight cryptography aims at wide variety of hardware and software spectrum in which an algorithm can be implemented. On the device spectrum in Figure 1 for example, servers and desktop computers occupy at the high end [1]. Tablets and smartphones are the next.

Servers and Desktops	Conventional cryptography
Tablets and Smartphones	
Embedded Systems	Lightweight cryptography
RFID and Sensor Networks	

Fig. 1. Device Spectrum.

Conventional cryptographic algorithms inherently perform well in these devices. Embedded systems, RFID devices and sensors networks can be found at the end in the spectrum. Highly resource-constrained devices are at the very end of the spectrum that has very limited processing capabilities and memory. Lightweight cryptography is principally motivated for those.

Microcontrollers of wide array of performance traits are available. 8-bit, 16-bit and 32-bit microcontrollers are more common but use of 4-bit microcontrollers for certain ultra-low cost applications are noticeable. There exist some instruction sets which only contain a small number of simple instructions. When executing common cryptographic algorithm, they take excess number of cycles. The intended application can get slower and energy-consuming. The amount of random-access memory (RAM) and read-only memory (ROM) of certain microcontroller can be very limited; ranging from 64 bytes to as little as 16 bytes. RFID and sensors are often used in applications which require very strict timing and power requirements [3]. They are for only dedicated purpose and their constraints are stringent. The algorithm they need must also fulfill their requirements.

It is important to understand that lightweight cryptography is not necessarily only for the lower end devices of the spectrum. Many resource-constrained devices work with server which is powerful. The server must support lightweight algorithm so that it can interoperate with the devices.

A. Motivation

Cryptography itself a challenging and interesting subject to study and especially to research on. It is impossible to think secure data communication without cryptography. In history, people won wars using cryptography as a weapon. It involves mathematics, algorithm, programming, understanding in data communication, etc. With the widespread use of small low powered devices, lightweight cryptography will play a vital role in future. A survey by HP states that more than 70% of resource-constrained devices are vulnerable [4]. It is necessary to make a balance between the security and performance.

B. Security Challenges in Resource-Constrained Devices

Resource-constrained devices have many application areas: automotive systems, smart parking, sensor networks, disaster/weather forecast, healthcare, distributive control systems, Internet of Things (IoT), cyber-physical systems, smart cities, smart grid, etc. Building the confidence among the user is necessary for the adoption of technology with small computing devices, especially about its privacy and security [5]. Resources constrained devices are intrinsically defenseless to many types of security threats.

Devices in IoT are extremely open to assaults [6]. As they remain unsupervised for long time there is a chance of physical attack on its components. Also eavesdropping is simple because of wireless communication medium. The constituents bear low competency in terms of energy and computational capability. If conventional security algorithms are used which require computations, their performance will be wasted [7]. IoT, used for monitoring purposes generates substantial amount of data, so their integrity and authentication are a matters of concerns.

The confidentiality of the data is retained in secure systems. It is important that data should retain its originality and no intentional or inadvertent changes are undetected by the system [8]. For example IoT is composed of many small devices such as RFIDs which remain unattended for a long time [9]; it is easier for any malicious entity to steal the data stored in the memory.

II. Literature Review

In [10], authors enhance the security of Ceaser cipher including sharing secret key using modified Diffie-Hellman technique. Shared key are made in the following way: Let A has a public key 10 and private key 14. A sends 140 (public key multiplied by private key) to B over unsecure channel. B also has a private key 16, so B sends 160 to A. A generate the value of shared secret key as 140 multiplied with 16 result is 2240. Similarly B generates the key value 160 multiplied with 14 which is same to 2240. They use the mod operation with 26 to get the value less than or equal to 26. For any character in the 'x' position the secret key is simply first multiplied with 'x' and then mod is done to get the cipher character. So 2nd character of the message is multiplied with 2, third character with 3 and so on. Then some light calculation to perform cipher.

Authors in paper [11] analyze the performance and security of different type of lightweight encryption algorithm, which are used in especially resource-constrained applications. Four lightweight algorithms TEA, HIGHT, KATAN and KLEIN are implemented on AVR Atmel ATtiny45 microcontroller to evaluate performance analysis on their memory efficiency and energy consumption and also evaluated degree of confusion and diffusion for security analysis.

In paper [12], the authors propose an encryption technique using simple mathematical operations and trivial authentication using unique id. The algorithm applies encryption on ASCII values. Each receiver has unique id and sender possesses a database of all receivers. A set of three keys are used. First a palindrome number is generated from receiver’s alphanumeric id and four random numbers. From the palindrome number an encoding matrix is generated. Data is encrypted using the encoding matrix and ASCII values of data. The decryption process is done using the inverse of the encoding matrix called the decoding matrix. But the entire process is questionable to security analysis. Here the encrypted data, keys and random number seed are sent to the receiver. It is possible for any intruder to perform a middle-man attack thus making the entire process vulnerable.

In [13], the Authors propose an algorithm based on combined concept of Genetic Algorithm (GA) and pseudorandom number sequence generation. Only two operators (Crossover and Mutation) of GA are used as a part of this algorithm. They used Blum Blum Shub to generate the pseudorandom sequences in order to select the crossover operators among three (single point, two points, and uniform crossover). Also, five keys are used for performing the encryption and decryption process. First key is a number that indicates a size of block to divide the plain text into blocks. Second and third keys are used to generate the random sequences. Fourth key indicates the modulating factor and Fifth key is used for mutation oparation. This algorithm ensures higher performance and security through the concept of GA and pseudorandom sequence generation.

In [14] The authors proposes a symmetric key block cipher that uses 64 bit key over 64 bit data. Block ciphers such as AES uses substitution-permutation (SP) network in order to integrate Shannon’s confusion and diffusion properties. Other ciphers such as Blowfish and DES use Feistel architecture using the advantage of having almost the same encryption and decryption operation.

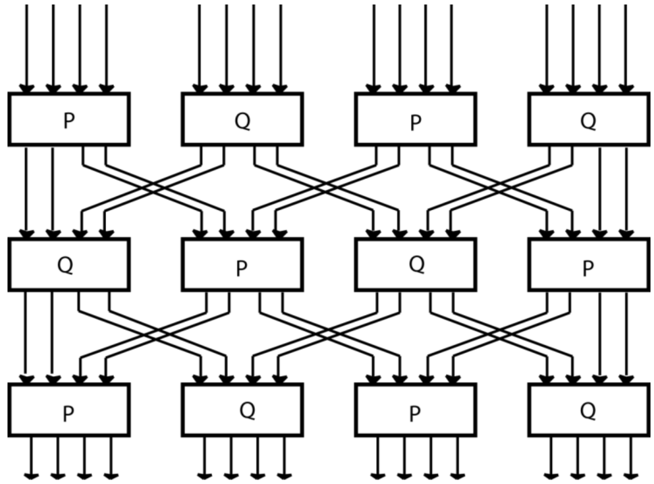


Fig. 2. F-Function of SIT algorithm.

Their proposal is a combination of both Feistel and SP networks (here F-function as SP network as in figure 2), using properties of the both to provide substantial security but keeping the computation complexities as minimum as possible. The algorithm has two parts: key expansion and encryption. A 64 bit key is input by user, divided into 4 blocks, supplied into F-functions, arranged in 4X4 matrices and new five unique keys are generated using some linear and non-linear transformations.

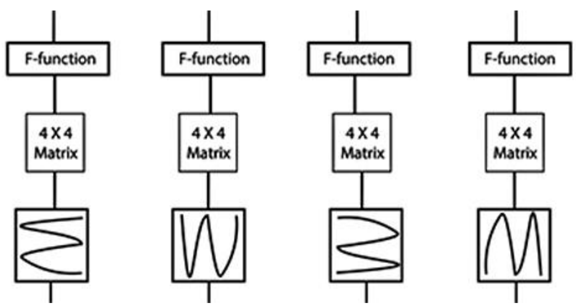


Fig. 3. 4x4 matrices formation during key scheduling.

Input from the f-function forms a matrix and a non-linear transformation occurs as show in figure 3. It can be observed that it is too time consuming and with a little tweak the operation can be minimized using in-place bit shuffling and introducing random number.

The encryption process consists of logical operations, shifting, and substitutions. Although other cipher uses 10 to 20 rounds but it uses Feistel network of 5 rounds that use the five unique generated keys but provides enough confusion and diffusion.

III. Proposed Algorithm

The proposed algorithm is a symmetric key block cipher. It constitutes 64-bit key. In any symmetric key algorithm the encryption process is made up of several encryption rounds. Some mathematical functions define each round to create confusion and diffusion. Increasing number of rounds will ensure better security but will increase the consumption of the device. A typical cryptographic algorithm usually consists of on average 10 to 20 rounds so that the encryption process is strong enough. But the proposed algorithm restricted to only five rounds. The algorithm utilizes the Feistel network. It creates sufficient confusion and diffusion of data so that attacks can be confronted.

The algorithm consists of two parts:

a) Key Scheduling

b) Encryption Process

Key is the most fundamental component in the process of encryption and decryption. The entire security of the data is dependent on the key. The secrecy of the data will be lost if an attacker happens to know the key. Therefore, the revelation of the key should be as difficult as possible. The Feistel network used here consists of five rounds each requiring five unique keys for the encryption/decryption purpose. On figure 4 the key scheduling block is illustrated.

The proposed algorithm requires a 64-bit key. A 64-bit of data can be encrypted or decrypted using that key. In order to guard against exhaustive search attack, the length of the first key must be large enough so that it becomes difficult for the enemy to perform key searching attacks. A cipher key is taken as an input which is 64-bit. The cipher key is input to the key expansion architecture. The block creates five unique keys after going through much confusion and diffusion. The modification that is made from the existing algorithm is shown in the dashed border. Inside the border there are four blocks called non-linear bit shuffling replacing conventional matrix operation. The non-linear bit-shuffling is efficient in creating more confusion and diffusion than the other non-linear operation.

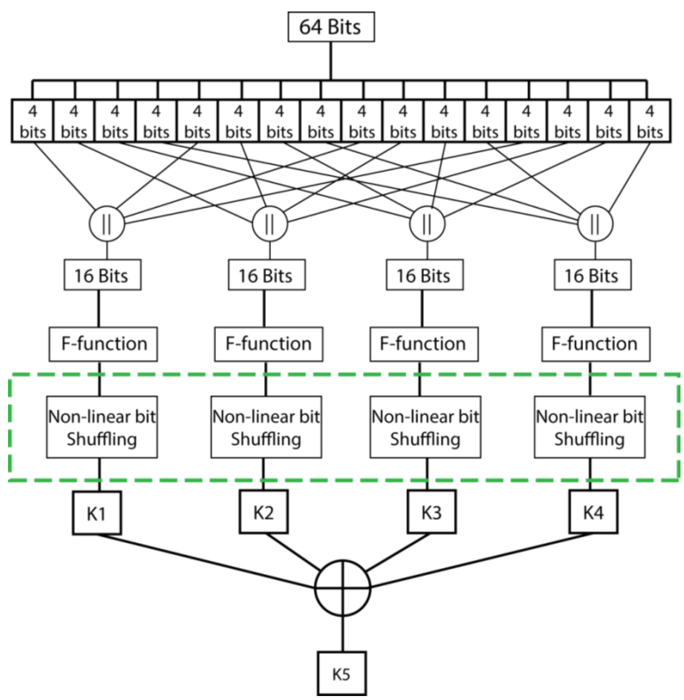


Fig. 4. Key Scheduling

The non-linear bit-shuffling block replaces matrix operation in existing method. A 16-bit input from F-function enters into the block. Taking that 16-bit data as a seed a pseudo random number is generated using linear feedback shift register. Then the two is XORed. The result is transferred to the bit shuffling block that performs a operation as shown in figure 6. The bit shuffling blocks perform an in-place conventional permutation. Figure 5 illustrates how these operations are performed.

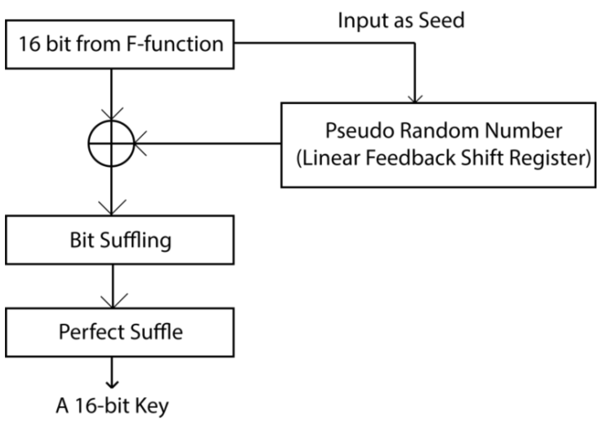


Fig. 5. Process of non-linear bit shuffling block.

Again the output enters into the perfect shuffling block. Figure 7 illustrates how a perfect shuffle is performed.

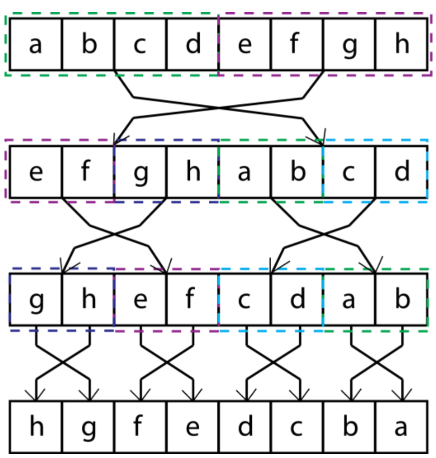


Fig. 6. Bit shuffling.

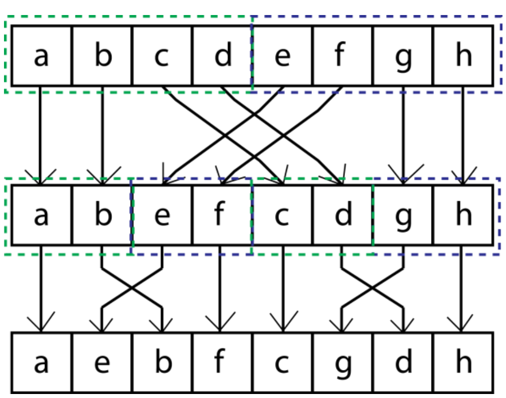


Fig. 7. Perfect shuffle

First four keys, K1, K2, K3, K4 are generated after non-linear bit shuffling. The fifth key K5 is computed by the XOR of the keys K1-K4.

The encryption process encrypts a 64-bit block of data in five rounds using five unique keys generated in the key expansion block. To create considerable confusion and diffusion this process is composed of some shifting, swapping, substitution, XOR, XNOR operations.

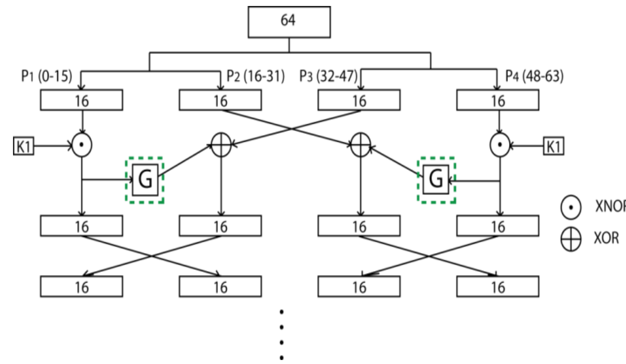


Fig. 8. One of the rounds encryption process.

For the first round an array (figure 8) of 64 bit plain text is first divided into four segments of 16 bits Px_{0-15}, Px_{16-31}, Px_{32-47}, and Px_{48-63}. As the bits progresses in each round the swapping operation is applied so as to diminish the data originality by altering the order of bits, essentially increasing confusion in cipher text. Bitwise XNOR operation is performed between the respective round key K_i obtained earlier from key expansion process. The output of XNOR operation is fed to the modified G-function. The rounds are repeated using the following equations.

$$Ro_{i,j} = \begin{cases} Px_{i,j} \odot K_i & ; \quad j = 1 \text{ and } 4 \\ Px_{i,j+1} \oplus EG_{li} & ; \quad j = 2 \\ Px_{i,j-1} \oplus EG_{ri} & ; \quad j = 3 \end{cases} \tag{1}$$

The results of the final round are concatenated to obtain Cipher Text (Ct).

The encryption process consists of five rounds and uses Feistel architecture. The data block is of 64-bit. The 64-bit data is divided into four 16-bit data. Each round utilizes one key; first round uses first key, second round uses second key and so on. Each key is used twice. In each round the innovated G-function is also used twice. This considerably reduces processing cycles. Figure 9 shows how five rounds of encryption looks like. Please note, after each round data blocks are exchanged except the last round. The decryption process is the opposite of the encryption process. This time last key is used first.

A. G-Function

Two fundamental concepts from genetic algorithm called crossover and mutation are used in the function. That is why the function is named as G-function. Figure 10 illustrates the process of a G-function. The block takes a 16-bit input. The input is divided into two 8-bit blocks. Middle four bits are substituted using a substitution box which is precomputed inside the program. Then a two-point crossover is performed over the two 8-bit blocks. Then a simple mutation is performed. It uses coin flip operation. In coin flip mutation only the first bit is flipped, that is 1 flipped to 0 and 0 is flipped to 1.

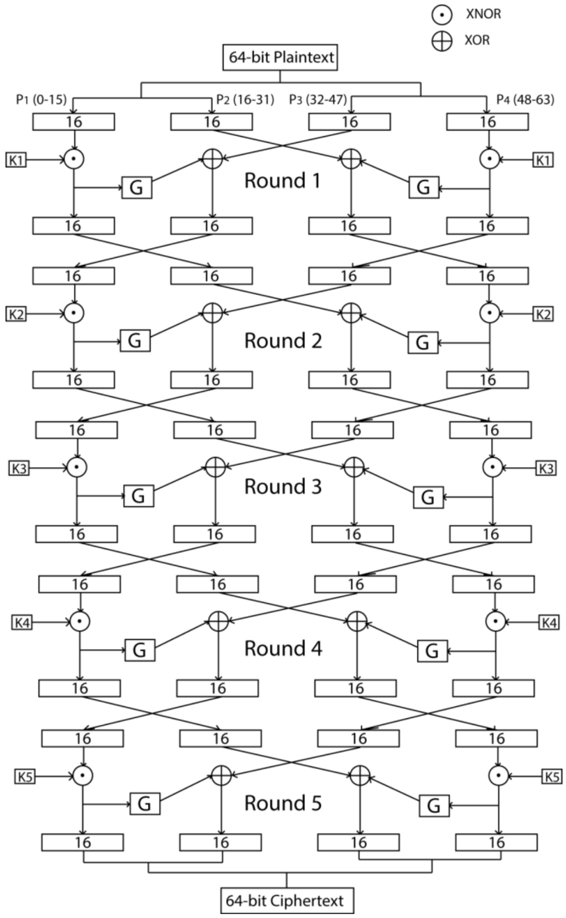


Fig. 9. Five rounds of the encryption process.

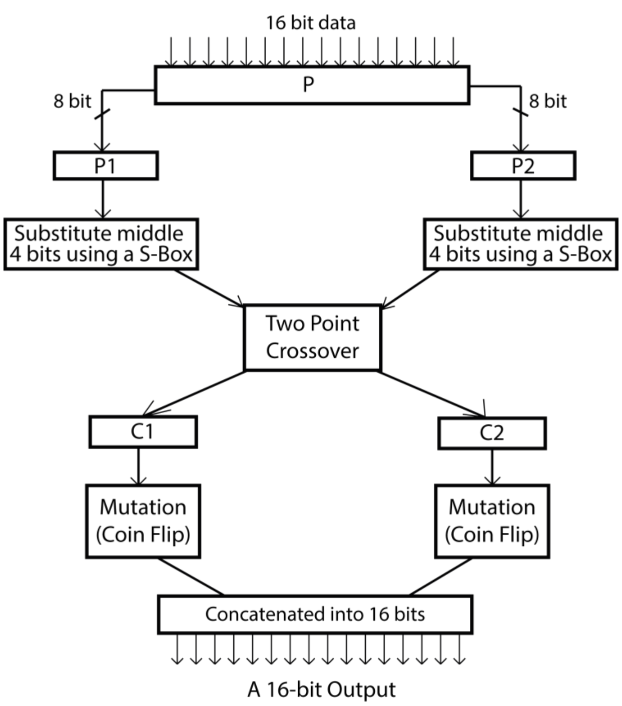


Fig. 10. Process of the G-function.

IV. Experimental Setup

The proposed algorithm is implemented in C programming language. C is an excellent choice for low-level operation as cryptographic algorithms require bit level operations. The algorithm is coded using Visual C++ Express 2010, although CodeBlocks is an excellent choice. The coding was independent of any machine specification. In order to measure execution cycles, memory usage a fantastic benchmark tool called FELICS (Fair Evaluation of Lightweight Cryptographic Systems) [15] is used. This tool incorporates already other standard and popular lightweight cryptographic algorithms like AES, PRESENT, HIGHT, SIMON, SPECK, among others. Many of these are implemented in different version optimized for different consideration in mind. FELICS provides interface to facilitate implementation of any new algorithm and comparing with standard ones. The tool is available to be downloaded. It runs on Linux Ubuntu. A virtual machine file incorporates both Linux Ubuntu and FELICS that saves the user from installing all prerequisites. In the experiment, the virtual machine file is used and that works excellently. The proposed algorithm is also implemented in MATLAB in order to analyze security strength by encrypting images.

The security strength of proposed algorithm is tested to evaluate the basis of following criterion: Key sensitivity, change of cipher entropy, histogram and correlation of the image. Main considerations for observation are the memory utilization and execution cycles for key generation, encryption and decryption of this algorithm.

A. Key Sensitivity

Key sensitivity ensures that the cipher must not decipher to original data if the key has even a bit difference from the actual key. The amount of change occurred in the ciphertext by the change of one bit of the key is evaluated by Avalanche test. According to Strict Avalanche Criterion, the test is to be perfect if 50% of the bits are changed effect of one bit change [16]. To practically observe this effect, an image is decrypted with a key which has only one bit difference from the actual key.

B. Execution Cycle

Most fundamental parameter for the evaluation of algorithm performance is the amount of cycle to perform encoding and decoding a particular data. The proposed algorithm developed for resource-constraint devices in mind must consume minimal cycle and should offer desired security. Execution cycle and power consumption can be correlated, in which case minimizing the cycle also tends to reduce the power consumption.

C. Memory Utilization

Limitation of memory is one of the major challenges for resource-constraint devices. Memory can be measured the number of registers and the number of bytes of RAM and ROM that are used. ROM is used to store the program code and fixed data such as S-boxes and hardcore round keys, while RAM is used to store the computational values. The proposed algorithm uses small amount of rounds that suitable and favorable for its deployments in resource-constraint devices.

D. Histogram

The image histogram is a powerful technique to observe the strength of security for a particular cryptographic algorithm. It is also a basic tool for quality control. However, a histogram can measure the randomness while encrypting an image. A cryptographic algorithm refers to enough secure if the calculated histogram after encryption is uniform.

E. Image Entropy

Image entropy is a quantity which is used to describe the amount of data that must be coded by an encryption algorithm. Higher entropy of an image after encryption refers to the higher security of encryption algorithm. An 8 bits gray scale image can have maximum entropy of 8 bits. Image entropy can be calculated using following equation as below.

$$Entropy,\ \ H(I) = -\sum_{i=1}^{2^8} P(I_i)\log_2 P(I_i) \qquad \text{..........................(2)}$$

Where $\boldsymbol{P(I_i)}$ is the probability that the difference between 2 adjacent pixels is equal to i.

F. Correlation

The Correlation is an effective way to measure the strength of a cryptography algorithm. However, the correlation between two values refers to the dependency. The cipher text of corresponding plaintext has no dependency on its original data or plaintext for an ideal block cipher. Hence, no information can be uncovered from the cipher text only [17]. Correlation coefficients for original and encrypted messages are calculated using the following equation.

$$\Upsilon_{x,y} = \frac{cov(x,y)}{\sqrt{D(x)}\sqrt{D(y)}} \qquad \text{..........................(3)}$$

Where $\boldsymbol{cov\ (x,y)}$, $\boldsymbol{D(x)}$ and $\boldsymbol{D(y)}$ are covariance and variances of variable *x* and *y* respectively. Also, $\boldsymbol{D(\alpha)}$ and $\boldsymbol{cov\ (x,y)}$ can be calculated as follows,

$$D(\alpha) = \frac{1}{N}\sum_{i=1}^{N}(\alpha_i - E(\alpha))^2 \qquad \text{..........................(4)}$$

$$cov\ (x,y) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - E(x)\right) * (y_i - E(y)) \qquad \text{..........................(5)}$$

Where $\boldsymbol{E(x)}$and $\boldsymbol{E(y)}$ are the expected values of variable x and y .Also $\boldsymbol{E(\alpha)}$ can be evaluated using the following equation.

$$E(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\alpha_i \qquad \text{..........................(6)}$$

Where N is the total number of pixels of the image, $\boldsymbol{\alpha_i}$ is a vector of length N and $\boldsymbol{\alpha_i}$ is the ith intensity values of the original image.

FELICES provides an command line interface like gcc (GNU Complier Collection) to test and build any lightweight cryptographic code. They have already provided standard and popular lightweight cipher. In order to test any other cipher the algorithm should be coded in specific format. They've provided documentation to facilitate the implementation. Anyone can compile his implementation and tests whether the algorithm is runnable in FELICS or not. It provides three scenarios against which one can test his code. It is a very convenient and highly advisable tool. Figure 11 illustrates an example of a run.

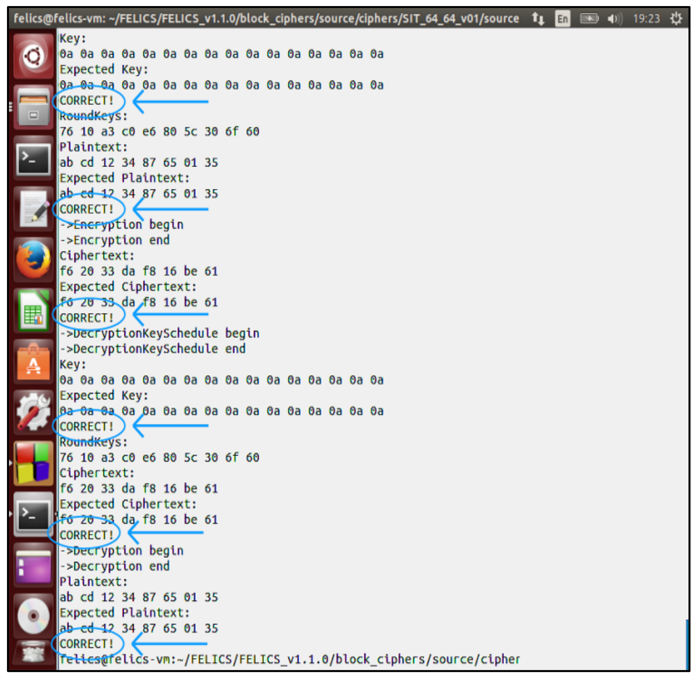


Fig. 11. Testing the implementation of the proposed algorithm on FELICS.

TABLE I. DATA TABLE FOR DIFFERENT CIPHERS IMPLEMENTED ON AVR ARCHITECTURE.

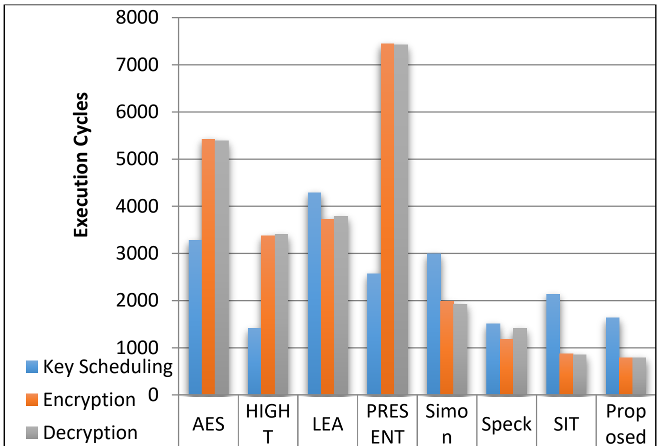
CIPHER	*DEVICE*	*Block Size*	*Key Size*	*Code Size*	*RAM*	*Cycles (key generation)*	*Cycles (encryption)*	*Cycles (decryption)*
AES	AVR	128	128	23090	720	3274	5423	5388
HIGHT	AVR	64	128	13476	288	1412	3376	3401
LEA	AVR	128	128	3700	432	4290	3723	3784
PRESENT	AVR	64	80	1738	274	2570	7447	7422
RC5	AVR	64	128	20044	360	26793	4616	4652
Simon	AVR	64	96	1370	188	2991	1980	1925
Speck	AVR	64	96	2552	124	1509	1179	1411
SIT	AVR	64	64	826	22	2130	876	851
PROPOSED	**AVR**	**64**	**64**	**1228**	**34**	**1630**	**792**	**789**

The simulation of the algorithm is performed by popular open source benchmark tool for lightweight cryptography

FELICS (Fair Evaluation of Lightweight Cryptographic Systems). It uses different platforms (such as AVR, MSP, ARM and PC) for performance evaluation and usually in different conditions. It can also evaluate execution cycles, RAM footprint and binary code size. It can easily compare new cipher with previous one.

Results comparison of different Lightweight Algorithm for Hardware Implementation are shown in TABLE I

In the bar chart in figure 12, comparisons are illustrated among different lightweight algorithm along with the proposed algorithm. The comparisons are made based on number of cycles taken by key scheduling, encryption and decryption individually. The chart shows that the proposed algorithm executes in fewer number of cycles, significantly improving over the others.



	AES	HIGHT	LEA	PRESENT	Simon	Speck	SIT	Proposed
Key Scheduling	3274	1412	4290	2570	2991	1509	2130	1630
Encryption	5423	3376	3723	7447	1980	1179	876	792
Decryption	5388	3401	3784	7422	1925	1411	851	789

Fig. 12. Execution Cycle Comparison for Hardware Implementation.

Two plots in figure 13 and figure 14 demonstrates encryption and decryption cycles of data size between 64 bits and 1024 bits. The proposed algorithm can be seen as a green line taking fewer cycles.

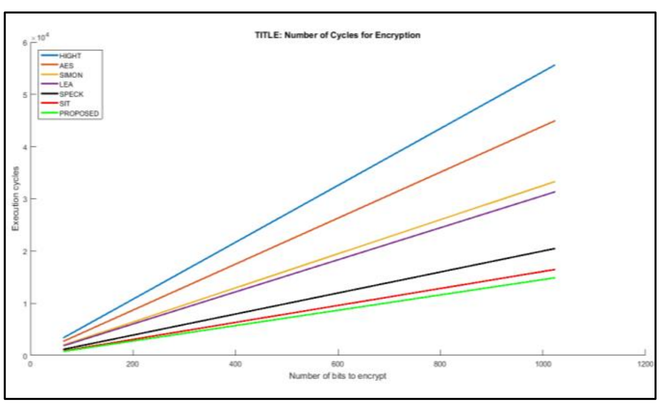


Fig. 13. Execution cycle curve for different cipher in different data sizes for encryption.

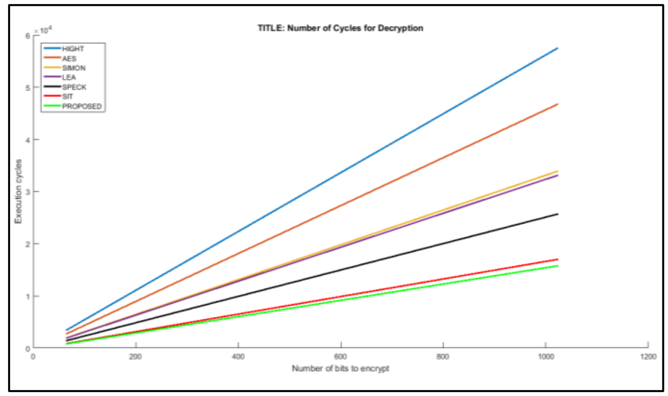


Fig. 14. Execution cycle curve for different cipher in different data sizes for decryption

For a visual observation of encryption-decryption demonstrate the code in MATLAB® which decrypted data using correct key.

The avalanche test of the algorithm, as in figure 15, implies that a single bit change in key, the plaintext brings around 49% changes in cipher bits. The decryption is non-recognizable if even one bit changed in original keys.

Original Image	Encrypted Image	Decrypted with the correct key	Decrypted with wrong key (1 bit difference)

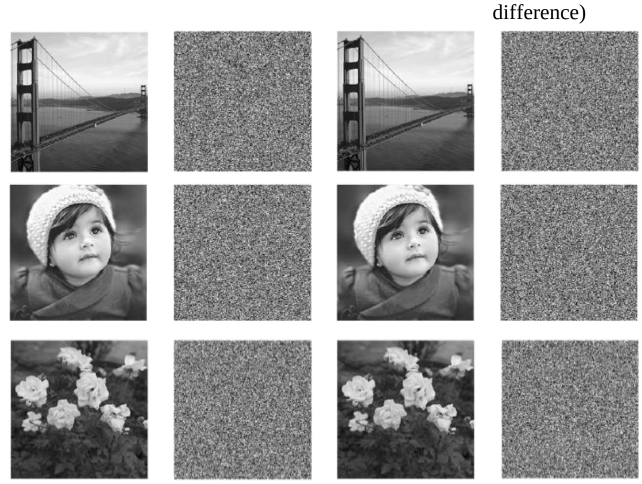


Fig. 15. Analysis of Key Sensitivity

In figures 16, 17, 18, the vertical lines indicate the number of pixels and the horizontal lines indicate the intensity value for each histogram. After encryption, uniform distribution of intensities indicates desired security.

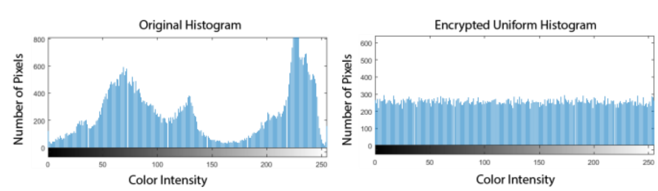


Fig. 16. Bridge histogram

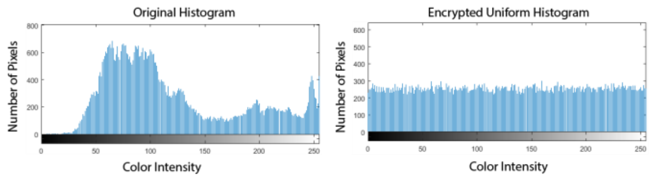


Fig. 17. Child histogram

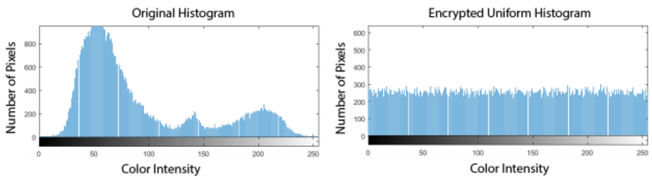


Fig. 18. Flower histogram

The correlation graphs in figures 19, 20, 21 demonstrate the comparison between the original images and the encrypted images. The original image demonstrates highly correlated value whereas the encrypted image seems to have negligible correlated value. Less correlation gives better security for the intended purpose.

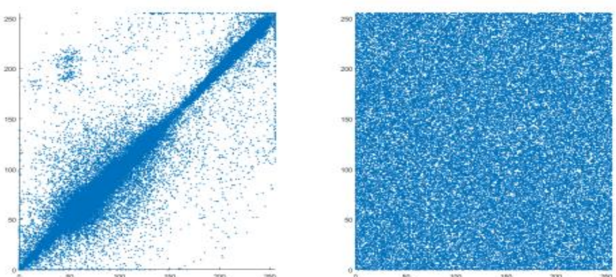


Fig. 19. Bridge correlations for encrypted and decrypted image.

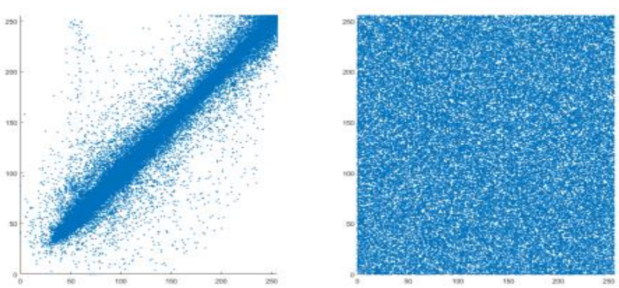


Fig. 20. Child correlations for encrypted and decrypted image.

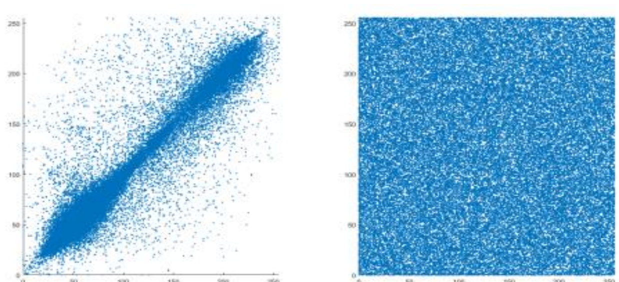


Fig. 21. Flower correlations for encrypted and decrypted image.

The performance of lightweight algorithms in term of memory efficiency is analyzed based on the size of the SRAM and In-System Programmable Flash. Figure 22 compares the memory usages with the existing algorithm. The size of SRAM and In-System Programmable Flash for Atmel ATmega128 microcontroller is 4KB and 128k bytes respectively. The program memory usage bases on the size of Assembly code for each algorithm.

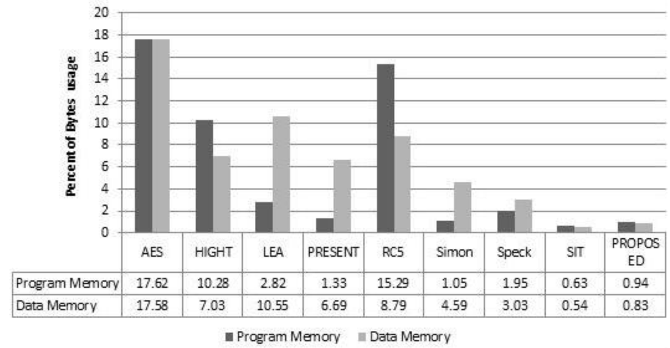


Fig. 22. The comparison of data and memory usage of ciphers.

If the CPU cycle is known, then the energy consumption of the algorithm can be measured. The equation [18] as follows:

$$E = I * VCC * Ʈ * N \tag{7}$$

Here, VCC is the supply voltage of the system and I is the average current in amperes in which is consumed of T seconds. Ʈ is the clock period and N is the number of clock cycle. So clock period is $Ʈ = {}^{1}/_{f}\ sec/cycle$.

Atmel Atmega128 generally uses operating voltage in range of 2.7~5.5, current 40mA on average, and also operates at 16 MHz. The figure 23 shows comparison of power consumption among existing ciphers with the proposed cipher.

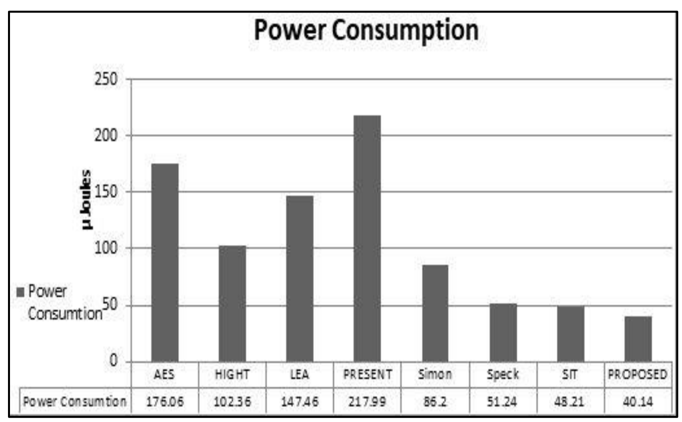


Fig. 23. Energy consumption comparison of ciphers.

V. Conclusion And Future Work

In the near future resource-constraint devices will be essential element of everybody’s daily lives with the blessing of modern electronics and internet. Those devices will be communicating with each other incessantly, so security of the data must be considered. For this purpose an effective lightweight cryptography algorithm proposed in this paper, with a **reduced 16.73% power consumption** than the existing cipher. The implementation shows promising performance making the algorithm a suitable candidate for resource-constraint devices. For future an indomitable challenge is taken to reduce computation cycle for sophisticated resource-constrained devices In-shah-Allah.

REFERENCES

[1] K. A. McKay, L. Bassham, M. S. Turan, N. Mouha, “Report on Lightweight Cryptography”, National Institute of Standards and Technology, USA, March 2017.

[2] B. Schneier, Applied Cryptography: protocols, algorithms, and source code in C, john wiley & sons, 2007.

[3] L. Khelladi, Y. Challal, A. Bouabdallah, N. Badache, “On Security Issues in Embedded Systems: Challenges and Solutions”, International Journal of Information Security, Inderscience, 2008, 2 (2), pp.140-174.

[4] S. A. Kumar, T. Vealey, and H. Srivastava, “Security in Internet of Things: Challenges, solutions and future directions”, in 2016 49th Hawaii International Conference on System Sciences (HICSS), IEEE, 2016, pp.5772-5781.

[5] H.J. Ban, J. Choi, and N. Kang, “Fine-grained support of security services for resource constrained internet of things”, International Journal of Distributed Sensor Networks, vol. 2016, 2016.

[6] P. Wang, Professor S. Chaudhry, S. Li, T. Tryfonas and H. Li, “The internet of things: a security point of view”, Internet Research, vol. 26, no. 2, pp. 337-359, 2016.

[7] S. Wang, Z. Zhang, Z. Ye, X. Wang, X. Lin, and S. Chen, “Application of environmental internet of things on water quality management of urban scenic river”, International Journal of Sustainable Development & World Ecology, vol. 20, no3, pp. 216-222, 2013.

[8] William Stallings. ”Cryptography and Network Security Principles and Practices”, Fourth Edition, Publisher: Prentice Hall, November 16, 2005

[9] T. Karygiannis, B. Eydt, G. Barber, L. Bunn, and T. Phillips, “Guidelines for securing radio frequency identification (RFID) systems”, NIST Special publication, vol. 80, pp. 1-154, 2007.

[10] Shreyank N Gowda,” Innovative Enhancement Of The Caesar Cipher Algorithm For Cryptography” 978-1-5090-3480-2/16/$31.00 ©2016 IEEE

[11] Vikash Kumar Jha,” Cryptanalysis of Lightweight Block Ciphers” Aalto University School of Science Degree Programme of Computer Science and Engineering, Master's Thesis, November 18, 2011

[12] J. Gitanjali, Dr. N. Jeyanthi, C. Ranichandra, M. Pounambal, “ASCII Based Cryptography Using Unique id, Matrix Multiplication and Palindrome Number” School Of Information Technology and Engineering, VIT University, India

[13] S. Dutta, T. Das, S. Jash, D. Patra, Dr. P. Paul, “A Cryptography Algorithm Using the Operations of Genetic Algorithm & Pseudo Random Sequence Generating Functions”, International Journal of Advances in Computer Science and Technology, ISSN 2320 – 2602, Volume 3, No.5, May 2014

[14] M. Usman, I. Ahmed, M. I. Aslam, S. K. and U. A. Shah, “SIT: A Lightweight Encryption Algorithm for Secure Internet of Things”, Iqra University, Defence View and Department of Electronic Engineering, International Journal of Advanced Computer Science and Applications, Vol. 8, No. 1, 2017.

[15] D. Dinu, A. Biryukov, J. Großschädl, D. Khovratovich, Y. L. Corre, L. Perrin, “FELICS – Fair Evaluation of Lightweight Cryptographic Systems”, University of Luxembourg, July 2015.

[16] Webster and S.E. Tavares, “On the design of s-boxes”, in Conference on the Theory and Application of Cryptographic Techniques. Springer, 1985, pp. 523-534.

[17] E. Shannon, “Communication theory of secrecy systems”, Bell system technical journal, vol. 28, no. 4, pp. 656-715, 1949.

[18] MOJTABA ALIZADEH, MAZLEENA SALLEH, MAZDAK ZAMANI, JAFAR SHAYAN, SASAN KARAMIZADEH. “Security and Performance Evaluation of Lightweight Cryptographic Algorithms in RFID”, Faculty of Computer and Information Systems, Advanced Informatics School Universiti Teknologi Malaysia.

A Hybrid Genetic Algorithm with Tabu Search for Optimization of the Traveling Thief Problem

Saad T Alharbi
Computer Science Department
Taibah University
Medina, Saudi Arabia

***Abstract*—Until now, several approaches such as evolutionary computing and heuristic methods have been presented to optimize the traveling thief problem (TTP). However, most of these approaches consider the TTP components independently, usually considering the traveling salesman problem (TSP) and then tackling the knapsack problem (KP), despite their interdependent nature. In this paper, we investigate the use of a hybrid genetic algorithm (GA) and tabu search (TS) for the TTP. Therefore, a novel hybrid genetic approach called GATS is proposed and compared with the state-of-the-art approaches. The key aspect of GATS is that TTP solutions are considered by firmly taking into account the interdependent nature of the TTP subcomponents, where all its operators are simultaneously implemented on TSP and KP solutions. A comprehensive set of TTP benchmark datasets was adopted to investigate the effectiveness of GATS. We selected 540 instances for our investigation, which comprised five different groups of cities (51, 52, 76, 100 and 150 cities) and different groupings of items, from 50 to 745 items. All types of knapsack (uncorrelated, uncorrelated with similar weights and bonded strongly correlated) with all different knapsack capacities were also taken into consideration. Different initialization methods were empirically investigated as well. The results of the computational experiments demonstrated that GATS is capable of surpassing the state-of-the-art results for various instances.**

Keywords*—Combinatorial; hybrid approaches; genetic algorithm; optimization; tabu search; TTP

I. Introduction

The traveling thief problem (TTP) is a benchmark problem recently introduced by [1]. It is an abstraction of real-world problems that consist of multiple components, such as vehicle routing problems and supply chain management. TTP has recently drawn researchers' attention, as its definition represents various aspects of real-world complexities. TTP combines two well-known problems, the traveling salesman problem (TSP) and the knapsack problem (KP). The underlying definition of the problem is that a thief has to visit a set of cities and pick some items from these cities and pack them in a knapsack, where each item has its own weight and value [1]. The challenging aspect of the problem is that the knapsack has a certain capacity, and the total weights of the picked items must not exceed this capacity; the thief also must pay rent for using the knapsack, the rent depending primarily on the total traveling time. Because the two problems (TSP and KP) are interdependent, the speed of the thief decreases when the knapsack gets heavier, which results in an increased total travel time and requires the paying of a higher rent. Therefore, the main objective of TTP is to maximize the total profit of the thief, comprising the total value of the picked items minus the rent of the knapsack.

Despite the TTP only recently being introduced, it has been rigorously considered. Various approaches have been introduced into the literature, adopting different types of techniques and algorithms. For instance, heuristics strategies are among the widely adopted methods for solving TTP, as seen in [2-6]. Searching for the best solutions in such approaches typically involves using classical greedy routines where an initial solution is generated and is iteratively improved. However, adopting heuristics to solve TTP can be computationally complex, especially with a large number of instances [7]. Evolutionary approaches such as genetic algorithms (GAs) and genetic programming have also been adopted for roughly solving the TTP, as in [6, 8-14]. The majority of these approaches try to improve TTP solutions by considering each subproblem (i.e., TSP and KP) independently, despite the interdependence between the subcomponents. Evolutionary operators such as crossover and mutation are normally implemented on tours, and then one of the known packing heuristics is implemented to obtain the best packing plan for the best tours. However, the shortest tours do not necessarily guarantee that the optimal TTP solution will be achieved, due to the nonlinear relationship in the solution's objective function [15]. Other approaches that have occasionally been adopted in the literature to solve TTP include swarm intelligence approaches, such as the ant colony or the artificial bees colony [16, 17].

It has not yet been proven which type of approach is most applicable for solving TTP. Several local search algorithms such as those presented in [18, 19] have been introduced. Most importantly, Packing Routine and PACKITERATIVE, presented by [20], have become key strategies in the literature. The former starts by sorting items according to their weight and then picks the most profitable items. The latter approach is considered an enhanced version of the former, working in the same way but in an iterative manner with some exponent values. Bitflip and Insertion are two local search operators that are also regularly adopted for solutions generated by the previous two methods, to achieve optimal solutions.

TABLE I. MATHEMATICAL REPRESENTATION OF TTP

Equation	**No**	**Description**
$v_{x_i} = v_{max} - C * w_{x_i}$	(1)	The speed (velocity) of the thief, where C is a constant value calculated using Equation 2, and w_{x_i} is the weight of the knapsack at the current city
$C = \frac{(v_{max} - v_{min})}{W}$	(2)	v_{max} is the maximum velocity , v_{min} minimum velocity, W is the total wieght of the knapsack
$g(z) = \sum_m p_m * z_m$	(3)	The total value of the collected items, where P_m is the profit of the item and z_m is a binary value indicating whether the item is available at a particular city (0 denotes that the item is unpicked. while 1 indicates that it is picked)
$f(x,z) = \sum_{i=1}^{n-1} t_{x_i,x_{i+1}} + t_{x_n,x_1}$	(4)	The total travel time from city x_i to x_{i+1}
$G(x,z) = g(z) - R * f(x,z)$	(5)	The objective value (i.e., the total profit gained by the thief)
$t_{x_i,x_{i+1}} = \frac{d_{x_i,x_{i+1}}}{v_{x_i}}$	(6)	The travel time between two cities, where $d_{x_i,x_{i+1}}$ is the distance between city x_i and x_{i+1}

In this paper, we introduce a hybrid GA [21, 22], called GATS, using one of the well-known local search methods, tabu search (TS) [23-25]. GA has been proven powerful in optimizing various types of problems in different domains, such as machine learning [26], network traffic control [27] and industry [28]. Similarly, TS has also been successfully adopted, both alone and hybridized with other approaches, for solving different problems in various fields, as in [29-31]. In fact, to our knowledge, TS has been hybridized with GA for solving various optimization problems, but the hybridization has not been adopted for TTP.

The contribution of this article to the literature is therefore twofold. First, it introduces a novel hybrid approach developed specifically for TTP. The key aspect of this approach is that TTP solutions are considered by firmly taking into account the interdependent nature of the TTP subcomponents. The adopted operators of the proposed approach are simultaneously implemented on tours and packing plans in the process of solution generation; specifically, GA operators, such as crossover and mutation, are implemented to modify tours, while TS is devoted to seeking the best corresponding packing plan. Second, the paper also presents a wide-ranging study taking into account different aspects of TTP, where the performance of the proposed approach was investigated on 540 datasets. These datasets differed in aspects such as size, knapsack capacity and type of knapsack. For instance, the number of cities in these datasets ranged from 51 to 150 cities, and the number of items ranged from 50 to 745 items. All types and capacities of knapsack were tested in this study, and the results were compared with the state-of-the-art approaches.

The rest of the paper is organized as follows. Section II briefly presents the definition of the TTP. In section III, a detailed description of the proposed approach is presented. The experimental design is discussed in section IV, and the proposed approach is tested on various instances and the experimental results are discussed in section VI . The paper concludes in section VII and directions for future work are outlined.

II. THE TRAVELING THIEF PROBLEM

The definition of TTP and its mathematical representation have been well introduced in the literature, for example in [1, 5, 20, 32]. Table 1 shows the mathematical representation of TTP, and we briefly summarize its underlying concept as follows:

- A thief has to travel among a set of cities n, visiting each city only once.
- The tour $\prod$ of the thief must start and end at the same city.
- The tour has a length that can be calculated from the distance $D = \{d_{ij}\}$ between cities.
- Each city contains some items m where each item has weight w_k and value p_k .
- The thief holds a knapsack that has a specific capacity W and a rent R.
- The thief is required to pick the most profitable items from cities during his tour, where the total weight of the picked items must not exceed the knapsack capacity.
- The knapsack rent R is based on the time unit.
- The speed of the thief depends on the knapsack weight, where the thief gets slower when the knapsack becomes heavier.
- The total profit gained by the thief is the total value of the picked items minus the rent.
- The ultimate objective is finding a tour $\prod$ and a packing plan z that maximize the total profit gained by the thief

III. DEVELOPMENT OF THE PROPOSED APPROACH

The aim of this paper is to empirically investigate the use of the GA to solve the TTP. Specifically, the hybridization of the GA with one of the well-known metaheuristic search algorithms, TS, is investigated. Therefore, a hybrid GA with a TS approach, called GATS, is proposed. Figure 1 shows the flow chart of this proposed approach. Its key modules are described in detail in the following sections. For instance,

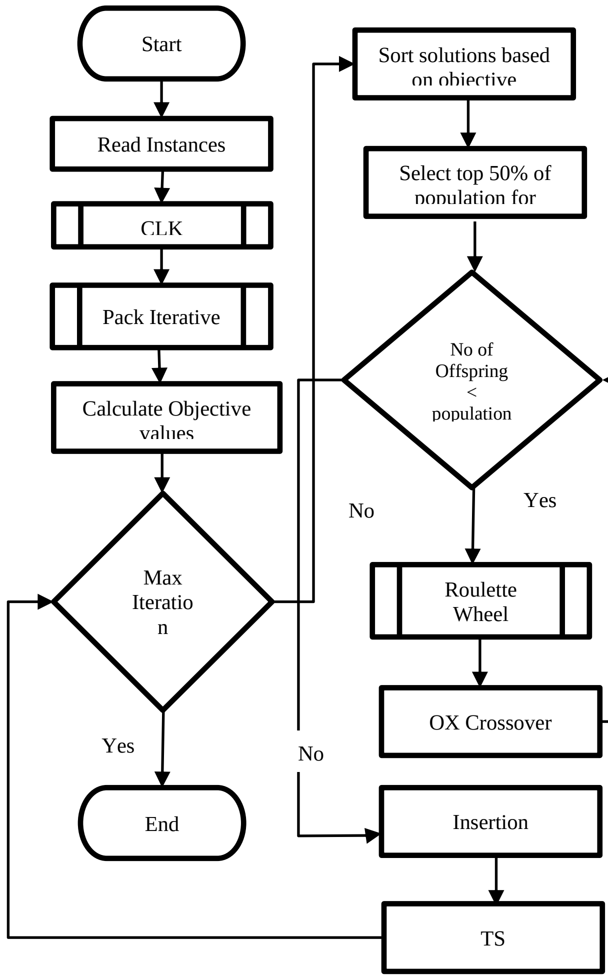


Fig. 1. GATS Flow Chart.

A. Genetic Algorithm for TTP

1) Coding

Tours are encoded directly, with integer numbers indicating city indices in sorted enumeration, starting from a city and ending with the same city. Figure 2 shows an example of a tour consisting of five cities, starting from city 4, traveling to cities 1, 3, 5 and 2 and then returning to the starting city.

n_1	n_2	n_3	n_4	n_5	n_1
4	1	3	5	2	4

Fig. 2. Tour Representation.

Similarly, a packing plan is encoded with an integer enumeration, where the length of the chromosome indicates the number of items, and the integer value in each position represents the number of the city from which the item was picked. Unpicked items are denoted by zero. Figure 3 shows a knapsack solution with 10 items (2 items per city), consisting of three picked items, 2, 3 and 9, at cities 1, 3 and 5, respectively.

Section A shows the adoption of GA and its operators, while Tabu Search C shows how TS was employed in the proposed approach to the TTP.

m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
0	1	3	0	0	0	0	0	5	0

Fig. 3. Packing Plan Representation.

$\prod_1$	9	2	3	8	4	5	6	1	7	9
$\prod_2$	4	5	2	1	8	7	6	9	3	4

Step 1

$\prod c_1$	0	0	0	1	8	7	6	0	0	0
$\prod c_2$	0	0	0	8	4	5	6	0	0	0

Step 2

{9,2,3,4,5}

$\prod c_1$	9	2	3	1	8	7	6	4	5	0

{2, 1,7,9,3}

$\prod c_2$	2	1	7	8	4	5	6	9	3	0

Steps 3 and 4

$\prod c_1$	9	2	3	1	8	7	6	4	5	9
$\prod c_2$	2	1	7	8	4	5	6	9	3	2

Step 5

Fig. 4. An Example of Crossover.

2) Initialization

GA starts with sets of routes and their corresponding packing plans, which are considered to be the initial candidate solutions (i.e., population). Various strategies of initialization were adopted and tested for this paper. Tours were either randomly or using the Chained Lin-Kernighan heuristic (CLK) [33]. The impact of each method was tested, and the results will be discussed in Section VI. The knapsack plan, on the other hand, was initialized using of one the well-known heuristics of TTP, which is PACKITERATIVE [20].

3) Crossover

An order crossover (OX) operator [34] was applied on tours to generate new solutions throughout the algorithm iterations. OX is one of the operators successfully used in combinatorial optimization problems, especially with TSP. In OX, a part of one parent is copied to the child. In GATS, two candidate solutions (tours) are selected using the well-known roulette wheel selection [35] method, which depends mainly on fitness values. Then OX is implemented as follows:

1) *Two random positions in each tour are selected that would be considered the starting and end of the part that will be copied to the new offspring.*

2) *The selected part in the first tour* $\prod_1$ *is copied to the second offspring* $\prod c_2$*, while the sub-tour in* $\prod_2$ *is copied to the first offspring* $\prod c_1$*.*

3) *Cities in* $\prod_1$ *that do not exist in* $\prod c_1$ *are recorded in the same order in which they occur in* $\prod_1$*.*

4) *Blank positions in* $\prod c_1$ *are filled, in order, with the cities recorded from* $\prod_1$*.*

5) *The last city of* $\prod c_1$ *is updated to be the same as the first one.*

6) *Steps 3–5 are repeated for* $\prod c_2$*.*

Figure 4 shows an example of new offspring generation using the OX crossover. Positions 4 and 7 are selected. Accordingly, then, the sub-tour {8, 4, 5, 6} from $\prod_1$ is copied to the second offspring $\prod c_2$, while $\prod c_1$ contains the sub-tour {1, 8, 7,6} from $\prod_2$. Blank positions in $\prod c_1$ are filled with cities {9, 2, 3, 4, 5}, in that order, because they occur in that order in $\prod_1$. Finally, the last city in $\prod c_1$ is updated to be the same as the starting city of the tour. Similarly, blank positions in $\prod c_2$ are filled with cities {2, 1, 7, 9, 3}, in that order, as they occur in

Algorithm 1: OX crossover pseudocode

```
P1 ← Π1  P2 ← Π2;
Pos1 ← r1  Pos2 ← r2 ;
where r1, r2 are random positions in Π1 & Π2 within [2,n-1] such that Pos1 ≠ Pos2
Initialize offspring
child1 ← ϕ; child2 ← ϕ;
child1(Pos1toPos2) ← P2(Pos1toPos2);
child2(Pos1toPos2) ← P1(Pos1toPos2);
Remaining cities from P1, C1 ← P1 − child1;
Remaining cities from P2, C2 ← P2 − child2;
child1 ← ϕ;
child2 ← ϕ;
foreach city in C1 do
    for i = 1 ← n do
        if child1 == ϕ then
            child1 = C1
        end
    end
end
foreach cityinC2 do
    for i = 1 ← n do
        if child2 == ϕ then
            child2 = C2
        end
    end
end
Adjusting last city to be the same of the first
child1(1 : n) ← child1(1 : 1);
child2(1 : n) ← child1(1 : 1);
```

Fig. 5. OX Crossover Pseudocode.

$\prod_1$, *and then the last city is updated to be the same as the starting city of the tour*. Figure 5 *shows the pseudocode of the implemented OX.*

4) Mutation

Opt-Mutation or insertion [34] was applied on the tours in GATS. Here, a predefined mutation probability determines the number of candidate solutions (tours) that will undergo the mutation process. Let us say, for example, that the mutation probability was set at 0.1, and the population (number of candidate solutions) is 100. Then insertion will only be performed on 10 tours, selected randomly. A randomly selected city is inserted into a randomly selected position in the tour. Figure 7 shows an example of insertion mutation on $\prod_2$. Notice that city 8 ($i = 5$) was selected to be moved to be the third city ($k = 3$), and subsequent cities are shifted. Consequently, $\prod_m$ is obtained. Figure 6 shows the insertion pseudocode.

B. Item Packing

In each iteration, after the crossover and mutation operators have performed on tours, the well-known PACKITERATIVE routine [20] is performed on each tour to generate its corresponding packing plan. PACKITERATIVE is well documented in [20], demonstrating good performance results. The main idea behind it is to pack the most profitable items into the pack by sorting them based on a score calculated based on the items' weights. Items are sorted based on their scores, and the algorithm sequentially checks whether adding an item increases the total profit obtained by the thief. After the generation of packing plans for all tours, the proposed GA calculates the objective value of each candidate solution using equation 5.

Algorithm 2: Insertion pseudocode

```
P1 ← Π;
Pos1 ← r1  Pos2 ← r2 ;
where r1, r2 are random positions in Π within [2,n-1] such that Pos1 ≠ Pos2
Initialize offspring
TempOfSpring ← Π;
TempOfSpring(Pos1) ← P1(Pos2);
foreach city bewteen Pos1 and Pos2 in TempOfSpring do
    shift city one postition forward;
end
P ← TempOfSpring;
```

Fig. 6. Insertion Pseudocode.

$\prod_2$	4	5	2	1	8	7	6	9	3	4
$\prod_m$	4	5	8	2	1	7	6	9	3	4

Fig. 7. An Example of Insertion.

C. Tabu Search

Tabu search is one of the known metaheuristic algorithms that has been efficiently employed in solving various optimization problems, especially combinatorial ones. For instance, it has been extensively used to solve classical job-shop scheduling, as in [36] [37], as well as various environmental problems, such as power system planning [38] and transportation [39]. It has also been adopted in optimizing different aspects of recent technology trends such as big data [40]. The main idea of TS is to generate new solutions from the neighborhood of a current solution. Similar to other metaheuristic algorithms, a certain number of iterations are performed to generate new solutions. However, TS selects the best of these. The most important feature of TS is the tabu list, which is used to store subsets of solutions that are not allowed to be visited again, as they would bring the search to areas that have already been visited. This feature helps the algorithm to avoid cycling and getting trapped by local optima. In our proposed GA, TS was adopted on packing plans for each tour to ensure that the best items were picked for the tour. In each iteration of the GA, after performing the crossover and mutation operators, the TS method is called for each tour. The current tour and the initial packing plan are passed to this method. The length of the tabu list is randomly initialized based on the number of items, and the method terminates when it reaches the maximum number of iterations value where the best packing plan (i.e., with highest objective value) is returned. The Bitflip routine introduced by [20] was adopted in the proposed algorithm as a method for searching the neighborhood (new packing plans). However, it was modified to adapt to the proposed algorithm, where all items in the pack are checked in order, and the status of each item is flipped where picked items become unpicked and vice versa. After flipping each item, the resulting packing plan is returned to the TS, and the objective value is calculated. Figure 8 shows the algorithm of TS.

Algorithm 4: Tabu Search pseudocode

```
Sbest ← ϕ;
TSList ← ϕ;
S − Tour ← Π;
S − Pack ← P;
while i < Maxiteration do
    foreach city item in S-Pack do
        NewPack = Bitflip(S − Pack, curentItem);
    end
    if NewPack Not in TSList AND Obj(NewPack) > Obj(S − Pack) then
        Sbest ← NewPack;
    end
    TSList.ADD(Sbest);
end
```

Fig. 8. TS Pseudocode.

TABLE II. INSTANCE CHARACTERISTICS

Parameters	Description
Number of cities	Ranged from 51 to 85900
Type of knapsack problem	- Uncorrelated (U) - Uncorrelated with similar weights (USW) - Bounded strongly correlated (BSC)
Items per city (F)	1, 3, 5 and 10
Knapsack Capacity (C)	Ranged from 1 to 10

TABLE III. BEST OBJECTIVE VALUES OBTAINED USING THE RANDOM AND CLK INITIALIZATIONS

	U		USW		BSW	
C	Random	CLK	Random	CLK	Random	CLK
1	4195	2251	-2770	1323	25033	3714
2	7718	1342	-3674	3410	22479	4639
3	5432	9111	-7162	3587	13055	6999
4	7143	1052	-5433	2260	10199	8866
5	19972	3892	-21274	884	-2714	5448
6	-19612	3737	-3674	1409	-28883	7678
7	-21893	4764	-3674	1478	-27672	9147
8	-19245	6019	-7162	2105	-33525	8607
9	-19611	6922	-21274	3609	-38895	8336
10	-21833.6	3965	-23798	5451	-35153	12386

IV. EXPERIMENTAL DESIGN

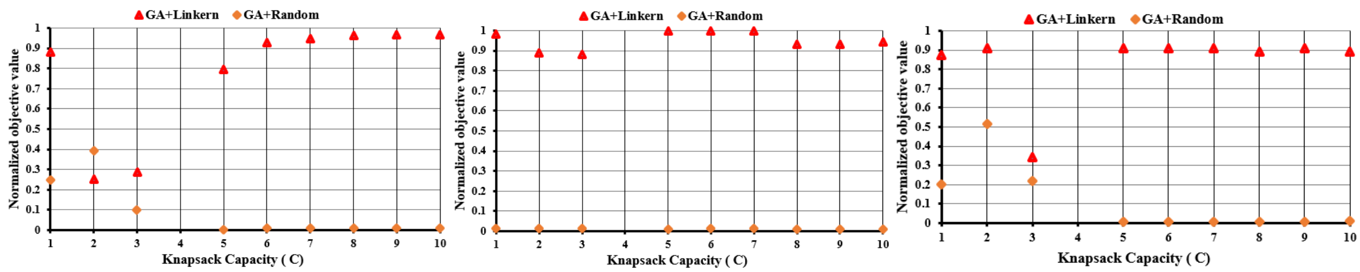
In order to perform an investigation on a TTP, the set of instances defined by [4] should be considered for performance evaluation. These instances were developed in such a way that the two sub-problems (i.e., TSP and KP) were considered. The total number of instances in this set is 9720, and the instances

have different characteristics based on different parameters, for example, the numbers of cities and items. Most of these characteristics are derived from the TSP library dataset [41]. Table 2 highlights the characteristics of the TTP instances. Because of the complexity of performing an experiment on the entire set of instances, a collection of 540 instances was selected for our investigation. These instances consisted of five different groups of cities, 51, 52, 76, 100 and 150 in number, as well as three different numbers of items per city (F), 1, 3, and 5. The selected instances also consisted of the three types of KPs (uncorrelated [U], uncorrelated with similar weights [USW] and bounded strongly correlated [BSC]) and all varieties of knapsack capacity (C). Our implementation was conducted using MATLAB R2014a, and all computations were performed on machines using an Intel Core i7-4790S 3.20 GHz processor and 12 GB RAM, running Windows 8.

Before performing the experiment, several parameters were determined. For instance, the maximum running time for an instance was 10 minutes. In addition, due to the randomization of evolutionary approaches, each instance would be tested 10 independent times. The results obtained from our experiment were compared with the best objective values obtained in the literature. At the beginning of the experiment, several variations on the proposed approach were empirically tested; for example, we tested the effect of the random initialization of tours and investigated the performance of GA with CLK initialization. The aim of these investigations was to adopt the best obtained methods in our approach. Then the proposed approach was tested, and the results were compared with some of the state-of-the-art approaches.

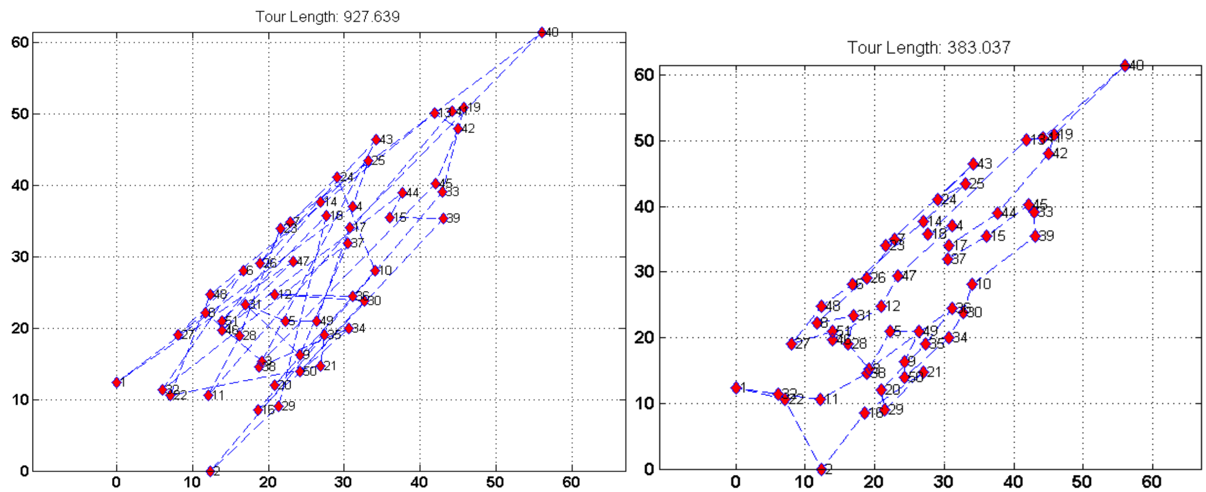
V. INITIALIZATION METHOD

In order to investigate the effect of the tour initialization method on the quality of the obtained solutions, two methods were tested, the basic random tour initialization and the CLK heuristic. These methods were implemented within a classic GA for solving TTP. The performance of the two methods was compared on 30 instances with relatively small numbers of cities and items. Surprisingly, the basic random method obtained higher objective values in some instances when compared with CLK, especially with a small knapsack capacity ranging from 1 to 3 and particularly with U and BSC knapsacks (see Table 3). In order to obtain an accurate result, a normalized objective value was calculated for each method in all instances, taking into consideration the values obtained from each run. This normalized value was calculated by taking the ratio between the best objective value for an instance and the average of the objective values for all runs for this instance. Figure 9 shows the results of the comparisons for the three types of knapsack (U, USW and BSC). It is clear that the CLK used for tour initialization is significantly better than the basic random method. Figure 10 also shows an example of a tour generated for an optimal solution obtained by each method; the obtained tour using CLK was significantly (more than 50%) shorter than that obtained by random means.



(a) (b) (c)

Fig. 9. Normalized objective values for random and CLK initializations in a) U, b) USW and c) BSW.



(a) (b)

Fig. 10. Tours of an optimal TTP solution obtained by a) random and b) CLK initializations.

TABLE IV. BEST OBTAINED OBJECTIVE VALUES FOR U KNAPSACK AND F=1

Instance	n	m	C										
				1	2	3	4	5	6	7	8	9	10
eil51	51	50	GATS	**2251**	**13420**	**9111**	**1596**	**3892**	**3737**	**3965**	**4764**	**6019**	**6922**
			EA	1720	3983	4162	1444	2997	3313	3201	3929	5297	6130
			RLS	1533	3983	4162	1444	2963	3313	3201	3929	5297	6130
berlin52	52	51	GATS	1354	2261	4335	3934	4265	5582	6280	6944	7752	7545
			EA	**2330**	**3220**	**4748**	**3963**	**4141**	**5372**	**6541**	**7532**	**8060**	**7979**
			RLS	2003	**3220**	**4748**	**3963**	**4141**	5356	**6541**	**7532**	**8060**	**7979**
eil76	76	75	GATS	**4528**	**6101**	**4467**	**4663**	**6355**	**7425**	**7161**	**7134**	**9013**	**9457**
			EA	3727	5173	4176	3958	5292	6175	5968	6078	7768	8825
			RLS	3412	5012	4114	3958	5292	6150	5968	6061	7768	8825
kroA100	100	99	GATS	**2790**	**6180**	**6246**	**10318**	**11362**	**12001**	**12508**	**14712**	**15662**	**16725**
			EA	1410	4437	5359	8104	8255	9069	9560	12084	13072	14141
			RLS	1193	4409	5357	8104	8255	9069	9560	12084	13072	14141
pr124	124	123	GATS	**1953**	**7410**	11031	14092	13299	**16996**	**16329**	**17096**	**18470**	**19378**
			EA	2180	7214	**11406**	**15324**	**14092**	16766	15436	16316	17659	18544
			RLS	2180	7175	**11406**	15315	**14092**	16766	15406	16316	17659	18546
ch150	150	149	GATS	3007	8645	11335	10207	12107	14304	15926	**17621**	**18860**	**20595**
			EA	**5033**	**10805**	**12964**	**10972**	**12402**	**14767**	**16182**	17475	18271	20171
			RLS	4982	**10805**	12890	**10972**	12393	**14767**	**16182**	17475	18271	20171

TABLE V. BEST OBTAINED OBJECTIVE VALUES FOR USW KNAPSACK AND F=1

Instance	n	m	C										
				1	2	3	4	5	6	7	8	9	10
eil51	51	50	GATS	**1323**	**3490**	3864	2354	1281	1726	**2183**	**3060**	**4114**	**5909**
			EA	1238	3318	**3890**	**2531**	**1373**	**1933**	2133	2977	3894	5420
			RLS	473	2035	3182	2486	**1373**	**1933**	2097	2977	3894	5404
berlin52	52	51	GATS	**1356**	**2605**	**3502**	**2890**	3647	3623	4345	5927	5785	6676
			EA	**1350**	2516	3254	2763	**3975**	**3847**	**4817**	**6716**	**6678**	**7231**
			RLS	754	1503	1831	2663	3965	3796	**4817**	**6716**	6674	7226
eil76	76	75	GATS	**1344**	1574	**2212**	**2481**	3556	1468	**3960**	**5496**	**7687**	**8166**
			EA	1227	**1717**	1847	2449	**3560**	**1601**	3722	5276	7360	7807
			RLS	-27.9	1470	1847	2284	3545	**1601**	3722	5282	7360	7807
kroA100	100	99	GATS	**2247**	**5784**	**7538**	**8276**	7821	**10097**	**11778**	**12903**	**14490**	**15607**
			EA	1642	5066	6994	7855	**7904**	9925	11187	11675	13146	1642
			RLS	482	4510	6994	7775	7893	9915	11187	11675	13146	482
pr124	124	123	GATS	**4151**	**6669**	**6594**	8507	**10743**	12619	14325	16519	16814	**17950**
			EA	3833	6474	6224	**8725**	10622	**12852**	**14533**	**16566**	**17010**	17942
			RLS	2717	6174	6152	8661	10593	12845	**14533**	**16566**	**17002**	17942
ch150	150	149	GATS	3187	3893	4680	6380	10005	11523	12555	16806	17504	19475
			EA	2786	4033	5110	6973	10034	12104	12919	16638	17048	18445
			RLS	696	4033	4914	6954	10019	12112	12919	16638	17048	18445

TABLE VI. BEST OBTAINED OBJECTIVE VALUES FOR BSC KNAPSACK AND F=1

Instance	n	m	C										
				1	2	3	4	5	6	7	8	9	10
eil51	51	50	GATS	**7046**	**5705**	**6210**	**5633**	**5763**	**7752**	**9209**	**8619**	**8431**	**12604**
			EA	3669	5076	4918	5528	3920	5701	7120	5980	5576	9613
			RLS	2077	3922	4383	5037	3918	5497	7103	5978	5576	9574
berlin52	52	51	GATS	3257	4063	4206	6421	8785	6834	8944	**11481**	**13777**	**10591**
			EA	**3927**	**5634**	**5393**	**7419**	**9884**	**7900**	9706	9048	11131	7875
			RLS	3104	5319	4978	7014	9524	7524	**9325**	9035	11131	7868
eil76	76	75	GATS	2505	5591	9698	**8990**	**6082**	**10586**	**9868**	**12179**	**9773**	**12999**
			EA	**3353**	**6220**	**9750**	8517	5469	9961	9779	11001	9238	10033
			RLS	2965	4891	9216	7855	5410	9678	9757	10973	9009	9920
kroA100	100	99	GATS	**3519**	**8510**	**12545**	**16345**	**21823**	**24768**	**27075**	**27222**	**27545**	**27733**
			EA	2936	7607	8463	10024	13983	14752	15548	14710	14665	14740
			RLS	3185	7233	7632	9824	13636	14728	15285	14456	14631	14762
pr124	124	123	GATS	**5279**	**9868**	**15891**	**22874**	**26369**	**27065**	**31564**	**31509**	**31748**	**28895**
			EA	4918	5977	8981	12518	15501	16410	21540	22435	23838	21770
			RLS	4610	5584	8744	11769	14618	16237	21517	22434	23838	21748
ch150	150	149	GATS	6003	9361	**14416**	**19712**	**21516**	**21474**	**26198**	**30004**	**28441**	**21983**
			EA	**7060**	**9930**	13363	17123	18267	17915	22443	25924	23967	17133
			RLS	5540	7890	13138	16969	18205	17915	22446	25736	23898	17139

TABLE VII. BEST OBTAINED OBJECTIVE VALUES FOR U KNAPSACK AND F=3

Instance			C										
	n	m		1	2	3	4	5	6	7	8	9	10
eil51	51	150	GATS	**12757**	6862	**10408**	**11631**	**10125**	**11404**	**12407**	**14898**	**18978**	**20940**
			EA	5866	8190	9496	6704	7405	9110	10039	12848	15883	18783
			RLS	5644	**8224**	9548	6706	7367	9110	10039	12848	15883	18783
berlin52	52	153	GATS	5653	6243	11573	13346	14434	12841	14244	15958	16055	18265
			EA	**9550**	**9446**	**14467**	**15717**	**15453**	**13783**	**16367**	**18420**	**18479**	**20755**
			RLS	**9550**	**9446**	**14467**	**15717**	**15453**	**13783**	**16367**	**18420**	**18479**	**20755**
eil76	76	225	GATS	9885	10755	9793	12910	**15712**	**18468**	**20719**	**21363**	**26461**	**26235**
			EA	**11215**	**11577**	**10285**	**13518**	14881	18011	19134	21004	25645	25696
			RLS	11153	11567	**10285**	**13518**	14881	18011	19135	21004	25645	25696
kroA100	100	297	GATS	5958	15285	22481	25766	**29504**	**31782**	**34282**	**40355**	**40065**	**45303**
			EA	**10305**	**19318**	**25500**	**26306**	26622	28964	32229	38375	37576	41997
			RLS	10294	**19318**	**25500**	**26306**	26622	28959	32229	38375	37576	41997
pr124	124	369	GATS	9438	15284	25219	30209	36136	39723	42801	47099	51937	57657
			EA	**16303**	**23105**	29438	37560	42888	**42589**	**43700**	**49025**	**52550**	**56691**
			RLS	**16303**	23068	**29462**	**37572**	**42899**	**42589**	**43700**	**49029**	**52550**	**56691**
ch150	150	477	GATS	5086	17325	26264	28929	31472	37347	40510	46622	51966	55714
			EA	16721	28115	35162	34578	37215	41684	44664	49488	54138	57236
			RLS	16721	28136	35162	34572	37217	41684	44683	49488	54138	57236

TABLE VIII. BEST OBTAINED OBJECTIVE VALUES FOR USW KNAPSACK AND F=3

Instance	n	m	C										
				1	2	3	4	5	6	7	8	9	10
eil51	51	150	GATS	**4116**	6622	8109	5967	6789	8302	9943	**12327**	**13402**	**16159**
			EA	3798	**7325**	**9966**	**6807**	**7657**	**8472**	**10397**	11070	12804	14844
			RLS	2545	7208	9879	6774	7653	**8452**	**10397**	11070	12804	14841
berlin52	52	153	GATS	3708	4905	3089	2231	5447	7705	10469	14224	15148	19844
			EA	**5122**	**8317**	8492	**7680**	**11501**	**12556**	**15218**	**18025**	**20323**	**23209**
			RLS	3002	5232	**8505**	**7680**	**11501**	**12556**	**15211**	**18025**	**20323**	**23209**
eil76	76	225	GATS	4860	3644	4897	6716	10454	**8483**	13552	**16530**	**21883**	**24085**
			EA	**5411**	**4969**	6958	**8684**	**10717**	8819	**14067**	16440	20802	23165
			RLS	5112	4898	**6971**	8653	**10717**	8814	**14067**	16440	20802	23165
kroA100	100	297	GATS	**8155**	11858	14149	15844	18600	22103	25725	**31424**	**39119**	**42713**
			EA	7145	**13163**	**15424**	17418	**20282**	**23721**	**27051**	31370	38089	39463
			RLS	6886	13020	15387	**17421**	**20282**	**23721**	**27051**	31370	38089	39463
pr124	124	369	GATS	11919	15691	20015	22816	27333	32262	36064	**43836**	**47830**	**55714**
			EA	**13643**	**20347**	**25838**	**27580**	**32387**	**37483**	**40047**	43151	47270	52732
			RLS	12388	20190	25787	27575	**32387**	37480	**40047**	43150	47272	52732
ch150	150	477	GATS	6144	8269	10327	13749	19747	23968	27928	35052	41202	47371
			EA	**9117**	**14249**	**17855**	**23513**	**30111**	**34003**	**35413**	41297	**45660**	**49657**
			RLS	5998	14221	17838	23497	30094	**34003**	**35413**	**41301**	**45660**	**49657**

TABLE IX. BEST OBTAINED OBJECTIVE VALUES FOR BSC KNAPSACK AND F=3

Instance	n	m	C										
				1	2	3	4	5	6	7	8	9	10
eil51	51	150	GATS	**14794**	10763	15472	**21741**	22798	20715	**21456**	**21686**	**25276**	**27136**
			EA	6664	**12796**	**15786**	21558	**22913**	**21012**	21188	20677	23414	24114
			RLS	4487	11673	15037	21337	22366	20957	21154	20647	23382	24124
berlin52	52	153	GATS	5590	10577	12932	21148	20540	26319	25721	**39668**	**35159**	31443
			EA	**9652**	**16485**	**19684**	**28125**	**26488**	31150	**30211**	34029	30659	31700
			RLS	6600	14304	18337	27379	26344	**31352**	29934	34021	30509	**31784**
eil76	76	225	GATS	6393	14296	24487	**26867**	**31103**	**38152**	**40316**	**43023**	**39945**	**45623**
			EA	**7644**	**14692**	22538	22502	25785	31633	32611	35101	30232	34183
			RLS	6806	14687	**22561**	22468	25669	31571	32584	35138	30226	34190
kroA100	100	297	GATS	12447	**29433**	**42359**	**52054**	**62991**	**69103**	**72403**	**68952**	**63789**	**67504**
			EA	**13725**	25080	32678	37156	43668	46954	48207	42549	36035	39067
			RLS	12600	24664	32720	37043	43618	46969	48211	42544	36017	39076
pr124	124	369	GATS	13088	**35206**	**52693**	**74355**	**83716**	**97891**	**102660**	**101580**	**90070**	**98197**
			EA	**16217**	29193	41810	57901	64792	79009	84810	82360	71276	78945
			RLS	15539	28777	40657	57477	64361	78972	84866	82291	71265	78899
ch150	150	477	GATS	16124	35081	**53125**	**60824**	**68400**	**75684**	**76558**	**71987**	**75781**	**69198**
			EA	**19600**	**37596**	52111	53592	57974	62374	61093	54987	55871	49642
			RLS	16039	36403	52106	53533	57939	62419	61021	54922	55891	49642

TABLE X. BEST OBTAINED OBJECTIVE VALUES FOR U KNAPSACK AND F=5

Instance	n	m	C										
				1	2	3	4	5	6	7	8	9	10
eil51	51	250	GATS	6256	13322	13482	15624	**18430**	**20628**	**19378**	**22934**	**26390**	**31520**
			EA	10683	16345	13885	**16367**	16375	19057	17871	21961	25208	29508
			RLS	**10688**	**16380**	**13887**	16366	16375	19057	17871	21963	25208	29508
berlin52	52	255	GATS	9805	13413	19636	22637	26302	27357	31294	30881	29606	35997
			EA	**18049**	**21247**	**26153**	**29547**	31348	**32109**	**36819**	**35942**	**36584**	**39386**
			RLS	18029	21222	**26153**	**29547**	**31354**	**32109**	**36819**	**35942**	**36584**	**39386**
eil76	76	375	GATS	7268	13883	16258	18624	24665	27517	31973	35890	42916	**46714**
			EA	**13023**	18070	**19554**	**22349**	**26986**	28412	**32510**	**36327**	**43008**	44770
			RLS	**13023**	**18080**	**19554**	**22349**	**26986**	**28424**	**32510**	**36327**	**43008**	44770
kroA100	100	495	GATS	4084	18049	31563	41507	50054	**52219**	**55082**	**64904**	**69106**	**78315**
			EA	**13670**	**27789**	**38737**	**45781**	51076	51588	53633	62457	66117	73998
			RLS	13674	**27789**	38726	45764	**51086**	51592	53633	62457	66115	73999
pr124	124	615	GATS	21880	32241	40368	45632	56115	65422	69301	76267	**91335**	**101453**
			EA	28416	**40811**	**45873**	**58503**	66018	70346	**76155**	**78626**	84805	91399
			RLS	**28422**	**40811**	**45873**	**58503**	**66029**	**70351**	**76155**	**78626**	84805	91399
ch150	150	745	GATS	22475	34033	37762	39908	57564	58785	63270	56226	**85023**	**92916**
			EA	33047	50046	**55532**	58683	**58739**	59979	**64561**	72084	79610	87000
			RLS	**33052**	**50049**	**55532**	**58688**	**58739**	**59985**	**64561**	**72085**	79610	87000

TABLE XI. BEST OBTAINED OBJECTIVE VALUES FOR USW KNAPSACK AND F=5

Instance	n	m	C										
				1	2	3	4	5	6	7	8	9	10
eil51	51	250	GATS	4835	7312	8816	9449	10831	14091	17749	17645	21245	28107
			EA	**5285**	9121	**11140**	**10455**	**12878**	**15574**	**18683**	**18184**	**21343**	**26483**
			RLS	3811	**9206**	11093	**10445**	**12878**	**15574**	**18683**	**18184**	**21343**	**26483**
berlin52	52	255	GATS	6097	6775	5847	6338	9443	15392	18098	22447	28650	33515
			EA	**9178**	**13505**	15700	**17132**	20587	**22873**	**26359**	31739	**36054**	**40537**
			RLS	6079	11347	**15714**	17085	**20590**	**22873**	26363	**31741**	**36054**	**40537**
eil76	76	375	GATS	8245	5761	10534	12509	14123	14392	20804	24840	30082	38073
			EA	**9883**	10290	13293	**18532**	19944	18801	**24721**	**27858**	**32213**	**36553**
			RLS	9867	**10368**	**13329**	18508	**19957**	**18806**	**24721**	**27858**	**32213**	**36553**
kroA100	100	495	GATS	7875	13426	18677	23964	29733	36743	43030	48959	**63755**	**70255**
			EA	**10530**	21104	**27042**	**33303**	**38855**	45102	**49565**	**52726**	63657	66751
			RLS	10401	**21145**	26978	**33303**	38846	**45105**	**49565**	**52726**	63657	66751
pr124	124	615	GATS	13735	18155	23084	29777	34953	52432	47900	58064	65419	94974
			EA	19906	26311	33455	37740	44301	52962	60710	65982	74340	85794
			RLS	16085	26311	33444	37740	44299	52961	60710	65982	74340	85795
ch150	150	745	GATS	7397	10199	13604	29700	36421	44233	**62350**	63635	65498	**95766**
			EA	**15609**	**24436**	29324	33000	**40922**	**47562**	55701	**64278**	**74430**	83275
			RLS	10107	24427	**29341**	**33004**	40919	47560	55701	**64278**	**74430**	83275

TABLE XII. BEST OBTAINED OBJECTIVE VALUES FOR BSC KNAPSACK AND F=5

Instance	n	m	C										
				1	2	3	4	5	6	7	8	9	10
eil51	51	250	GATS	8215	17551	29452	**33230**	**33350**	**40293**	**40203**	**44120**	**45228**	**43007**
			EA	**10397**	**19722**	**30181**	32102	30482	35749	34191	37373	37247	33697
			RLS	9173	18991	29998	31560	30497	35752	34230	37058	37209	33697
berlin52	52	255	GATS	8107	20247	32764	35018	41165	53501	56404	74836	63133	50526
			EA	**14528**	**28835**	**44516**	**48130**	**52655**	**63329**	65358	67873	**65713**	55935
			RLS	11398	24981	40838	48127	52559	63221	**65362**	**67882**	**65713**	**55952**
eil76	76	375	GATS	12387	28397	43974	**53637**	**63576**	**64705**	**72170**	**74670**	**73161**	**72539**
			EA	**14684**	**29705**	**41728**	45733	53648	52825	59137	60762	58755	57441
			RLS	14522	29765	41568	45454	53625	52853	59172	60658	58752	57435
kroA100	100	495	GATS	17871	**42303**	**62852**	**78784**	**97322**	**109370**	**122500**	**111540**	**103920**	**108070**
			EA	**20171**	37407	52409	60664	71889	80653	91242	77821	69296	73190
			RLS	19592	37101	52544	60497	71853	80656	91106	77849	69300	73190
pr124	124	615	GATS	24295	57282	90776	**113290**	**129060**	**142430**	**157270**	**143380**	**143720**	**132110**
			EA	52341	**79028**	**93081**	106169	115370	133307	118764	119662	107580	52341
			RLS	**52492**	78851	92919	106100	115370	133342	118773	119662	107549	52492
ch150	150	745	GATS	26247	61770	88715	**107560**	**121410**	**132560**	**132380**	**121960**	**129690**	**126520**
			EA	**34374**	**69088**	**91584**	99919	108119	114990	111872	98960	104834	100531
			RLS	31445	68036	91362	99873	108125	114974	111876	98962	104836	100525

VI. GATS RESULTS

The best obtained objective values for GATS were recorded and compared with the best obtained by two state-of-the-art approaches, EA and RLS. Tables 4–6 show the results for instances with one item per city and a knapsack capacity ranging from 1 to 10. Table 4 shows that GATS obtained higher objective values than EA and RLS for all knapsack capacities for the instances consisting of 51, 76 and 100 cities with uncorrelated (U) item weights. However, the results showed that GATS was not able to record higher objective values in instances consisting of 52 cities; it achieved better values in only two instances, with medium knapsack capacity

($C = 5\ and\ 6$). Nevertheless, the differences between the highest objective values obtained by the three approaches were not substantial in these instances; for example, EA and RLS achieved objective values, for the knapsack with $C = 4$, that were only 1% higher than GATS. GATS also outperformed RLS and EA in instances with a large number of cities (i.e., 124 and 150) and knapsacks with a high capacity (8, 9 and 10). With USW, Table 5 shows that GATS was able to achieve better objective values than EA and RLS, with the majority of instances having a small knapsack capacity (ranging from 1 to 3). Similarly, it outperformed them for most of the large-capacity instances, ranging from 8 to 10, except instances with 52 cities (i.e., berlin52). On the other hand, EA and RLS outperformed GATS with a medium KP, ranging from 4 to 7 in most instances, although the differences between the objective values obtained by the three approaches were not significant. For example, the best value reached by GATS with the instance consisting of 76 cities (i.e., eil76) and a knapsack capacity of 5 was 3556, whereas it was 3560 in RLS. In BSC, GATS surpassed RLS and EA in almost all instances, with medium- and large-capacity knapsacks ranging from 4 to 10, except in the instances with 52 cities (i.e., berlin52), where EA and RLS were able to achieve better objective values in the majority of the instances (see Table 6). GATS was also able to obtain better objective values with a small-capacity knapsack in various instances, particularly the ones for 51, 100, 124 and 150 cities, while the best objective values were achieved by EA in the rest of the instances for 52 and 76 cities. Tables 7–9 show the best objective values obtained from instances consisting of three items per city. Table 7 shows that GATS achieved better objective values in several uncorrelated instances with different knapsack capacities. For instance, it outperformed EA and RLS in all knapsack capacities in instances consisting of 51 cities (i.e., eil51). In contrast, lower objective values were achieved by GATS in all instances consisting of 52, 124 and 150 cities. Although the values obtained by GATS fluctuated among instances, it performed better with medium- and large-capacity knapsacks ranging from 5 to 10 instances containing 76 and 100 cities. The best objective values were achieved interchangeably with USW by the three approaches. However, GATS was able to surpass EA and RLS with large-capacity knapsacks, specifically when $C = 8\ ,9\ and\ 10$, in instances containing 51, 76, 100 and 124 cities. But EA and RLS achieved better values in the other two instances with the same knapsack capacities (see Table 8). Table 9 shows that GATS also obtained better objective values with BSC items, especially with large knapsack capacities ($C = 7, 8, 9\ and\ 10$).

As observed in Table 10, with uncorrelated items, GATS was unable to exceed the objective values obtained by EA and RLS when increasing the number of items to five per city. This became apparent with small and medium knapsacks ($C =$ 1 to 4). However, GATS revealed the highest objective values with large-capacity knapsacks in various instances. For instance, GATS outperformed RLS and EA in all instances consisting of large knapsacks, ranging from 6 to 10, for 51 and 100 cities (eil51 and kroA100). It also reached the highest objective value in most uncorrelated item instances with a knapsack capacity equal to 10, except for those that contained 52 cities (berlin52). But GATS was not able to outperform the two state-of-the-art approaches in all instances with USW, except instances with the largest knapsack capacity ($C = 10$). In fact, it recorded significantly higher objective values in four sets of various instances, particularly those composed of 51, 100, 124 and 150 cities (see Table 11). The performance of GATS for a BSC knapsack was notably better than that of the two state-of-the-art approaches in the majority of instances (see Table 12). However, this improvement became most evident with medium- and large-capacity knapsacks ranging from 4 to 10.

Based on the obtained results, the following findings were observed:

- The method used for tour initialization significantly affects the obtained solutions.
- Our proposed approach (GATS) performed better, in terms of objective values achieved, than two of the well-known state-of-the-art approaches (RLS and EA) in the majority of instances.
- GATS performed better, especially with instances of a large-capacity knapsack.
- GATS' performance significantly decreased when increasing the number of items and cities, particularly with a small-capacity knapsack.
- GATS had some issues with one set of instances, for berlin52, where it struggled to achieve better objective values in almost all instances.

VII. Conclusion

This paper investigates one of the recent NP-hard problems called the traveling thief problem, a multicomponent problem consisting of the two well-known problems TSP and KP. The optimization of TTP is challenging because of the interdependence between its components, where finding an optimal solution for one problem independently does not guarantee obtaining an optimal TTP solution. The aim of this paper was to investigate the use of hybrid GAs for the TTP. Therefore, we proposed a hybrid genetic approach with TS, a combination called GATS. The key aspect of GATS is that TTP solutions are considered by taking into account the interdependent nature of the TTP subcomponents. The performance of GATS was analyzed and compared with that of two state-of-the-art approaches, EA and RLS. A comprehensive set of TTP benchmark datasets was adopted in this experimental work, and 540 instances were selected for our investigation. These instances consisted of five different groups of cities, 51, 52, 76, 100 and 150 in number, as well as groups of items ranging in number from 50 to 745. The selected instances also consisted of the three types of KPs (U, USW and BSC) and all varieties of knapsack capacity (C). The obtained results were analyzed based on several factors, such as the type of knapsack, knapsack capacity and number of items per city.

The obtained results revealed that GATS was able to outperform EA and RLS in terms of objective values for several instances. This became more apparent with a large-capacity knapsack. However, some limitations of GATS were observed, for example, that its performance significantly decreased when the number of items and cities increased, particularly with a small-capacity knapsack. It was also noted that GATS had some issues with one set of the datasets, berlin52, where it struggled to achieve better objective values in almost all instances. Therefore, in the future, further experiments should be conducted to tackle such issues. Larger numbers of instances should also be investigated.

REFERENCES

[1] Bonyadi, M.R., Z. Michalewicz, and L. Barone. The travelling thief problem: The first step in the transition from theoretical problems to realistic problems. in 2013 IEEE Congress on Evolutionary Computation. 2013.

[2] Gupta, B.C. and V.P. Prakash. Greedy heuristics for the Travelling Thief Problem. in 2015 39th National Systems Conference (NSC). 2015.

[3] Martins, M.S.R., et al., HSEDA: a heuristic selection approach based on estimation of distribution algorithm for the travelling thief problem, in Proceedings of the Genetic and Evolutionary Computation Conference. 2017, ACM: Berlin, Germany. p. 361-368.

[4] Polyakovskiy, S., et al., A comprehensive benchmark set and heuristics for the traveling thief problem, in Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation. 2014, ACM: Vancouver, BC, Canada. p. 477-484.

[5] Chand, S. and M. Wagner, Fast Heuristics for the Multiple Traveling Thieves Problem, in Proceedings of the Genetic and Evolutionary Computation Conference 2016. 2016, ACM: Denver, Colorado, USA. p. 293-300.

[6] El Yafrani, M., et al., A hyperheuristic approach based on low-level heuristics for the travelling thief problem. Genetic Programming and Evolvable Machines, 2017.

[7] Mei, Y., X. Li, and X. Yao. Improving Efficiency of Heuristics for the Large Scale Traveling Thief Problem. 2014. Cham: Springer International Publishing.

[8] Bonyadi, M.R., et al., Socially inspired algorithms for the travelling thief problem, in Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation. 2014, ACM: Vancouver, BC, Canada. p. 421-428.

[9] Karder, J., et al. Solving the Traveling Thief Problem Using Orchestration in Optimization Networks. 2018. Cham: Springer International Publishing.

[10] Mei, Y., et al. Heuristic evolution with genetic programming for traveling thief problem. in 2015 IEEE Congress on Evolutionary Computation (CEC). 2015.

[11] Moeini, M., D. Schermer, and O. Wendt. A Hybrid Evolutionary Approach for Solving the Traveling Thief Problem. 2017. Cham: Springer International Publishing.

[12] Vieira, D.K.S., et al. A Genetic Algorithm for Multi-component Optimization Problems: The Case of the Travelling Thief Problem. 2017. Cham: Springer International Publishing.

[13] Wu, J., et al., Evolutionary Computation plus Dynamic Programming for the Bi-Objective Travelling Thief Problem. arXiv preprint arXiv:1802.02434, 2018.

[14] Lourenço, N., F.B. Pereira, and E. Costa. An Evolutionary Approach to the Full Optimization of the Traveling Thief Problem. 2016. Cham: Springer International Publishing.

[15] Mei, Y., X. Li, and X. Yao, On investigation of interdependence between sub-problems of the Travelling Thief Problem. Soft Computing, 2016. 20(1): p. 157-172.

[16] Alharbi, S.T., The Design and Development of a Modified Artificial Bee Colony Approach for the Traveling Thief Problem. International Journal of Applied Evolutionary Computation (IJAEC), 2018. 9(3): p. 32-47.

[17] Wagner, M. Stealing Items More Efficiently with Ants: A Swarm Intelligence Approach to the Travelling Thief Problem. 2016. Cham: Springer International Publishing.

[18] El Yafrani, M. and B. Ahiod, A local search based approach for solving the Travelling Thief Problem: The pros and cons. Applied Soft Computing, 2017. 52: p. 795-804.

[19] Araujo, R.P., et al., A novel List-Constrained Randomized VND approach in GPU for the Traveling Thief Problem. Electronic Notes in Discrete Mathematics, 2018. 66: p. 183-190.

[20] Faulkner, H., et al., Approximate Approaches to the Traveling Thief Problem, in Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation. 2015, ACM: Madrid, Spain. p. 385-392.

[21] Srinivas, M. and L.M. Patnaik, Genetic algorithms: a survey. Computer, 1994. 27(6): p. 17-26.

[22] Davis, L., Handbook of genetic algorithms. 1991.

[23] Glover, F. and M. Laguna, Tabu Search, in Handbook of Combinatorial Optimization: Volume1–3, D.-Z. Du and P.M. Pardalos, Editors. 1999, Springer US: Boston, MA. p. 2093-2229.

[24] Glover, F., Tabu Search—Part I. ORSA Journal on Computing, 1989. 1(3): p. 190-206.

[25] Glover, F., Tabu Search—Part II. ORSA Journal on Computing, 1990. 2(1): p. 4-32.

[26] Shapiro, J., Genetic Algorithms in Machine Learning, in Machine Learning and Its Applications: Advanced Lectures, G. Paliouras, V. Karkaletsis, and C.D. Spyropoulos, Editors. 2001, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 146-168.

[27] Barolli, A., et al. Application of Genetic Algorithms for QoS Routing in Mobile Ad Hoc Networks: A Survey. in 2010 International Conference on Broadband, Wireless Computing, Communication and Applications. 2010.

[28] Sivanandam, S.N. and S.N. Deepa, Genetic Algorithm Optimization Problems, in Introduction to Genetic Algorithms, S.N. Sivanandam and S.N. Deepa, Editors. 2008, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 165-209.

[29] Brusco, M.J. and P. Doreian, Partitioning signed networks using relocation heuristics, tabu search, and variable neighborhood search. Social Networks, 2019. 56: p. 70-80.

[30] Fauziah, N.F. and Y.H. Putra, Scheduling Regular Classrooms using Heuristic Genetic and Tabu Search Algorithms. IOP Conference Series: Materials Science and Engineering, 2018. 407(1): p. 012116.

[31] Shafahi, A., Z. Wang, and A. Haghani, SpeedRoute: Fast, efficient solutions for school bus routing problems. Transportation Research Part B: Methodological, 2018. 117: p. 473-493.

[32] Wagner, M., et al., A case study of algorithm selection for the traveling thief problem. Journal of Heuristics, 2017.

[33] Applegate, D., W. Cook, and A. Rohe, Chained Lin-Kernighan for Large Traveling Salesman Problems. INFORMS Journal on Computing, 2003. 15(1): p. 82-92.

[34] Simon, D., Evolutionary optimization algorithms. 2013: John Wiley & Sons.

[35] Holland, J. and D. Goldberg, Genetic algorithms in search, optimization and machine learning. Massachusetts: Addison-Wesley, 1989.

[36] Abdul–Razaq, T.S., Solving Composite Multi objective Single Machine Scheduling Problem Using Branch and Bound and Local Search Algorithms. Al-Mustansiriyah Journal of Science, 2017. 28(3): p. 200-208.

[37] Tamssaouet, K., S. Dauzère-Pérès, and C. Yugma, Metaheuristics for the Job-Shop Scheduling Problem with Machine Availability Constraints. Computers & Industrial Engineering, 2018.

[38] Cherukupalli, K., P.R. Chinda, and S. Peddakotla, Security Constrained Optimal Power Flow by Hybrid SATS Algorithm. Journal of Advanced Research in Dynamical and Control Systems, 2018(09-Special Issue).

[39] Wang, J. and Y. Wu, Optimal Design of Urban and Rural Public Transportation Network Based on Spatio - Temporal Constraints. Journal of Applied Science and Engineering, 2018. 21(1): p. 51-58.

[40] Lu, Y., et al., A Tabu search based clustering algorithm and its parallel implementation on Spark. Applied Soft Computing, 2018. 63: p. 97-109.

[41] Reinelt, G., TSPLIB—A traveling salesman problem library. ORSA journal on computing, 1991. 3(4): p. 376-384.

A Simple Approach for Representation of Gene Regulatory Networks (GRN)

Raza-ul-Haq, Javed Ferzund, Shahid Hussain
COMSATS University Islamabad
Sahiwal, Pakistan

***Abstract*—Gene expressions are controlled by a series of processes known as Gene Regulation, and their abstract mapping is represented by Gene Regulatory Network (GRN) which is a descriptive model of gene interactions. Reverse engineering GRNs can reveal the complexity of gene interactions whose comprehension can lead to several other details. RNA-seq data provides better measurement of gene expressions; however it is difficult to infer GRNs using it because of its discreteness. Multiple other methods have already been proposed to infer GRN using RNA-seq data, but these methodologies are difficult to grasp. In this paper, a simple model is presented to infer GRNs, using RNA-seq based coexpression map provided by GeneFriends database, and a graph-based database tool is used to create regulatory network. The obtained results show that it is convenient to use graph database tools to work with regulatory networks instead of developing a new model from scratch.**

Keywords—Graph theory; graph database; gene regulatory networks; RNA-seq; Genes Co-Expression; Neo4j

I. Introduction

The required information to make proteins and other molecules is stored in DNA. The functional circuitry of all living organisms is formed by genes [1]and synergistic actions between inter related genes is the reason of all biological reactions inside a cell. Gene regulation is a mechanism of increasing or decreasing production of gene products. In this process, genes are regulated by regulators to produce proteins or RNA, which produces a complex network of regulatory relationships. To understand the cellular process, it is critical to understand the regulatory relationships between genes. These relationships are expressed by gene regulatory network and are used to understand functions of genes. A gene regulatory network consists of nodes which represent genes, and edges between nodes represent relationships between genes. GRNs play important role in cell transduction, metabolism, cell differentiation, cell cycle and every other biological mechanism. In-depth and comprehensive understanding of complex biological processes can be provided by gene regulatory networks. The study of gene regulatory networks not only unveils the dynamics of organisms but also reveals their behavior in different scenarios and shows how their fate is controlled. The interaction of genes can be understood by reconstructing gene regulatory networks. The representation of genetic data, reverse engineering of GRNs and performing analytics on regulatory networks to retrieve information is challenging tasks. It can not only help to diagnose diseases but can also shed light on those changes which became reason for a disease and what changes have occurred because of that disease. Different people can be on different stages of a diseases, have different medical history which can lead to different response to the treatment from others; GRNs can be helpful in individualized treatment [2]. Determining or knowing drug sensitivity on target is an important aspect of the process of individualized treatment and is a research area in which GRNs can be used to detect drug sensitivity easily. Using GRNs we can answer the question that if a gene is mutated can its function be restored or not? The main genes responsible for a differentiation of cell into an organ or responsible for a disease can be identified using GRNs. In the process of regulating wide range of activities which include cellular, physiological and behavioral, circadian rhythm plays fundamental role. The researchers know a small number of genes which play a key role in circadian rhythm, however by using these genes, the existence of other key genes and how they work can be unveiled through GRNs. There are two main types of data which are used to infer GRN: microarray and RNA-seq. Continuous probe intensities are measured by microarrays, but discrete digital sequencing read counts which are aligned to sequence and are quantified by RNA-seq. Genes differential expression is measured more accurately with transcriptome sequencing (RNA-seq) than with microarrays. Different splice variants and non-coding RNA (ncRNA) can play important role in regulation of gene expression. The measurement of levels of transcripts provided by RNA-seq is far more precise than other methods [3]. In a single experiment RNA-seq can identify novel isoforms, novel transcripts allele specific expression, alternative splice sites and rare transcripts beyond gene expression analysis. It provides abilities to perform such types of experiments which traditional microarray-based methods cannot provide. Huge size of RNA-seq data increases challenges to interpret results, due to which processing mechanism and interpretation of results from RNA-seq experiments was often impeded. There was no database available for bioscience community which could provide RNA-seq data in a form through which any useful information could be extracted without going through time consuming processes of analyzing raw RNA-seq data before GeneFriends [4]. It allows researchers to identify those genes which are poorly annotated and associated with genes under study. Genes that are responsible for lung cancer is used in this research work to infer regulatory network. In this research work, GeneFriends database is used to obtain the genes co-expression network and an existing graph database tool is used to infer GRN. Deployment method shows that it is much simpler than other existing methods.

II. Methodology

A. Data Set Selection

Genes which are involved in commonly identified genomic/genetic alterations of lungs including chromosomal fusion rearrangements, nonsense or missense mutations, alternative splicing and small deletions or insertions are *EGFR, KRAS, MET, LKB1, BRAF, PIK3CA, ALK, RET*, and *ROS1* or in other words these genes are the most relevant to lung cancer [5]. These genes are selected in this research work to infer regulatory network.

B. Data Extraction and Transformation

Genomic databases are produced by a project named Ensembl for vertebrates. It is used to extract Ensembl transcript IDs of selected lung cancer genes Table I.

TABLE I. Lung Cancer Genes and Their Corresponding Ensemble ID

Gene	Ensemble ID
EGFR	ENSG00000146648
KRAS	ENSG00000133703
MET	ENSG00000105976
LKB1	ENSG00000118046
BRAF	ENSG00000157764
PIK3CA	ENSG00000121879
ALK	ENSG00000171094
RET	ENSG00000165731
ROS1	ENSG00000047936

These IDs are used in GeneFriends database to generate co-expression network of all the genes. Connection strength of direct partners within the seed list is 1, the genes having 0.75 connection strength are strongly co-expressed, 0.5 is connection strength of other direct partner and indirect partners have strength of 0.25. More than 1700 gene entries were generated as co-expression network in CSV format.

C. Tool

In our research scenario, graph databases are appropriate tools to infer network. These are database engines which model nodes and edges as first-class entities. Complex interactions between nodes can be represented in natural form. There are multiple graph databases available to work with and ArangoDB, Neo4j, Oracle Spatial and Graph, IBM System G Native Store and OrientDB are a few of them. Neo4j [6] graph database is used in this research work because it protects data integrity while providing fast reads and writes, and it is easy to learn.

D. GRN Construction

Cypher query language is used to query dataset. The network data available in CSV form was loaded into network. This data file contains genes and their connection strengths. The first gene was treated as source node, second one was treated as destination node and connection strength between two genes was used to create an edge between them, Figure 1. While creating nodes MERGE function was used instead of CREATE as values in data file are repeating multiple times and if a value is repeating, CREATE function always creates a new node having same label and property value which is not required in this research work. We could not afford repetition because multiple nodes with same properties will add extra work to get information about a single gene and it will also need more computing resources. On the other hand, only one node is created by MERGE against all the repeating values. But it does not happen in every case as if unbounded elements of an existing graph are being used to MERGE with a pattern it will start repeating nodes so always use bounded elements if an existing graph is being used. As in this research work there was no existing graph available, this issue was not considered. A value from data file was required to create a relationship, the default mechanism provided by neo4j could not be used as it does not provide this functionality. There is another function APOC.CREATE.RELATIONSHIP provided by apoc plugin, which was used for this purpose.

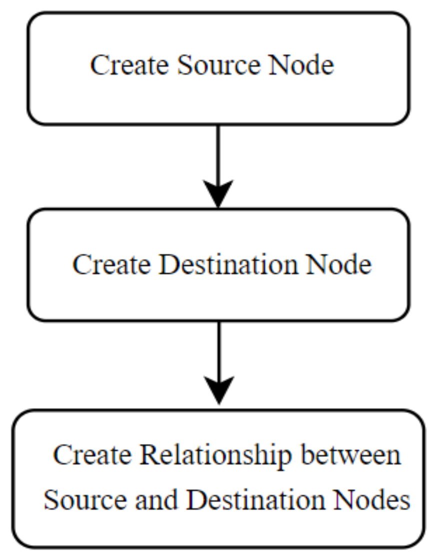


Fig. 1. Gene Regulatory Network Inference Model.

III. Results and Discussion

Using neo4j, a network of more than 1700 genes was created in which nodes indicate genes; relationships among nodes indicate connection strength between genes, and Ensemble IDs were assigned as property values of nodes. In this network, first value is treated as source node and is indicated by blue nodes, second value is treated as relationship type, third value is treated as destination node and is indicated by green nodes. The whole network is shown in Figure 2. There are 322 source nodes and 852 destination nodes having 1791 relationships, out of which three source nodes are connected to three destination nodes with connection strength 1. There are 65 source nodes connected to only 8 destination nodes with connection strength of 0.75. On the other hand, 9 source nodes are connected to 313 destination nodes with 0.5 connection strength and 0.25 is the connection strength of 1599 relationships among 1068 nodes out of which 322 are source nodes and 746 are destination nodes.

Fig. 2. Lung Cancer Regulatory Network Having 1791 Relationships, 322 Source Nodes and 852 Destination Nodes.

There is a node in the entire network which has the most relationships with other nodes among all. It means that, the gene named ALK associated with this node ID is the most important gene in the entire network.

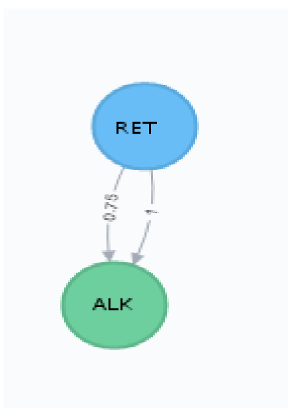


Fig. 3. *ALK* and *RET* Gene are connected with 0.75 and 1 Connection Strengths.

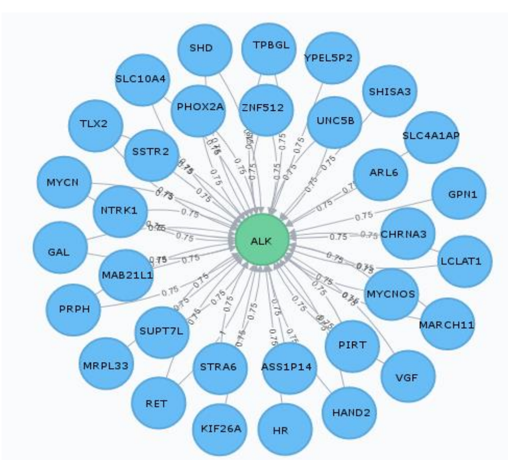


Fig. 4. *ALK* Gene is Connected to 32 Genes as Destination Node.

A gene named RET is a direct partner to ALK within seed list Figure 3. The production of protein involved in signaling within cells receives instructions from RET gene. Several kinds of nerve cells are developed by this protein. Mutation in this gene is the reason for Hirschsprung disease, pheochromocytoma, and most importantly lung cancer.

There are 32 genes associated with ALK with 0.75 connection strength, Figure 4. These 32 genes act as sources for ALK, which means it is co-expressed with these 32 genes. On the other hand, 39 genes are connected as destination nodes with ALK having 0.5 connection strength, Figure 5, which means it is direct partner with these 39 genes. It has 8 relationships of strength 0.25 as source node with 4 destination nodes, Figure 6.

Since it is present in every type of connection strength, it the most important node in the entire Network. There are strong evidences available to prove that it is the driving force of different types of cancers, including Non-Small Cell Lung Cancer and neuroblastoma [7] and inferred network in this research confirms this as well.

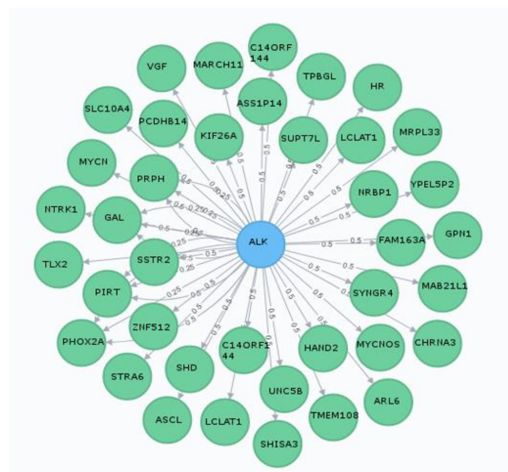


Fig. 5. *ALK* Gene is Source Node for 39 Genes Having 0.5 Connection Strength.

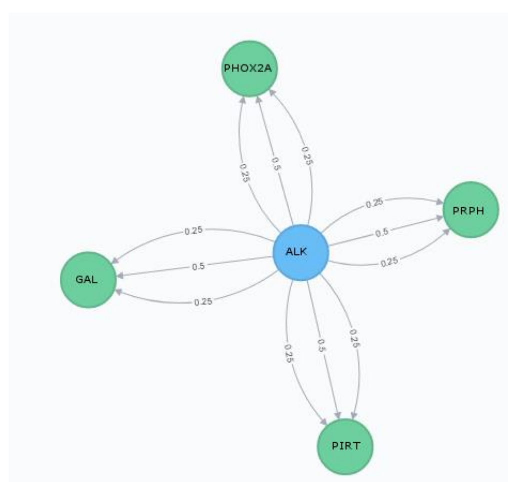


Fig. 6. *ALK* Has 12 Relationships of 0.25 Connection Strength with Four Other Genes.

TABLE II. GENES WHICH ARE COMMON IN BOTH 0.75 AND 0.5 CONNECTION STRENGTHS

Sr. No.	**0.75**	**0.5**
1	*TPBGL*	*TPBGL*
2	*HR*	*HR*
3	*MRPL33*	*MRPL33*
4	*YPEL5P2*	*YPEL5P2*
5	*GPN1*	*GPN1*
6	*MAB21L1*	*MAB21L1*
7	*CHRNA3*	*CHRNA3*
8	*ARL6*	*ARL6*
9	*SHISA3*	*SHISA3*
10	*LCLAT1*	*LCLAT1*
11	*STRA6*	*STRA6*
12	*PHOX2A*	*PHOX2A*
13	*PIRT*	*PIRT*
14	*TLX2*	*TLX2*
15	*SSTR2*	*SSTR2*
16	*GAL*	*GAL*
17	*ZNF512*	*ZNF512*
18	*SHD*	*SHD*
19	*UNC5B*	*UNC5B*
20	*HAND2*	*HAND2*
21	*MYCNOS*	*MYCNOS*
22	*NTRK1*	*NTRK1*
23	*MYCN*	*MYCN*
24	*SLC10A4*	*SLC10A4*
25	*VGF*	*VGF*
26	*PRPH*	*PRPH*
27	*KIF26A*	*KIF26A*
28	*ASS1P14*	*ASS1P14*
29	*SUPT7L*	*SUPT7L*
30	*MARCH11*	*MARCH11*
31	*LCLAT1*	*LCLAT1*

A list of 31 genes is given in Table II, these are common genes connected to ALK gene with both connection strengths 0.75 as well as 0.5. But when these genes have 0.75 strength they are being source genes of ALK and when they have 0.5 strength they are being destination genes of ALK. It means that if ALK is being directly co-expressed with these gene then it is also their direct partner at the same time. These genes are not only connected with mentioned genes but there are multiple other genes interacting with them at the same time in the entire network. We have discussed multiple techniques already available to infer regulatory network. There is one thing common in above mentioned methods and even most of the other methods proposed in last two decades to infer GRN, that is, they need a highly expert person in graph theory or mathematics in general to implement these methods and infer GRNs; however, the method and already existing tool presented in this research work does not need a researcher to be highly expert in such areas. Even though graph theory is working at backend of the tool, but researchers don't need to know how it is working and how it is inferring networks. So instead of spending time on comprehending those methods and then implementing them, researchers can easily infer GRN and spend their most of the time on analysis on inferred regulatory network which is the actual process of unveiling underlying mechanisms of genes interactions.

IV. RELATED WORK

For the integration of graph database with the analytical process of transcriptome data, a platform is presented in [8] through which data coming from Affymetrix platforms on rhesus, rat, mice and humans can be analyzed. An algorithm bLARS which is based on regression is used to construct GRNs using steady state gene expression data [4], which allows different genes to have different regulatory mechanisms. It uses bootstrapping for scoring purpose. Based on FA, PSO, BA-PSO which are swarm intelligence techniques, RNN formalism is used to investigate reverse engineering of GRNs from time series microarray datasets [1]. For refinement of classical network thresholding, a GRN post processing tool is represented in [9], linking nodes that belong to the same cluster with nodes that have higher weight, get favor by this method. It uses random walker to compute an optimal gene and select an optimal edge jointly. GRN inference is improved when clustering process is introduced in edge selection process. A novel technique for discovery of gene regulatory network is proposed [10] in which discovery process is integrated into heuristic information. To construct large scale gene regulatory networks a dynamic multi-agent genetic algorithm is represented [11], which is based on FCM. The method proposed in [12] uses low rank property to construct a common GRN structure from other inferred GRNs, drug effect is also inferred and estimated by this method. Through anti-diabetic drug, Metformin, the benefits to target tumor cell metabolism are investigated using simulations [13]. The use of S-System modelling formulation is proposed in [14], by combining standard system identification procedures with this modelling formalism, the type of regulation between each gene is established and then a model which is suitable for designing a synthetic genetic feedback controller is derived. To predict transcription factors and gene interactions, a method which uses iterative SVM and clustering is implemented in [15]. To discover relationships between genes this method [16] combines PCA-CMI and GA algorithms. For a target gene to obtain the best predictor, GA was performed and PCA-CMI method was used to create initial population to reduce search space. For encoding dynamics of multi-valued network, use of an extension of ASP named FASP as the language in continuous domain is proposed [17]. A methodology [18] in which observed count data is modelled as being negative binomially distributed is proposed to infer gene regulatory networks using RNA-seq time series data. To identify multivariate gene interaction in RNA-seq data, an application of BEE and OBC is demonstrated [19] to differentiate biological phenotypes. To infer gene regulatory network Legendre neural network (LNN) is proposed [20] and to optimize the parameters of Legendre neural network, Firefly algorithm is used.

V. Conclusion

Complex biological mechanisms are widely elucidated by gene expression information, which is expressed by interacting along with one or multiple other genes and this interaction with other genes formulates a regulatory network. Reverse engineering of this regulatory network is an important task to get the insight of biological mechanisms. To study cell trascriptome at system level, RNA sequencing is a revolutionary technique. In this research work, a graph database neo4j is used as a tool to construct GRN and RNA-seq data of lung cancer genes and is used as dataset provided by GeneFriends database. Many techniques are available to infer a GRN but most of them implement complex mathematical models during the process. In this research work, we have used an already existing tool, even though graph theory is working behind the scene to infer network in this tool as well, but researchers don't have to pay any attention on background process instead they can focus on network analysis part so that, the complex underlying mechanism can be understood.

References

[1] Abhinandan Khan, et al. "A swarm intelligence based scheme for reduction of false positives in inferred gene regulatory networks," IEEE Congress on Evolutionary Computation (CEC), pp. 40-47, Jul 2016.

[2] Baiyi, An, and Shen Wei. "A Novel Gene Regulatory Network Construction Method Based on Singular Value Decomposition." 2016 IEEE International Conference on Big Data Analysis (ICBDA), 2016, doi:10.1109/icbda.2016.7509844

[3] Lehrach, Hans. "Faculty of 1000 Evaluation for RNA-Seq: a Revolutionary Tool for Transcriptomics." F1000 - Post-Publication Peer Review of the Biomedical Literature, Sept. 2013, doi:10.3410/f.1128830.793481779.

[4] Nitin Singh, et al. "bLARS: An Algorithm to Infer Gene Regulatory Networks," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, pp. 301-314, Jun 2015.

[5] El-Telbany, A., and P. C. Ma. "Cancer Genes in Lung Cancer: Racial Disparities: Are There Any?" Genes & Cancer, vol. 3, no. 7-8, Jan. 2012, pp. 467–480., doi:10.1177/1947601912465177.

[6] López, Félix Melchor Santos, and Eulogio Guillermo Santos De La Cruz. "Literature review about Neo4j graph database as a feasible alternative for replacing RDBMS." Industrial Data, vol. 18, no. 2, 2015, p. 135., doi:10.15381/idata.v18i2.12106.

[7] Ardini, E., et al. "Anaplastic Lymphoma Kinase: Role in specific tumours, and development of small molecule inhibitors for cancer therapy." Cancer Letters, vol. 299, no. 2, 2010, pp. 81–94., doi:10.1016/j.canlet.2010.09.001.

[8] Costa, Raquel L., et al. "GeNNet: An Integrated Platform for Unifying Scientific Workflow Management and Graph Databases for Transcriptome Data Analysis." Dec 2016, doi:10.1101/095257.

[9] Aurelie Pirayre, et al, "BRANE Clust: Cluster-Assisted Gene Regulatory Network Inference Refinement," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1-1, Mar 2017.

[10] Armita Zarnegar, et al, "A heuristic gene regulatory networks model for cardiac function and pathology," Computing in Cardiology Conference (CinC), pp. 353-355, Sept 2016.

[11] ing Liu, et al, "A Dynamic Multiagent Genetic Algorithm for Gene Regulatory Network Reconstruction Based on Fuzzy Cognitive Maps," IEEE Transactions on Fuzzy Systems, vol.24, pp. 419-431, Jul 2015.

[12] Young Hwan Chang, et al. "Retrieving common dynamics of gene regulatory networks under various perturbations," 54th IEEE Conference on Decision and Control (CDC), pp. 2531-2536, Dec 2015.

[13] Osama A. Arshad, et al. "Using Boolean Logic Modeling of Gene Regulatory Networks to Exploit the Links Between Cancer and Metabolism for Therapeutic Purposes," IEEE Journal of Biomedical and Health Informatics, vol.20 pp. 399-407, Nov. 2014.

[14] Mathias Foo, et al. "Modelling and Control of Gene Regulatory Networks for Perturbation Mitigation," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1-1, Jan 2018.

[15] Jisha Augustine, et al. "Gene regulatory network inference: A semi-supervised approach," Electronics, Communication and Aerospace Technology (ICECA), Dec. 2017.

[16] Sima Iranmanesh, et al. "Inferring gene regulatory network using path consistency algorithm based on conditional mutual information and genetic algorithm," Computer and Knowledge Engineering (ICCKE), Dec. 2017.

[17] Mushthofa Mushthofa, et al. "Computing attractors of multi-valued Gene Regulatory Networks using Fuzzy Answer Set Programming," IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Nov. 2016.

[18] Thomas Thorne, "Approximate inference of gene regulatory network models from RNA-Seq time series data," Cold Spring Harbor Laboratory, bioRxiv journal, Jan. 2017.

[19] Jason Knight, et al. "Detecting Multivariate Gene Interactions in RNA-Seq Data Using Optimal Bayesian Classification," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1-1, Oct. 2015.

[20] Bin Yang, et al. "Inference of Gene Regulatory Network based on Legendre Neural Network," International Conference on Information Technology in Medicine and Education, pp. 192-194, July 2017

A Methodology for Identification of the Guilt Agent based on IP Binding with MAC using Bivariate Gaussian Model

B. Raja Koti[1]
Research Scholar
Department of Information Technology,
Gitam Institute of Technology,
GITAM, Visakhapatnam, Andhra Pradesh, India

Dr. G. V. S. Raj Kumar[2]
Associate Professor
Department of Information Technology,
Gitam Institute of Technology,
GITAM, Visakhapatnam, Andhra Pradesh, India

Dr. Y. Srinivas[3]
Professor
Department of Information Technology,
Gitam Institute of Technology,
GITAM, Visakhapatnam, Andhra Pradesh, India

Dr. K. Naveen Kumar[4]
Assistant Professor
Department of Information Technology,
Gitam Institute of Technology,
GITAM, Visakhapatnam, Andhra Pradesh, India

***Abstract*—Enormous increase in data in the current world presents a major threat to the organization. Most of the organization maintains some sort of data that is sensitive and must be protected against the loss and leakage. In the IT field, the large amount of data will be exchanged between the multiple points at every moment. During this allocation of the data from the organization to the third party, there are enormous probabilities of data loss, leakages or alteration. Mostly an email is being utilized for correspondence in the working environment and from web-based like logins to ledgers; thereby an email is turning into a standard business application. An email can be abused to leave organization's elusive information open to trade off. Along these lines, it might be of little surprise that muggings on messages are normal and these issues need to be addressed. This paper completely focuses on the concept of data leakage, the technique to detect the data leakage and prevention.**

Keywords—Data leakage; sensitive information; data leakage detection; bivariate normal distribution; probability density function

I. INTRODUCTION

In the current digital era, the usage of internet has increased rapidly, and every office and organizations were connected to the internet to simplify the works like saving files and sharing information from one to another in order to execute works very fast. In this process, of exchange of data due to the large storage systems, huge data collection is accumulated which is being stored in the server for getting access to a user and reuse the data. But in the large organizations, the data will be handled by the third party members, and sometimes it leads towards certain issues pertaining to data modification, either by technical men within the organization or unknowingly by the users and at times intentionally done by the few. So, in the current scenario, if the data is shared by an unauthorized user, it may create threat/ problem to the organization especially if it is a sensitive data and thereby incurs a huge loss to the organization. To overcome this type of issues, we need to identify the cause of sharing information, whether it was done purposefully or unintentionally, and also we need to provide security to data. If the data is already shared, then we have to protect the data which is already leaked. Protection of data is the very essential thing for any enterprise related to digital data, several methods are adapted to protect the data, based on the priority of information on the demand.

Among the recent challenges, network information expulsion is a vital concern for the business organizations. Unauthorized communication could have serious consequences for a corporation. To discontinue from the unwanted dealings of an organization, it is required to regulate the information flow within and outdoors of the organization and this process helps to secure the data. Recent news and reports indicate fifty percent of organizational data's are leaked within the sector either part or absolutely [1]. This can be terribly troublesome to spot the precise details of leaked information and therefore the informant. However, the information discharge has several channels to leak. Thus observing each channel is a challenging possible task, and also creates several serious problems. However, the principles will be profaned from completely different accessible channels like email, instant electronic communication, and via alternative social media attachments.

Threats to an e-Mail are that employees put organization's sensitive information and resources in danger in spite of arrangements that characterize adjust methodology. The accompanying illustrations indicate how representatives deliberately and unexpectedly spill delicate information. The some of the points referred to are (1) Usage of Unauthorized applications: In organizations, utilization of individual messages can put delicate information and individual data in danger. As indicated by a study report, 63 percent of workers concede that they utilize work PC for individual utilization. These applications don't take after corporate security gauges.

Thus, information spillage by a representative is high [14]. (2) Misuse of Corporate Computers: Employees purposefully utilize organization's PCs in numerous ways that undermine IT security arrangements which incorporates sharing of work gadgets and delicate data with non-workers. These practices can bring about spilling out the IP of an organization, which presents genuine dangers to an organization's security and productivity. (3) Misuse of Passwords and Login/Logout Procedures: When a worker leaves a framework logged in and with a secret phrase joined to it that welcomes an attacker to take the delicate information at their relaxation. On the off chance that worker utilized that PC for individual utilize which implies data is presently eagerly accessible to the attacker [15]. In this article, a model is presented to identify the appropriateness of the user and an error is generated as an unauthorized user if he/she is not authenticated.

The rest of the paper is organized as follows; section 2 of the paper deals with the current scenarios, Data Leakages and Guilt Agents identification. Section 3 of the paper highlights about bivariate normal distribution, KL- Divergence and the proposed algorithm deals with the methodology developed in section 3. Results derived are presented in section 4. In section 5 the conclusions are summarized.

II. Related Work

Data Leakage is one of the necessary assets to several organizations having information, and for that matter, the protection of this information should take the primary priority [11]. Although several organizations have placed unavoidable security mechanisms and technical systems significantly firewalls, virtual personal networks (VPNs), and intrusion detection systems/intrusion bar systems (IDSs/IPSs; still information expulsion will occur [2]. Say again that if the information leak happens and once sensitive data is unconcealed to unauthorized users or parties either purposely or not. The information leakage will cause serious consequences or several threats to an organization. Let’s say, the loss of the confidential or sensitive information will have an unembellished or confrontational impact on a company’s name and quality, customers, worker confidence, competitive advantage, and in some cases, this will result in the closure of the organization. In additionally, information leak is a key concern for the business organizations during this progressively networked world today and for that matter, any unauthorized speech act of sensitive or confidential knowledge might have serious consequences for a corporation in each long and short-term [3].

In addition, consistent with the problem of information, a leak could be a growing concern among organizations and agents. Alneyadi et al [4] indicated that a lot of leakages occurred within the business sectors that they were within the government sector. Consistent with a report in 2014, the statistics stand at 500th within the sector and 200th within the government sector. They have declared that though in some cases the information leaks weren't harmful to organizations, however, others have caused many lots of dollars’ value harm. More so, the quality of many businesses or organizations area unit comprised one such sensitive information as trade secrets, project documents, and client profiles area unit leaked to their competitors. Take it more than government sensitive information values political selections, enforcement, and national security data may also be leaked. A typical example of presidency sensitive information that was leaked was the United States diplomatic cables by Wiki Leaks. “The leak consisted of concerning 250,000 United States diplomatic cables and 400,000 military reports observed as ‘war logs’. This revelation was administered by an enclosed entity exploitation an external disk drive and regarding 100,000 diplomatic cables were labeled confidential and 15,000 cables were classified as secret”, this incident received high public criticisms from among civil rights organizations everywhere the global. In another development, hackers scarf 160 million credit and open-end credit numbers that targeted 800,000 bank accounts in the United States, that were thought of joined of the most important hacking incident that has occurred [5].

The need to deal with such serious problems culminated with the implementation of bound security management mechanisms similar to firewalls, VPNs, IDS, and IPSs by many organizations [6]. Consistent with these systems work satisfactorily once the information is well outlined, structured and constant. Alneyadi et al [7] Any such explicit attack; that once information is either changed, tag otherwise or compressed, these systems become naive and confidential information will still be leaked. As an example, a firewall will have rules to dam access to confidential information; however, constant information may be accessed through many means that compared to an email attachment and instant electronic messaging (IM). This suggests that the normal security mechanisms (firewall, VPNs, IDSs / IPSs) is incapacitated and lack the understanding of information linguistics [8]. To beat this deficiency in protecting sensitive information, a brand new paradigm shift known as information leak interference systems are introduced. Security and privacy problems have enhanced by the rate, volume, and style of bachelor's degree, comparable to large-scale cloud infrastructures, diversity of information sources and formats, streaming nature of information acquisition, and large amount information inter-cloud migration [9]. Maybe sensitive or non-sensitive, and irrespective of a leak of information may be expensive for any businesses or users. Parenthetically, a client MasterCard record that is leaked may be expensive to reach the bank and therefore the client. [10] Typically, information leak happens because of data sharing with users internally or outwardly to the organization, exchanging emails that contain sensitive data, publically cathartic information on the web or cloud, data that is taken with illicit motives or inadvertently. The guilt can be identified by using the MCA-IP address [12] bound to the particular log file and the timestamps allotted by the records and the leaked data is encrypted [13]. Information sensitivity varies comparable to banking data, MasterCard data, criminal and justice information, financial data, health records, etc., to feature to the current; the appearance of sensitive data has caused various information security challenges that need completely different mechanisms in managing things. Also, because of the voluminous of information that is generated and used recently by organizations, there should be subtle technologies and methodologies that may handle the voluminous of information firmly and with efficiency and to stop data leak. Finally, many DLP ways are designed, however, there is very little done stop information leak in Sensitive data exploitation; the preventive approach which might facilitate organizations prevent the leak before they happen.

III. Methodology

Data leakages are considered to be one of the most targeting problems for each and every organization. With the latest technology evolution, most of the industries are migrating towards the development of a framework that suits their organization. This professional difference towards the latest technology changes has given to fold both job seeker and job provider. The number of technical jobs in this area research has been increased enormously and thereby creating a lot of ventures to the software industry and also given a scope for the budding entrepreneurs to start their own entrepreneurs by mean of establishing start-up companies. On the other hand, due to professional opportunities available, a lot of challenges also arose because of the security constraints. The security issues may be either with respect to the attacks on the organization with the only purpose of hacking the valid information is to surplus the information available at an organization with the objective of professional rivalry. These attacks are turned as outside attacks were apart from hacking and attacking the information other means of attacks such as spreading malicious virus, worms etc., with the purpose of destroying the meaningful information other types of attacks are also victims in the current day challenges and these types of attacks are called as the internal attacks were the objective is to leak the content from the source organization working with and supply the information to the third party for financial and other sort of benefits. Many articles are being proposed and developed in the literature to identify the guilt agent. However, this issue remains to be a still in challenge state and leads directions to make the work progress. Hence we propose a Bivariate Gaussian distribution model and KL divergence which helps to identify the leakage more appropriate.

A. *Bivariate Normal Density*

The bivariate normal distribution can be defined as the probability density functions (pdf) of two variables X and Y that is linear functions of the same independent normal random variables of the function is

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}Q(x,y)\right) \tag{1}$$

with the quadratic form

$$Q(x,y)=\frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\frac{(x-\mu_x)(x-\mu_y)}{\sigma_x\sigma_y}\right] \tag{2}$$

Given the joint density function of a bivariate normal distribution, with the parameters σ_x^2,σ_y^2 and ρ , the 2-by-2 symmetric matrix is given by

$$\sum=\begin{bmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{bmatrix},\quad x=\begin{bmatrix}x\\ y\end{bmatrix}\quad \text{and}\quad \mu=\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix} \tag{3}$$

The quadratic form can be expressed as

$$Q(x,y)=(x-\mu)^T\sum(x-\mu)=\begin{bmatrix}x-\mu_x & y-\mu_y\end{bmatrix}\begin{bmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_x\\ y-\mu_y\end{bmatrix} \tag{4}$$

Suppose that two random variables X and Y has the bivariate normal distribution (1). The following properties are to find out the bivariate normal distribution.

1) Their marginal distributions

$f_X(x)$ and $f_Y(x)$ become

$$f_X(x)=\frac{1}{\sqrt{2\pi\sigma_x}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right\} \tag{5}$$

and

$$f_Y(y)=\frac{1}{\sqrt{2\pi\sigma_y}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\} \tag{6}$$

Thus, X and Y are normally distributed with respective parameters (μ_x,σ_x^2) and (μ_y,σ_y^2) .

2) If $\rho=0$, then the quadratic form (2) becomes and consequently, we have $f(x,y)=f_X(x)f_Y(y)$. Thus, X and Y are independent when the $\rho=0$.

$$Q(x,y)=\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{x-\mu_y}{\sigma_y}\right)^2 \tag{7}$$

3) If X and Y are not independent (that is, $\rho \neq 0$), we can compute the conditional density function $f_{Y|X}(y|x)$ given $X=x$ as

$$f_{Y|X}(y|x)=\frac{1}{\sqrt{2\pi\sigma_y}\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu_y-\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)}{\sigma_y\sqrt{1-\rho^2}}\right)^2\right\} \tag{8}$$

Which is the normal density function with a parameter $(\mu_y+\rho\frac{\sigma_y}{\sigma_x}(y-\mu_x),\ \sigma_y^2(1-\rho^2))$ similarly, the conditional density function $f_{X|Y}(y|x)$ is given $Y=y$ becomes this is the normal density function with parameter $(\mu_x+\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y),\sigma_x^2(1-\rho^2))$

$$f_{X|Y}(x|y)=\frac{1}{\sqrt{2\pi\sigma_x}\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_x-\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)}{\sigma_x\sqrt{1-\rho^2}}\right)^2\right\} \tag{9}$$

B. *Covariance*

Let (X,Y) be a bivariate normal random variables with a parameters $(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2\rho)$. The covariance is given by

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi\sigma_x}\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu_x-\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)}{\sigma_x\sqrt{1-\rho^2}} \right)^2 \right\} \sigma_y^2(1-\rho^2) \quad (10)$$

The correlation coefficient of X and Y is defined by

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \quad (11)$$

It implies that the parameter ρ of the bivariate normal distribution represents the correlation coefficient of X and Y.

C. KL-Divergence

Kullback - Leibler Divergence (KL Divergence) is considered to identify the divergence between two probability density functions and the formula for computing the same is given by

$$= \int_{-\infty}^{\infty} (x-\mu_X)^2 f(x)dx = \rho\frac{1}{2} \quad (12)$$

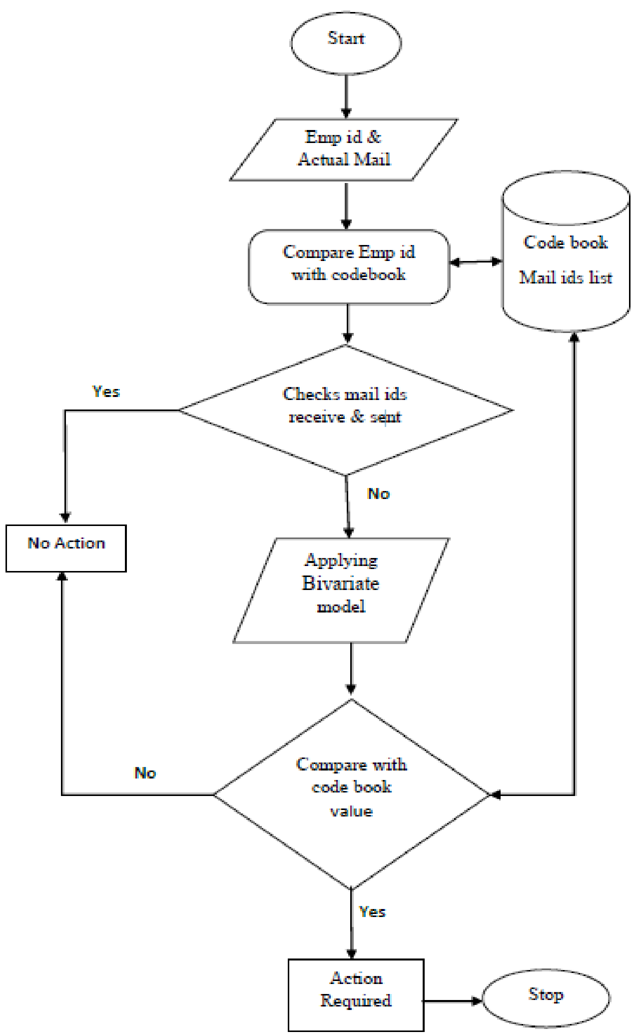


Fig. 1. Data Flow Diagram.

D. General Architecture

In this paper, we have considered a Bivariate Gaussian model, with the assumption that as it considers two attributes into consideration, the identification of the guilt agent becomes easier. In this method, we have considered the unique id of an employee within the organization, together with the MAC address of the system being used and the corresponding probability density function are mapped which is shown in fig 1. Every data that is transmitted from the source are identified as the super user or administrator and against each of the probability density function, the corresponding MAC id and system IP are notified. During the transmission process whenever a data is transferred from source to destination, three parameters are considered and checked against the e-mail ids. Every employee performing for an organization is entitled to transmit data are receiving data from a set of valid users. Every e-mail i.e. received or transmitted is cross-checked against the validity of the e-mails to which they are sent or receive in order to identify the guilt agent.

E. Proposed Algorithm

Algorithm: Checking Codebook Algorithm
Input: T1, T2 (codebook), Dept, and Mail id
//T2**:** Original codebook data
//Dept**:** department of an organization
Output: L, display miss match data of T1 and T2

1. **begin**
2. L = 0;
3. **If** match (Mail. To, organization) AND match (Mail. Dept) **then**
4. Return L;
5. **end if**
6. **If** match ((Mail. To, organization) AND not match (Mail, Dept)) OR not match (Mail. To, organization emp id) **then**
7. **If** match (receive, mail) AND match (send, mail) then
8. Return L;
9. **end if**
10. **end if**
11. **If** not match (receive, mail) not AND match (send, mail) then
12. **for all** t1 ∍ Mail **do**
13. **for all** t2 ∍ T2**do**
14. **If** match (t1, t2) **then**
15. L = L compare (t1 ↔ t2);
16. **end if**
17. **end for**
18. **If** L ≠ 0 **then**
19. Apply Action ();
20. **end if**
21. Return L;
22. **end for**
23. **end if**
24. **end**

IV. Experimental Results

To evaluate the performance of the proposed model that considers the sensitive data leakage problem. We implemented the proposed framework by using Python environment. In our experimental the developed model is presented in two phases, in the first phase we have considered a model that is developed by considering the MAC and along with employ id and output derived are given as the input to the model based on Bivariate Gaussian mixer model in the second phase instead of MAC address we assumed that IP is static and linked to the employ id to get a unique value based on the consider model. The two scenarios have experimented with different typed of employ ids and their MAC addresses, IP addresses.

For this implementation we use the data set which has unique employ ids, for each employ id we have created a codebook which contains emails ids list for that particular employee has to receive and sent, this experiment was done with 150 emails ids which are unique, from those values we generated mean, variance, standard deviation, correlation and to calculate Bivariate values all those sample values are given in below from table I to table III.

TABLE I. Mean, Variance and Standard Deviation for the dataset

	EMP ID	MAC	IP
Count	150	150	150
Mean	75.5	649.2267	849.88
Variance	1887.5	1578.499	1483.999
Standard Deviation	43.44537	39.73032	38.52271

TABLE II. Correlation and Covariance for EMP, MAC and IP

	EMP, MAC	EMP, IP	MAC, IP
Correlation	0.565266	0.571449	0.3615
Covariance	975.7047	956.396	553.2824

TABLE III. Bivariate Gaussian model values for EMP and MAC, EMP and IP

EMP IDs	Bivariate Gaussian model for EMP and MAC	Bivariate Gaussian model for EMP and IP
5	0.000299749	0.000501601
10	0.000511379	0.000194190
19	0.000628811	0.000644221
27	0.000729978	0.000177475

From the fig 2 Bivariate Gaussian model values which are one employ id, for that, we generated values for the taken data set to substitute in the equation 1. Here we consider x and y values are for the first phase employ id is x and MAC is y, in the same way in the second phase we consider employ id is x and IP is y. After getting two phases pdf values are stored in the codebook which is on the server side to compare with the test data.

We performed our experiments in 2 cases; based on IP and EMP id, MAC and EMP id categories. The results in Table IV show that the accuracy of MAC and EMP id category is the best case when compared to the other.

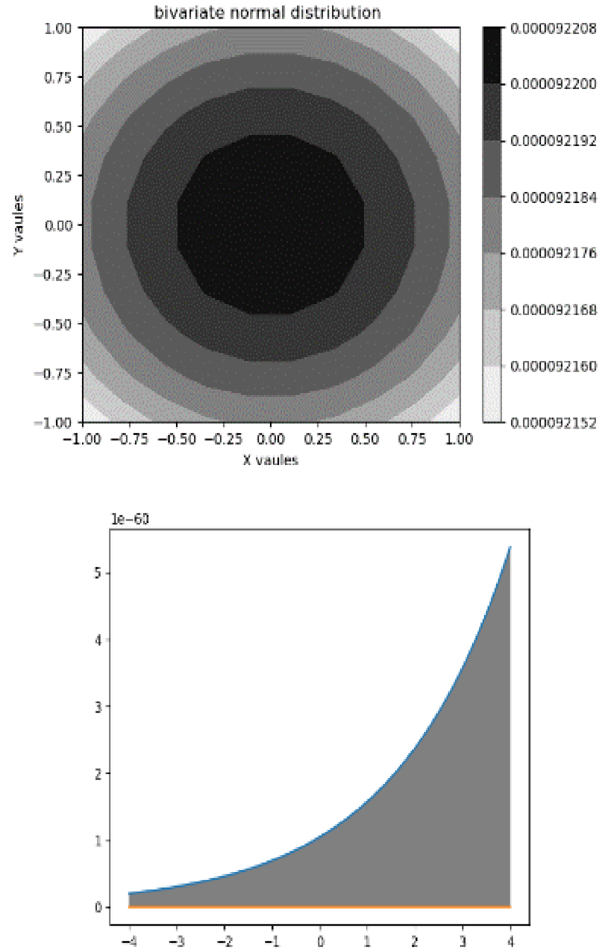


Fig. 2. Bivariate Gaussian Model Values.

In order to evaluate the procedure, we have considered the evaluation metrics based on Precision and Recall and based on these values the accuracy measure is evaluated, and the formula for estimating the accuracy is given by

$$\text{Accuracy} = (TP + TN)/N \quad (13)$$

$$\text{Precision} = TP/(TP+FP) \quad (14)$$

$$\text{Recall} = TP/(TP+FN) \quad (15)$$

TABLE IV. Precision , Recall and accuracy

	MAC and EMP ID	IP and EMP ID
TP	120	110
TN	245	255
FP	30	40
FN	10	15
Precision	0.80	0.73
Recall	0.92	0.88
Accuracy	0.90	0.87

where N = TP + FP + TN + FN, and where TP is True positive, FP False positive, TN True negative, FN False negative.

In order to identify the output, KL Divergence is considered, KL Divergence is used to find relevance in between two generated probability density function values. Here the KL Divergence is to compare and relevance of the training set and test data, if the divergent between the test and train is more than this, says there is a diversion in the data and they may be a possible scope for data leakage.

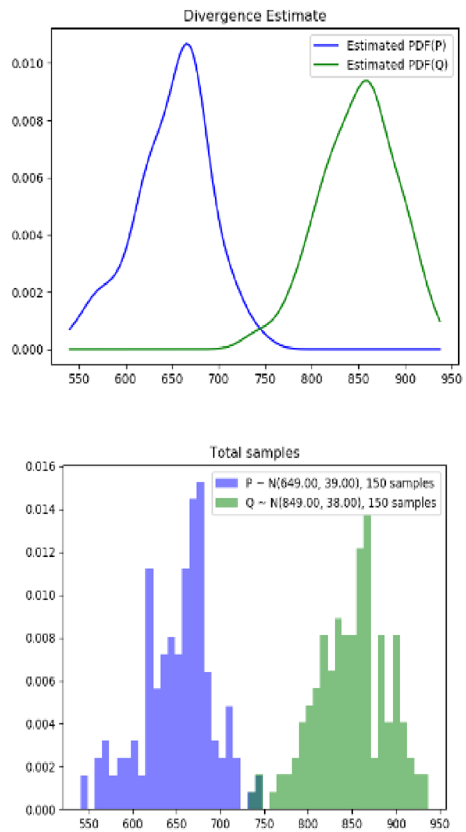
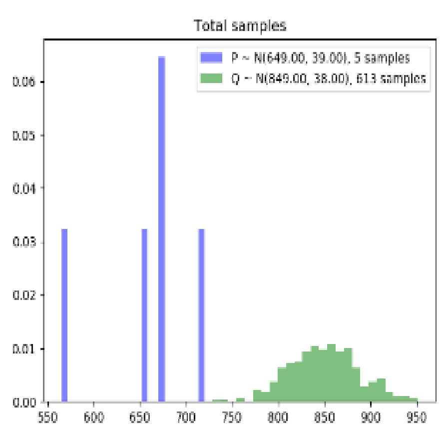


Fig. 3. Total Samples that we are taken for P, Q and Estimated PDF's for P and Q.



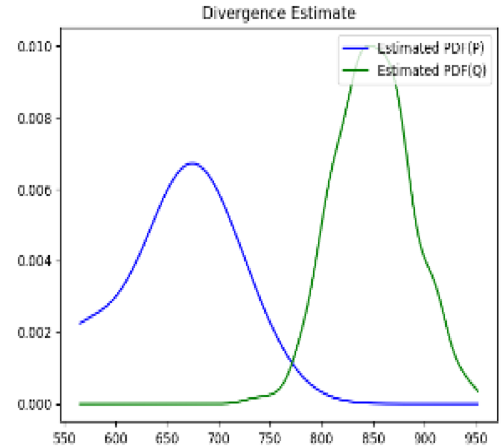


Fig. 4. PSamples that we are taken for test with mean 649.00 and Estimated PDF's for P and Q.

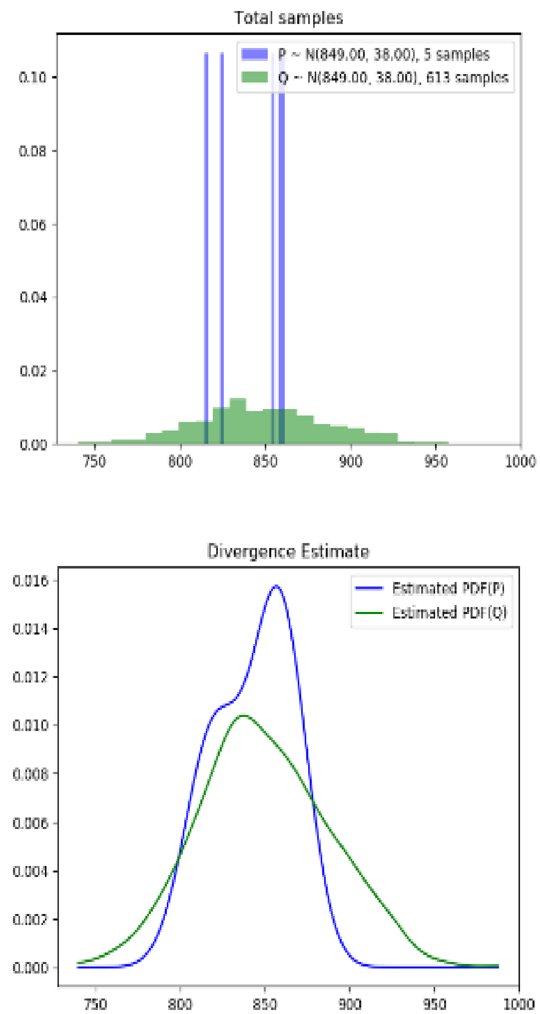


Fig. 5. PSamples that we are taken for test with mean 849.00 and Estimated PDF's for P and Q.

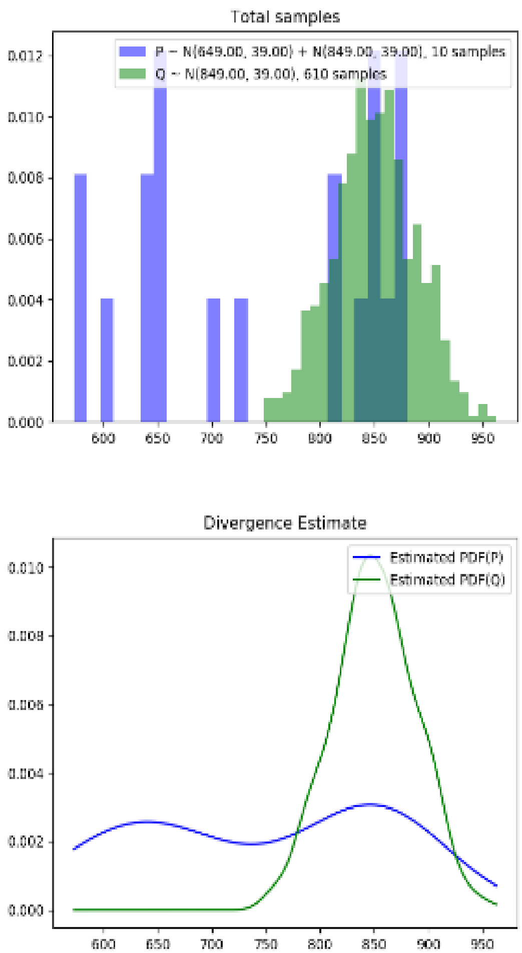


Fig. 6. KL Divergence graphs for the P, Q and Estimated PDF's for P and Q.

The above fig -3 to fig - 6 the pdf values of the two data sets, viz trained data pdf values that stored in server side and the original values of the acquired are considered for each employ id separately and then another figure denotes the test data pdf values which we have collected after the employ communicated with mail id. Then the KL Divergence is used to find relevance in between two generated probability density functions values and find out odd one out in the organization.

V. Conclusion

Data leakage is a current drawback within the field of IT mostly in the data security. There are various researches from numerous domains who are endlessly working towards developing and safeguarding the sensitive information from the data leak detection and finding ways to reduce this drawback. Protective confidential and sensitive data is a lot of vital importance. The aim was to develop the model with a known the scope of leakage and apply a technique to detect the data leakage and to identify the guilty agent(s) in a particular organization. We can't predict an organization that does not utilize an email for correspondence and similarly it is hard to consider an organization that would not get any advantage from an email security. So, we proposed a Bivariate Gaussian distribution model that helps to identify the guilty agent and leakage more appropriately. Thus, it provides the necessary security to the sensitive information and finding the guilt agent also, which is very helpful in various areas that hold sensitive data, especially where data is shared through emails. The proposed system can disconnect inbound or outbound emails. Future endeavors can be made in actualizing this technique that can handle the present reality situation thereby helping to identify the guilt agents and also protect the leaked data with more accuracy.

References

[1] Data loss db. Data loss statistics. Retrieved from ⟨http://datalossdb.org/⟩; 2015.

[2] Tahboub, R & Saleh, Y. (2014), Data Leakage / Loss Prevention Systems (DLP), NNGT Journal: International Journal of Information Systems, vol. 1, pp. 13-18.

[3] Soumya, S. R. &Smitha, E. S. (2014), Data Leakage Prevention System by Context based Keyword Matching and Encrypted Data Detection, International Journal of Advanced Research in Computer Science Engineering and Information Technology, vol. 3, issue 1, pp. 375-384.

[4] Alneyadi, S., Sithirasenan, E. &Muthukkumarasamy, V. (2016), A survey on data leakage prevention systems, Journal of Network and Computer Applications, vol. 62, issue C, pp. 137-152.

[5] Vadsola, R., Desai, D., Brahmbhatt, M. &Patanwadia, A. (2014), Data Leakage Prevention by Using Word Gram Based Classification and Clustering, International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, issue 9, pp. 8040-8041.

[6] Kale, A. V., Bajpayee, V. & Dubey, S. P. (2015), Analysis of Data Leakage Prevention Solutions, International Journal for Engineering Applications and Technology (IJFEAT), vol. 1, issue, 12, pp. 54-57.

[7] Alneyadi, S., Sithirasenan, E. and Muthukkumarasamy, V. (2015), Detecting Data Semantic: A Data Leakage Prevention Approach, In the Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA, August 20 - 22, IEEE Computer Society Washington DC, USA, vol. 1, pp. 910-917.

[8] Alneyadi S., Sithirasenan E. &Muthukkumarasamy V. (2014), A Semantics-Aware Classification Approach for Data Leakage Prevention, In: Susilo W., Mu Y. (eds) Information Security and Privacy, ACISP 2014, Lecture Notes in Computer Science, vol. 8544, pp.413-421, Springer, Cham.

[9] Shirudkar, K. &Motwani, D. (2015), Big-Data Security. International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, issue 3, pp. 1100-1109.

[10] Tidke, P., Wagh, A., Bharade, D. &Dongre, A. G. (2015), Data Leakage Prevention with E-Mail Filtering, International Journal of Advance Foundation and Research in Computer (IJAFRC), vol. 2, issue 2, pp. 28-32.

[11] B. Raja Koti, Dr. G.V.S. Raj Kumar, and Dr. Y. Srinivas, "A Comprehensive Study and Comparison of Various Methods On Data Leakages", International Journal of Advanced Research in Computer Science, Volume 8, No.7, July – August 2017, pp-627-631, ISSN No: 0976-5697

[12] B Raja Koti, GVS Raj Kumar, Y Srinivas, "Identification of Guilt Agent and Leaked Data by Using MAC-IP", International Journal of Applied Engineering Research, 2017, Volume 12, Issue 22, pp 12237-12245.

[13] B Raja Koti, G V S Raj Kumar, "Information leakage detection and protection of leaked information by using the MAC-IP binding technique", International Journal of Engineering &Technology, Volume 7, Issue 1.7, 2018, pp230-235.

[14] P. Zilberman, S. Dolev, G. Katz, Y. Elovici and A. Shabtai, "Analyzing group communication for preventing data leakage via email," Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics, Beijing, 2011, pp. 37-41.

[15] R. Tahboub and Y. Saleh, "Data Leakage/Loss Prevention Systems (DLP)," 2014 World Congress on Computer Applications and Information Systems (WCCAIS), Hammamet, 2014, pp. 1-6.

Short-Term Load Forecasting for Electrical Dispatcher of Baghdad City based on SVM-FA

Aqeel S. Jaber[1], Kosay A. Satar[2], Nadheer A. Shalash[3]
Department of Electrical Power Engineering Techniques
Al-Mamoun university college
Baghdad, Iraq

***Abstract*—The improvement of load forecasting accuracy is an important issue in the scientific optimization of power systems. The availability of accurate statistical data and a suitable scientific method are necessary for a perfect prediction of future occurrences. This research deals with the use of a regression forecast model (Support Vector Machine, SVM) for the prediction of the vector data for electrical power loading and temperature in Baghdad city. The Firefly algorithm was used to optimize the parameters of the SVM to improve its prediction accuracy. The quantitative statistical performance evaluation measures (absolute proportional error (MAPE)) were used to evaluate the performance of the optimization methods. The results proved that the modification method was more accurate compared to the basic method and PSO-SVM.**

***Keywords*—SVM; FA; Load forecasting; PSO**

I. Introduction

In the context of increasing power demand and development of the power market, load forecasting is a major challenge in terms of power system planning and operation [1-3]. The advantage of a precise load forecasting is to help the operators make decisions for the commitment unit, reduce the reserve capacity, and make a proper schedule for maintenance planning [4]. Apart from holiday forecasting, the power system experts have divided forecasting into long, medium, and short-term forecasting based on the duration of the term.

The intelligent methods of power system forecasting have been increased significantly over the last decades to reduce the error in the expected power system changes [5-9]. The Autoregressive Moving Average (ARMA) with the Artificial Neural Network (ANN) is accurate in predicting small power and depends only on the data [10]. Jie Shi has applied a hybrid forecasting model based on grey relational analysis and wind speed distribution features [3]. Many solutions have been introduced in electrical problems since 2008 when the Firefly algorithm (FA) was proposed by Xin-She Yang [11]. Bayesian Neural Networks, Deep neural networks, and Deep neural networks have been assumed as a feasible method for presaging the power in this year [7-9].

From the other hand, the SVM is a regression technique which was feigned by Vladimir Vapnik in 1995 [11]. Since its development, its usage in forecasting has increased gradually [11]. Xing et al suggested many hybrid prediction methods based on SVM, and one of the methods depend on the use of least squares support vector machine [12]. The PSO-SVM was developed by many researchers such as: Nadtoka I for short forecasting [13, 14].

However, in addition to the non-suitable parameters for SVM method, there is no powerful method yet in the area of short-term prediction [15]. This problem is because the SVM accuracy depends on the suitable values of two parameters – the Lagrange multipliers (α) and the regularization parameter (C)[16].

In this paper, a new method which depends on two methods (Firefly optimization and SVM (FA-SVM)) was used to predict the power load of Bagdad city. Firefly optimization was used to optimize α and C parameters to ensure an optimal load forecasting. This forecasting was performed using one-year real data (power and weather) of Baghdad and the performance of the proposed method on the forecasting accuracy was compared to that of two forecasting methods. For each iteration of SVM, many iterations were performed to select the values of α and C. A comparison of the absolute proportional error between SVM, PSO-SVM, and the proposed methods was performed, and the results showed that the proposed method was more efficient in prediction compared to the other methods in terms of forecasting the system within a short duration.

II. Theoretical Background

A. SVM

Basically, any learning method depends on a driven data which is based on the multi-disciplinary ideas that combine with the facility of a computer to perform an expected action on other datasets [16]. Fig. 1 provides a schematic flowchart of any simple learning model [17].

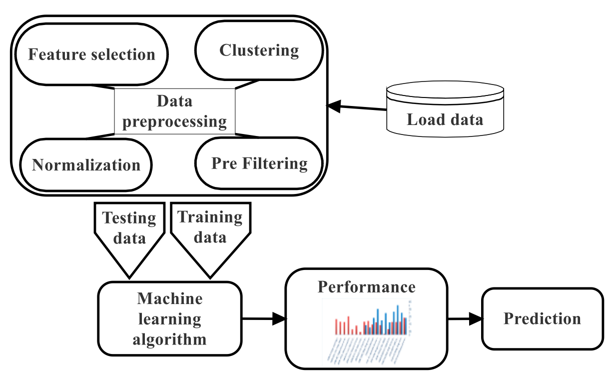


Fig. 1. Machine Learning basic Model.

Vladimir proposed one of the most famous artificial intelligent learning machines and named it as Support Vector Machines (SVM) [18]. Basically, the SVM optimization equations can be expressed as [13]:

$$\min J = \frac{1}{2}\|\omega\|^2 + \frac{1}{2}C\sum_{i=1}^{l} e_i^2 - \sum_{i=1}^{l}\alpha_i(\omega^T\varphi(x_i) + b + e_i - y_i) \quad (1)$$

Where:- e_i is the initial error, $e \in R^{l\times 1}$ is the vector error, C is the regularization parameter, α are the Lagrange multipliers, $\alpha \in R^{l\times 1}$. Also, the nonlinear forecasting model can be expressed as [13]:

$$P_{i+1} = \sum_{i=1}^{m}\alpha_i K(x_i, x) + b \quad (2)$$

Where:- $i = 1,\ldots,m$; x_i is the input vector, $P(t)$ = power, $T(t)$= temperature, and $L(t)$ = natural illumination; x = coordinates of the center of the scattering vector; α_i, b = linear coefficients; m = dimension of the input vector; $K(x_i, x\ )$ = Kernel function which performs a nonlinear mapping of the input data space.

B. Firefly Algorithm

The flashing behavior of fireflies inspired Xin-She Yang to develop a novel swarm-intelligence-based method of optimization in 2008 [19]. The light signal from a firefly serves as a signal that attracts the other fireflies towards itself. Mathematically, all the fireflies in the FA are assumed to be unisex and can be attracted to each order based on the intensity of the emitted light from each firefly [20]. The objective function is the leader of any combination of the two brightest fireflies in one population on every algorithmic iteration [21].

As a firefly's attractiveness is proportional to the light intensity which is noticed by adjacent fireflies, the variation of attractiveness with the distance r can be described in terms of the attractiveness $at\ r$ by [22]:

$$\beta = \beta_o e^{-\gamma r^2} \quad (3)$$

The change in the position of firefly *i* which is attracted to firefly *j* is calculated by [22]:

$$x_i^{\ t+1} = x_i^{\ t}\beta = \beta_o e^{-\gamma r_{i,j}^2}(x_j^{\ t} - x_i^{\ t})\alpha_t \epsilon_i^t \ + \pi r^2 \quad (4)$$

```
Objective function f(x), x=(x1,.....xd)T
 Generate initial population of fireflies xi (i=1,2...,n)
 Light intensity Ii at xi is determined by f(xi)
 Define light absorption coefficient ?
 While (t<MaxGeneration)
    for i= 1:n all n fireflies
      for j= 1:n all n fireflies (inner loop)
      if (Ii < Ji), Move firefly I towards j; end if
        Vary attractiveness with distance r via exp[-?r]
        Evaluate new solutions and update light intensity
      End for j
    End for i
 Rank the firefly and find the current global best g*
 End while
 Post process results and visualization
```

Fig. 2. Firefly Algorithm.

Where:- ϵ_i^t is a vector of random numbers on time t, which depends on Gaussian and The third term is randomization with α_t. The firefly algorithm is as shown in Fig 2 [18].

III. The Proposed Method

One of the main challenges in Baghdad load forecasting is the low accuracy of prediction due to the high complexity of the power system and the climate data. Also, the accuracy of SVM forecasting depends on two parameters which are the regularization parameter and the Lagrange multipliers. These parameters represent the degree of correlation on the SV learning specimen input spaces. Therefore, a new hybrid approach based on SVM and FA is introduced in this study to deal with these two issues. The employment and selection of the optimal values of the two parameters of SVM (σ , C) were based on the FA algorithm. Normally, any evolutionary algorithm requires a fitness function which will serve as the optimization objective [15]. In the area of forecasting, there are many criteria-based methods to gauge the accuracy of a system, such as the relative percentage error, the mean absolute percentage error, the root mean square error, and the absolute proportional error which is selected to gauge the proposed method. The proposed method is shown in the following algorithm:

1. **Collect the data of Bagdad city (climate, load power).**
2. **Separate the date as seasons.**
3. **Do the training set according to:**
4. $S\left\{\begin{matrix} \boldsymbol{P_1(power1), T_1(temperature1),} \\ \boldsymbol{P_m(power\ m), T_m(temperature\ m)} \end{matrix}\right\}$.
5. **Determine the α and C using FA and the kernel function $K(x_i, x)$.**
6. **Construct the objective function (6); where b is determined as:**
7. $b = P_i - \sum_{i=1}^{m}\alpha_i K(x_i, x) \quad \alpha_i \in (0, C) \quad$ **(7)**
8. **Calculate the expected power of the next days of unknown load.**

The flow chart of the FA-SVM for the mid-season is shown in Fig. 3.

The selected limits of the initial values of C and σ are [0.1, 250] and [0. 1, 15], respectively, m = 25. The main FA parameters are set to the optimal settings $\beta 0$ = 0.25, γ = 1.0, α = 0.3, and the number of Fireflies = 25, number of iterations = 80. For the PSO, the bird step = 25, ω = 0.7; acceleration constants c_1 and c_2 = 1.8 as earlier recommended by [13].

The real data of Bagdad city was taken for the period of 1st March 2017 – 31st December 2017 and used for the calculation of daily forecasts using the three compared methods. The power data was obtained from the Iraqi national dispatch center, while the climate data were obtained from the Iraqi meteorological organization and seismology. the comparison between the methods depended on the prediction errors (mean absolute error percentage mape) calculated as follows:

$$\text{MAPE} = \frac{1}{24 \cdot N} \cdot \sum_{i=1}^{24}\sum_{k=1}^{N}\frac{\left|P_{i,k}^{forcast} - P_{i,k}^{actual}\right|}{P_{i,k}^{actual}} \cdot 100\% \quad (5)$$

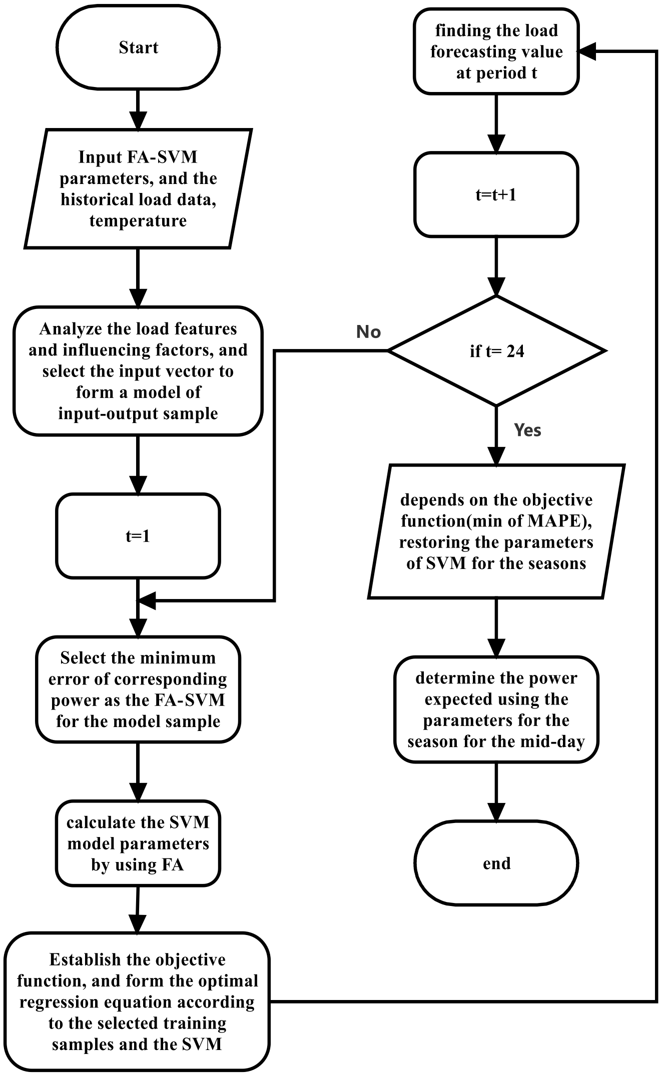


Fig. 3. FA-SVM Flowchart.

IV. Simulation Results

The aim of the work is to examine the capability of the proposed FA-SVM in reducing the mean absolute error percentage (MAPE) in high complex prediction problems. The simulation was conducted in a MATLAB platform for the three algorithms during the forecast of the power load of Baghdad city. The SVM was first used before using the hybrid technique (PSO-SVM) to select SVM parameters. Finally, the FA was used to optimize the selected SVM parameters. The prediction results were compared with the real data based on the mean absolute error percentage. Figs. 4-7 showed the real data and forecasted load for a single day in spring, summer, autumn, and winter seasons. From these Figs, it was evident that the proposed method achieved results that closely matched the real data.

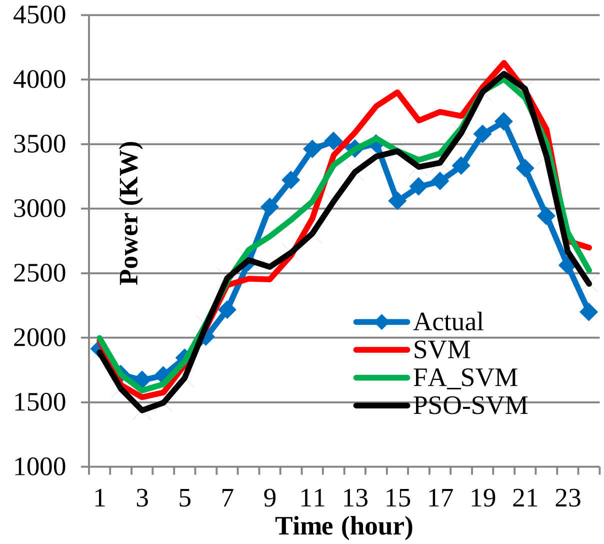


Fig. 4. Daily Forecasting Load Curve of Mid-Spring.

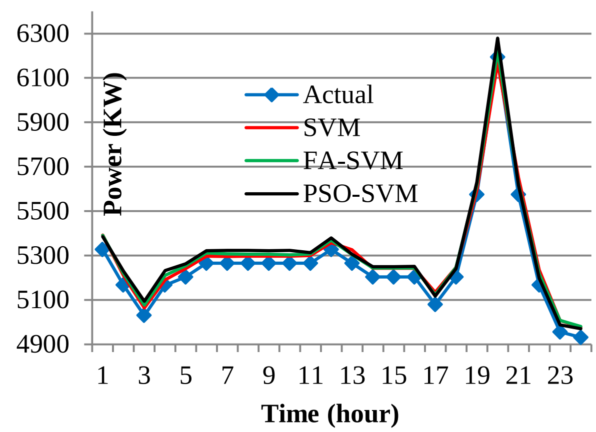


Fig. 5. Daily Forecasting Load Curve of Mid-Autumn.

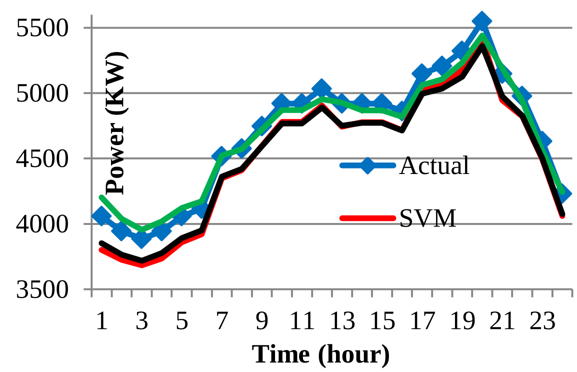


Fig. 6. Daily Forecasting Load Curve of Mid-Summer.

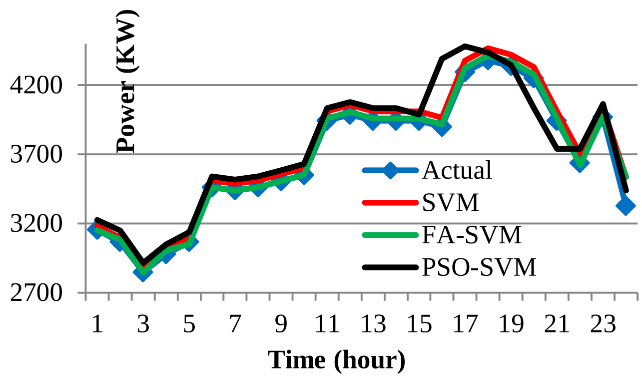


Fig. 7. Daily Forecasting Load Curve of Mid-Winter.

TABLE I. MAPE FOR SVM

SVM				
date	21-3 2017	21-9 2017	21-6-2017	21-12 2017
C	30	30	30	30
σ	2	2	2	2
MAPE%	4.55	0.8	3.65	1.76
Overall (MAPE) 2.69 %				

TABLE II. MAPE FOR PSO-SVM

PSO-SVM				
date	21-3 2017	21-9 2017	21-6 2017	21-12 2017
C	2.7	180	14	205
σ	7.011	2.54	2.64	10.12
MAPE%	4.5	0.8	3.51	1.39
Overall (MAPE) 2.55 %				

TABLE III. MAPE FOR FA-SVM

FA-SVM				
date	21-3 2017	21-9 2017	21-6 2017	21-12-2017
C	2.0601	150.0000	2.9746	149.9249
σ	6.4242	2.7100	0.1000	0.7775
MAPE%	3.61	0.75	1.27	0.62
Overall (MAPE) 1.5875 %				

Tables I, II, and III show the results of percentage error be for different days and methods. Form this tables, it can be concluded that FA-SVR has more ability to reach the load characteristics than SVM and PSO-SVM. Also, the tables are explaining the selected values of VSM parameters for each method.

V. CONCLUSION

An SVM-FA hybrid algorithm was proposed in this study for the forecast of the daily load power of Baghdad city for the period of 1st March 2017 – 31st December 2017. This data covered the four available seasons of the year in Iraqi. In this proposed method, the FA was employed to adjust the values of the SVM model parameters in order to achieve more promising data. The results obtained from the proposed SVM-FA method showed an acceptable percentage error margin with the real data. The accuracy of the methods was verified and compared using the absolute proportional error as a criterion.

REFERENCES

[1] Q. Wu, "Product demand forecasts using wavelet kernel support vector machine and particle swarm optimization in manufacture system," J. Comput. Appl. Math., vol. 233, no. 10, pp. 2481–2491, 2010.

[2] G. Sreenivasan, "PSO Based Short-Term Hydrothermal Scheduling with Prohibited Discharge Zones," IJACSA - Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 9, pp. 97–105, 2011.

[3] J. Shi, Z. Ding, W. Lee, Y. Yang, Y. Liu, and M. Zhang, "Hybrid Forecasting Model for Very-Short Term Wind Power Forecasting Based on Grey Relational Analysis and Wind Speed Distribution Features," IEEE Trans. Smart Grid, vol. 5, no. 1, pp. 521–526, 2014.

[4] C.-M. Huang, C.-J. Kuo, and Y.-C. Huang, "Short-term wind power forecasting and uncertainty analysis using a hybrid intelligent method," IET Renew. Power Gener., vol. 11, no. 5, pp. 678–687, 2017.

[5] Y. Jiang, X. Chen, K. Yu, and Y. Liao, "Short-term wind power forecasting using hybrid method based on enhanced boosting algorithm," J. Mod. Power Syst. Clean Energy, vol. 5, no. 1, pp. 126–133, 2017.

[6] J. Shi et al., "Hybrid Forecasting Model for Very-Short Term Wind Power Forecasting Based on Grey Relational Analysis and Wind Speed Distribution Features," vol. 5, no. 1, pp. 521–526, 2014.

[7] S. Zheng, Q. Zhong, L. Peng, and X. Chai, "A Simple Method of Residential Electricity Load Forecasting by Improved Bayesian Neural Networks," Math. Probl. Eng., vol. 2018, 2018.

[8] M. Easley, L. Haney, J. Paul, K. Fowler, and H. Wu, "Deep neural networks for short-term load forecasting in ERCOT system," in 2018 IEEE Texas Power and Energy Conference, TPEC 2018, 2018, vol. 2018–Febru, pp. 1–6.

[9] T. Vantuch, A. G. Vidal, A. P. Ramallo-Gonzalez, A. F. Skarmeta, and S. Misak, "Machine learning based electric load forecasting for short and long-term period," in IEEE World Forum on Internet of Things, WF-IoT 2018 - Proceedings, 2018, vol. 2018–Janua, pp. 511–516.

[10] S. Zhang, Y. Wang, M. Liu, and Z. Bao, "Data-Based Line Trip Fault Prediction in Power Systems Using LSTM Networks and SVM," IEEE Access, vol. 6, pp. 7675–7686, 2017.

[11] P. Vähäkyla, E. Hakonen, and P. Léman, "Short-term forecasting of grid load using Box-Jenkins techniques," Int. J. Electr. Power Energy Syst., vol. 2, no. 1, pp. 29–34, 1980.

[12] W. H. NIU Dong-xiao, GU Zhi-hong, XING Mian, "Study on Forecasting Approach to Short-term Load of SVM Based on Data Mining," in Proceedings of the CSEE, 2006, pp. 1–6.

[13] I. I. Nadtoka and M. Al-Zihery Balasim, "Mathematical modelling and short-term forecasting of electricity consumption of the power system, with due account of air temperature and natural illumination, based on support vector machine and particle swarm," in Procedia Engineering, 2015, vol. 129, pp. 657–663.

[14] G. Ren, S. Wen, Z. Yan, R. Hu, Z. Zeng, and Y. Cao, "Power Load

Forecasting Based on Support Vector Machine and Particle Swarm Optimization," no. 1, 2016.

[15] A. Kavousi-Fard, H. Samet, and F. Marzbani, "A new hybrid Modified Firefly Algorithm and Support Vector Regression model for accurate Short Term Load Forecasting," Expert Syst. Appl., vol. 41, no. 13, pp. 6047–6056, 2014.

[16] L. Demidova, E. Nikulchev, and Y. Sokolova, "The SVM Classifier Based on the Modified Particle Swarm Optimization," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 2, 2016.

[17] S. N. Fallah, R. C. Deo, M. Shojafar, M. Conti, and S. Shamshirband, "Computational intelligence approaches for energy load forecasting in smart energy management grids: State of the art, future challenges, and research directions," Energies, vol. 11, no. 3, 2018.

[18] S. Dai, D. Niu, and Y. Han, "Forecasting of Power Grid Investment in China Based on Support Vector Machine Optimized by Differential Evolution Algorithm and GreyWolf Optimization Algorithm," Appl. Sci., vol. 8, no. 636, pp. 1–19, 2018.

[19] I. Fister, X. S. Yang, and J. Brest, "On the randomized firefly algorithm," Stud. Comput. Intell., vol. 516, pp. 27–48, 2014.

[20] X. S. Yang and X. He, "Firefly algorithm: recent advances and applications," Int. J. Swarm Intell., vol. 1, no. 1, p. 36, 2013.

[21] N. Ali, M. A. Othman, M. N. Husain, and M. H. Misran, "A review of firefly algorithm," ARPN J. Eng. Appl. Sci., vol. 9, no. 10, pp. 1732–1736, 2014.

[22] A. S. Jaber, "A Novel Tuning Method of PID Controller for a BLDC Motor Based on Segmentation of Firefly Algorithm," Indian J. Sci. Technol., vol. 10, no. 6, 2017.

A Novel Student Risk Identification Model using Machine Learning Approach

Nityashree Nadar[1]
Bharathiar University, Coimbatore

Dr.R.Kamatchi[2]
Amity School of Engineering and Technology
Panvel, Mumbai

***Abstract*—This research work aim at addressing issues in detecting student, who are at risk of failing to complete their course. The conceptual design presents a solution for efficient learning in non-existence of data from previous courses, which are generally used for training state-of-art machine learning (ML) based model. The expected scenario usually occurs in scenario when university introduces new courses for academics. For addressing this work, build a novel learning model that builds a ML from data constructed from present course. The proposed model uses data about already submitted task, which further induces the issues of imbalanced data for both training and testing the classification model. The contribution of the proposed model are: the design of training the learning model for detecting risk student utilizing information from present courses, tackling challenges of imbalanced data which is present in both training and testing data, defining the issues as a classification task, and lastly, developing a novel non-linear support vector machine (NL-SVM) classification model. Experiment outcome shows proposed model attain significant outcome when compared with state-of-art model.**

***Keywords*—Classification; imbalanced data; machine learning; virtual learning environment**

I. Introduction

Student dropout is an important problem across various levels such primary school, higher secondary, graduation level and the scenario is much worse in Massive Open Online Courses (MOOCs). As per the research conducted in [1], [2], the number of student not completing graduation in USA is 20% and in Europe it is around 20% to 50% fail to finish their studies on time [3]. For online or distance education, these statistics are even worse with 78% of students not completing the graduation [4]. Further, it gets even worse for student who gets registered with MOOCs, the percentage of student who enrolled and successfully finished the course is only 5% as reported in [5] or 15% as reported in [6]. The issues of identifying student that re expected to fail the course has been extensively analyzed across various research community in recent times [7], [8], [9]. It was also a major subject of the KDD'CUP 2015 competition that mainly aimed on forecasting student withdrawing from online courses.

Establishing student, who are at chance or risk of withdrawing or failing from their respective course, is the initial step towards provisioning them with remedial (material) support. Generally, supportive measures are carried out by instructor/professor, who obtains the information/outcome of forecasting [7], [8]. In other way, the forecasting model may build email messages that communicate directly to the student [10]. The preliminary objective is to enhance the student learning, to keep student engaged in course, and aid them completing the research or study programs.

In distance or online courses, most material are delivered through Virtual Learning Environment (VLE). In VLE each and action are recorder and stored. Along with, student information such as assessment, task results, and demographic information etc, are also kept. These data are cleansed and ML is applied to build a forecasting/predictive model. These model are then used to offers online course provider to forecast student at-risk of completing it on time. A generic way of building a predictive model is to train the models using legacy data from a history or previous task submitted information of the course [8]. Further, it is applied to the present presentation. However, adopting these methods will not be efficient when applied to new type of courses that has no history. For such case, it is important to find new solution.

From extensive survey carried out by MOOCs [11] and Higher Education (HE) courses [8] shows that the highest amount of dropout occurs during first year's courses, and many student dropouts even with a month, first few weeks of the course presentation. The cause may be also due to fee payment toward courses. Therefore, the objectives are to establish or find student who are at-risk of dropping out or failing to complete on time as early as possible. It must also be noted that the same behavior or pattern may not be same across different university/education institution or course design, rapid student dropping out of course may also arise in late stage of course [9].

For overcoming research challenges this work, this work aimed at designing a forecasting model that identify student at-risk of failing or completing on time by presenting a novel non-linear support vector machine (NLSVM) classification model.

The contribution of work is as follows

- Presenting a non-linear enhanced support vector machine classification model for identifying student risk of failure. The NLSVM can be used as both binary classifier as well as multi-level classifier.
- Our model attain good accuracy performance when compared with state-of-art model.
- Experiment outcome shows good performance in terms of ROC, F-measure, and precision and recall.

The paper is organized as follows: In section II, extensive survey is presented. Experimental study are discussed in section III. Finally section IV the paper is concluded and future work of research is described.

II. Litterature Survey

Machine learning technique is composed of supervised, semi-supervised and non-supervised is widely applied and used across various state-of-art models [11], [12], [13], [14], [15], [16], [17], [18], [20], and [21] for identifying risk of student failing to complete course on time. The basic conception is to utilize legacy data to learn the forecasting models and to utilize these approaches to perform forecasting on current courses. The data can aid the course provider who is aiming to address or build policies to enhance the student performance (student retention rate) and student dropping out of courses or failing to finish on time. In [12], the approaches for finding failure or success of student were trained using data of their prior study result. it can be seen that forecasting failure for the first term of courses is very important, since the dropout rate is generally higher but with suitable policies or strategies (help) many student can be saved [19].

Behavior of students [20], [21] in the VLE can be used to construct forecasting models for online courses. These could be just simple summary statistics [15]. When neither the students virtual learning environment activities nor the student prior study results are available, demographic information can be used as the major foundation of information [16].

The proposed learning model is constructed using state-of-art models at the OU [13], [17], [18], [19]. Initially, using decision tree that is trained using data labeling student behavior in the virtual learning environment complemented by the scores of the past assessments/tasks [17]. Further, [18] used demographic features for enriching the input data for training model. The significant discovery in [19] was the prominence of the early establishment or finding of students at risk, even prior to the first task/assessment in the course. The students who do not submit or fail to complete the assessment are very likely to fail or withdraw the entire course. Further, number of approached [7], [22], and [23] for solving problem of classification with presence of imbalanced data in forecasting or identifying student at-risk of failing. However, they neglected student who haven't shown any interest in performing tasks and only focused on active students. For overcoming research challenges this work present a novel non-linear based supervised classification model.

III. Proposed Non Linear Support Vector Machine based Student Risk Identification Learning Model

This manuscript present a novel learning design that use data from running presentation for training forecasting model. The fundamental objectives is to use the information of students who have already completed and submitted the future task and analyze the behavior pattern of find the students who are at risk of failing to submit the assignment. It is assumed that the behavior pattern of student who are about to submit are expected to follow identical behavior pattern as those who already completed the completed and submitted the task similarly, the behavior pattern will different for student who don't complete or submit their task. Number of machine learning based classification model is available to utilize and attain efficient learning model. However, in this work, we present a classification model as a binary classification problem. However, it can work even for solving multi-label classification problem. That is, for a given day (present), which is k days before deadline data, the objective of this work is to build a binary classification algorithm that forecast whether the student will submit the assignment or not on/before time (i.e., within the future k days). If $k = 0$, forecasting are done on the deadline day. Only students that are enrolled in course and haven't finished the task yet are considered for the forecasting.

A. System Model

Let's consider the deadline data and the date when the forecasting is done, which is $k$days prior to the deadline day, as forecasting date. For able to construct a forecasting model for period [forecasting date; deadline date] such that d deadline date is equal to forecasting date. The k forecasting date and d deadline day can be established as a template forecasting and deadline days, respectively.

Here, the deadline is within three days from the present day and we want to forecast if set of student submit their assessment or task either today or within next 5 days. The information for the present day are inaccessible, so the training data will come from the days [presentation initilaizaed+5] = 10 with the labels of submission in [present+4; present + 1] = [9; 6].

It shows the virtual view of the days for training and testing data, day = 0 depicts the present day, negatives keys shows to known information and positive keys to new/unknown data. This aids, that we have more days vacant when applying the forecasting model, some previous/older days cannot be utilized as they were not present in training stage.

B. Long-Term vs Short Term Labelling Window Tradeoff Modelling

Based on system model described, using long-term history means the window sampling for labels is growing. The more days prior to the deadline date, the more days is required for training labels. The condition for the present day being 0 to 5 days prior to the deadline date. For k days prior to the deadline, the size of the window for both training and testing labels will be $k + 1$.

C. Feature Selection for Learning

The data available for efficiently learning is composed of information such as activities and demographics in the virtual learning environment. For extensive analysis this work carried out it can be seen the demographic data is static in nature, it is important to carryout transformation of these information, such as standardization for numerical data and vectorization of categorical data. Similarly, the virtual learning environment data are generally are composed of very rich information such as daily click events clustered by precise action, i.e., student X has viewed 15 times a particular document or presentation research material. All the events/actions are clustered into actions types such as video, resources, blogs, etc. For a given day (present) when the algorithm (model) is learned, the virtual learning environment features are aligned in reverse with respect to time on a particular days, i.e., day 0 is the present day, day 1 is depicted as yesterday and so on. The oldest day utilized for training is the day that the course is initialized. In addition to virtual learning environment daily counts, it's likely to obtain various statistical information of student behavior pattern in the virtual leaning environment, such as the how long (days) a student is active in the virtual learning environment (i.e., when a person (student) has last accessed or has logged in).

D. Addressing Imbalanced Data Problem in Classification

The ML algorithm are generally modelled to learn objective parameter from data when the classes in the training information are balanced. However, considering real-world environment, the data are generally imbalanced (i.e., some classes data will have significantly less data than other classes). As a result, the state-of-art algorithm [24], [25], [26], [27], and [28] performs very poor in identify probability of risk of failure of student that has been modelled so far [10]. For addressing the problem of imbalanced data the following two stages must be considered such as: **Algorithm stage: -** on-class or linear classification models, cost-sensitive learning, and various kind of ensemble algorithm model are some of the design are generally used. **Data stage:-**by applying sampling window for modifying the class label distribution in such way the training data becomes more balanced. The key functionality of cost-sensitive based learning model is to penalize the cost parameter error on marginal class variable during training stage, which is done by using a cost matrix. However, for attaining fine-grained binary classification model it is better to fix the weight parameter for minority classes (i.e., by considering weight of majority class will be 1). Further, number of approached [7], [22], and [23] for solving problem of classification with presence of imbalanced data in forecasting or identifying student at-risk of failing. However, they neglected student who haven't shown any interest in performing tasks and only focused on active students.

E. Forecasting Model using Machine Learning Model

For training the learning algorithms and for evaluation of our model, this work conducted survey of various exiting machine learning based classification models such as logistic regression, Naïve Bayes, support vector machine, Tree Boosting XGBoost [29] and so on. Further, number of approach used XGBoost for forecasting student's dropout as described in KDD-CUP15. However, these model are not efficient when the data is linearly non-separable. As a result, incur degradation in accuracy of classification performance. Further, very few algorithm provision probabilistic forecasting. As this aids in ordering students based on their likeliness to fail, and then use the resources constraint.

For overcoming research challenges, this work present non-linear support vector machine (NLSVM) classification model. The NLSVM first extract features of student behavior considering different assessment which are labeled (Completed: 0, and Failed: 1). If the task or assessment is successfully completed by student the class is labelled as 0 or if fails to submit on time then it is labeled as 1. The NLSVM then trains itself using labelled parameter and obtains support vector among parameters that maximize the distance among varied classes. Lastly, the NLSVM constructs a decision boundary (DB) from the support vectors. If the computed outcome from the DB is different from its known label, the decision is considered as training error. For such cases, this work considers soft-margin support vector machine which can fix boundary even when the datasets are mixed and cannot be disjointed. This work introduced slack parameters to maximizing the margin and reduce training error. For computing support vector using proposed NLSVM model is obtained as follows

$$\min \mathbb{D} \sum_{z=1}^{\mathbb{O}} \mu_z + \frac{1}{2}\langle u_t, u_t \rangle, \tag{1}$$

Such that $\mathbb{U}_z(\langle u_t, z_z\rangle + \mathbb{c}_{\mathbb{t}}) \geq 1 = \mu_z$ for $z = 1,2,3,\dots,\mathbb{O}$, and $\mu_z \geq 0$

Where $\mathbb{O}$ is the number of vectors, $\mathbb{D}$ is regulation variable, u_t is the weight vector, μ_z is the slack parameter and $<.,.>$ is the inner product function. The $\mathbb{U}_z$ is the z^{th} target parameter, $\mathbb{c}_{\mathbb{t}}$ is the bias, and z_z is the z^{th} input parameter. The SVM decision boundary $\mathbb{G}_z$is expressed as follows

$$\mathbb{G}_z = \langle u_t^*, z\rangle + \mathbb{c}_{\mathbb{t}}^* = 0 \tag{2}$$

Where $\mathbb{c}_{\mathbb{t}}^*$ represent bias and u_t^* represent weight vector, and z is the input feature. By transpiring the z_z and z term to $z_z \rightarrow (z_z)$ and $z \rightarrow (z)$, the non-linear support vector machine can be transformed to linear based support vector machine as follows

$$\mathbb{U}_z(\langle u_t, (z_z)\rangle + \mathbb{c}_{\mathbb{t}}) \geq 1 \tag{3}$$

Further, to perform classification using non-linear support vector machine, a kernel function $\mathbb{L}_{\mathbb{t}}(.,.)$, which is a dot-product in the transpired feature vector space as follows

$$\mathbb{L}_{\mathbb{t}}(z_z, z_{z'}) = \langle (z_z), (z_{z'})\rangle \tag{4}$$

where$z_{z'} = 1,2,\dots,\mathbb{O}$.

Further, the proposed NLSVM model is evaluated considering various ICT data which is composed of different behavior of different student data considering various task or assessment. Our model accurately classify these imbalanced data compared to state-of-art model which is experimentally shown in next section below.

IV. Experimetal Result and Analysis

This section evaluates performance evaluation of proposed student risk identification leaning model over state-of-art models. For experiment analysis various experiment are considered and date used for experiments are publically available. The experiment is conducted using windows 10 operating system, Intel I-5 class 64 bit processor, 16 GB RAM, 4GB Nvidia CUDA enabled GPU. For experiment analysis this work used publicly available dataset obtained from OULAD [30], [31] which composed of different courses with student enrollment around 1200 to 2500. The objective of this work is to forecast the submission of first assessment of a particular course within deadline time around 20 to 30 days. The course is composed of wide variety of fields such math's, history, engineering and so on. For completing the course, the student has to attain some minimum scores for a given task or assessment and then pass the final exam. The proposed student risk learning model for forecasting dropout has been aimed at attaining following objectives, Firstly, carryout analysis daily using ML algorithm to evaluate classification model. Secondly, analyses and identify the effects and problems of imbalanced data. Then, compare our proposed model over state-of-art model trained using legacy data. Fourthly, experiment is conducted for different *k* and courses and evaluate performance attained by proposed model over existing model in terms of precision, recall, F-measure, and ROC. The Table I, shows performance outcome attained by proposed model over existing model in terms of precision, and recall. Further, fig. 1 shows F-measure attained by both proposed and existing model. The overall result attained shows proposed learning model improves F-measure score by 24.45%, 26.65%, and 18.96% over existing learning model. An average improvement of 23.35% is attained by proposed learning model over existing model. Further, experiment are conducted to evaluate ROC performance attained by proposed NL-SVM over exiting SVM and XGBoost as shown in Fig. 2. The outcome shows NL-SVM attain an ROC performance improvement of 35.33%, 25.56% over SVM and XGBoost, respectively.

TABLE I. PERFORMANCE EVALUATION OF PROPOSED MODEL OVER EXISTING CLASSIFICATION MODEL

Top K	Precision SVM	Precision NL-SVM	Recall SVM	Recall NL-SVM
5	0.5751	0.88	0.8093	0.9
10	0.5044	0.89	0.8848	0.9
25	0.6698	0.88	0.8847	0.89

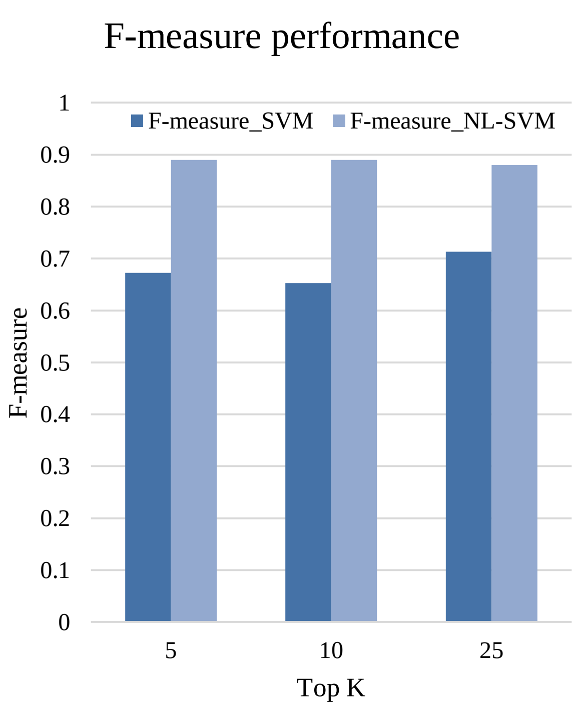


Fig. 1. F-Measure Performance for Different Top K.

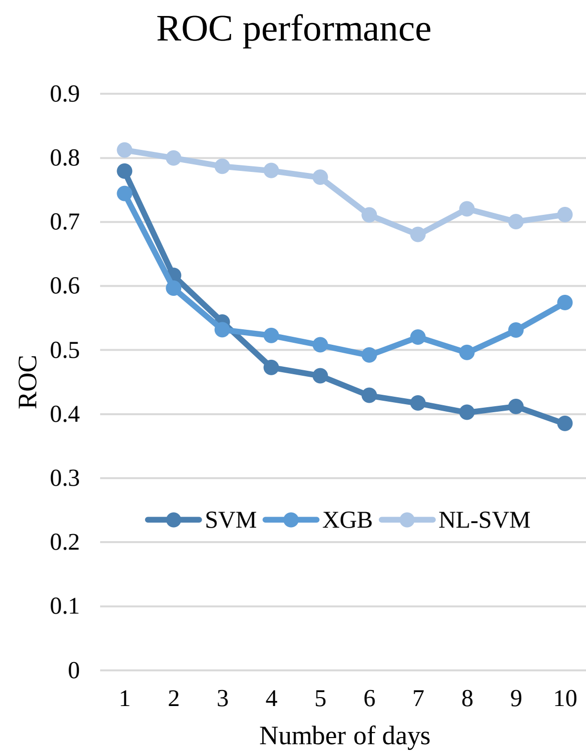


Fig. 2. ROC Performance for Number of Days.

V. CONCLUSION

This manuscript introduced a novel design for early finding of student who are at risk of failing or completing the course on time without using legacy data. The proposed model uses the significance factor of fist task being important factor in the progress of course work. The best way is to extract the student behaviour who already submitted their task and learn its pattern. This work defines the problem as a binary classification task with objective to learn and forecast daily using forecasting window. The proposed model is evaluated using publicly available OULAD dataset. The outcome shows the proposed model can predicts accurately even for early day (i.e. for 0 and 1 days), also predicts efficiently for later days of course completion, and attain better outcome that training using legacy data. From overall experiment analysis, it can be seen feature selection VLE is important for forecasting student at risk of failing. The proposed NL-SVM based classification model attains good recall, precision, and F-measure performance. An average F-measure improvement of 23.35% is attained by proposed learning model over existing model. Further, NL-SVM attains an ROC performance improvement of 35.33%, 25.56% over SVM and XGBoost, respectively. The future work we would consider experiment analysis considering different dataset and also considering building a hybrid model for enhancing forecasting model.

REFERENCES

[1] Peter J. Quinn. Drop-out and completion in higher education in europe among students from under-represented groups. Technical report, European Commission, Oct 2013.

[2] H. Vossensteyn, A. Kottmann, B. Jongbloed, and F. Kaiser. Drop-out and completion in higher education in europe executive summary. Technical report, European Commission, 2015.

[3] G. Kena, J. W. X. R. A. Musu-Gillette, Laurenand Robinson, J. Zhang, S. Wilkinson-Flicker, A. Barmer, and E. D. V. Velez. The condition of education 2015. Technical Report 2015-144, NCES, May 2015.

[4] O. Simpson. 22% - can we do better? In The CWP Retention Literature Review, 47, 2010.

[5] K. Jordan. Mooc completion rates: The data. http://www.katyjordan.com/MOOCproject.html, 2015. Accessed: 2017-10-10.

[6] D. Koller, A. Ng, C. Do, and Z. Chen. Retention and intention in massive open online courses: In depth. EDUCAUSE, http://www.educause.edu/ero/article/retention-and-intention-massive-open-online-courses-depth-0, Jun 2013.

[7] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauria, J. R. Regan, and J. D. Baron. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. Journal of Learning Analytics, 1(1):6{47, 2014.

[8] A. Wol_, Z. Zdrahal, D. Herrmannova, J. Kuzilek, and M. Hlosta. Developing predictive models for early detection of at-risk students on distance learning modules. In Machine Learning and Learning Analytics workshop at LAK14, 24-28 March 2014, Indianapolis, Indiana, USA, 4, 2014.

[9] H. He and E. A. Garcia. Learning from imbalanced data. IEEE Trans. on Knowl. and Data Eng., 21(9):1263{1284, Sep 2009.

[10] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. Predicting mooc dropout over weeks using machine learning methods. In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, 60{65, 2014.

[11] C. Taylor, K. Veeramachaneni, and U. O'Reilly. Likely to stop? predicting stopout in massive open online courses. CoRR, abs/1408.3382, 2014.

[12] Huang, S. & Fang, N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education, Issue 61, pp. 133-145, 2013

[13] Hlosta, M. et al., 2014. Modelling student online behaviour in a virtual learning environment. Indianapolis, LAK 2014.

[14] Pandey, M. & Sharma, V. K. A Decision Tree Algorithm Pertaining to the Student Performance. Analysis and Prediction. International Journal of Computer Applications, 61(13), pp. 1-5, 2013.

[15] Romero, C., López, M., Luna, J. & Ventura, S., 2013. Predicting students' final performance from participation in on-line discussion forums. Computers & Educaton, Issue 68, pp. 458-472.

[16] Wladis, C., Hachey, A. C. & Conway, K., 2014. An investigation of course-level factors as predictors of online STEM course outcomes. Computers & Education, Issue 77, pp. 145-150.

[17] Wolff, A., Zdrahal, Z., Nikolov, A. & Pantucek, M. Improving retention: predicting at-risk students by analysing behaviour in a virtual learning environment. s.l., LAK 2013.

[18] Wolff, A., Zdrahal, Z., Herrmannova, D. & Knoth, P., 2013. Predicting student performance from combined data sources. In: A. Peña-Ayala, ed. Educational Data Mining: Applications and Trends. Verlag: Springer International Publishing, pp. 175-202, 2013.

[19] Wolff, A. et al., 2014. Developing predictive models for early detection of at-risk students on distance learning modules. Indianapolis, LAK 2014.

[20] Wang, Rui & Chen, Fanglin & Chen, Zhenyu & Li, Tianxing & Harari, Gabriella & Tignor, Stefanie & Zhou, Xia & Ben-Zeev, Dror & T. Campbell, Andrew. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 10.1145/2632048.2632054, 2014.

[21] Wang, Rui & Chen, Fanglin & Chen, Zhenyu & Li, Tianxing & Harari, Gabriella & Tignor, Stefanie & Zhou, Xia & Ben-Zeev, Dror & T. Campbell, Andrew. StudentLife: Using Smartphones to Assess Mental Health and Academic Performance of College Students. 7-33. 10.1007/978-3-319-51394-2_2, 2017.

[22] J. He, J. Bailey, B. I. Rubinstein, and R. Zhang. Identifying at-risk students in massive open online courses. In AAAI, 1749{1755, 2015.

[23] N. Thai-Nghe, A. Busche, and L. Schmidt-Thieme. Improving academic performance prediction by dealing with class imbalance. In Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy , November 30-December 2, 2009, 878{883, 2009.

[24] C. Ye and G. Biswas. Early prediction of student dropout and performance in moocs using higher granularity temporal information. Journal of Learning Analytics, 1(3):169{172, 2014.

[25] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. 1909{1918, 2015.

[26] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In LAK '15, 93{102, New York, NY, USA, ACM,2015.

[27] J. Bainbridge, J. Melitski, A. Zahradnik, E. Lauria a, S. M. Jayaprakash, and J. Baron. Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education. The JPAE Messenger, 21(2):247{262, 2015.

[28] S. Jiang, M. Warschauer, A. E. Williams, D. ODowd, and K. Schenke. Predicting mooc performance with week 1 behavior. In EDM14, 273{275, 2014.

[29] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. CoRR, abs/1603.02754, 2016.

[30] Kuzilek, M. Hlosta, and Z. Zdrahal. Open university learning analytics dataset. In Data literacy for Learning Analytics workshop at LAK16, 26th April 2016, Edinburgh, UK, 9, 2016.

[31] Kuzilek, Jakub & Hlosta, Martin & Zdráhal, Zdenek. Open University Learning Analytics dataset. Scientific Data. 4. 170171. 10.1038/sdata.2017.171, 2017.

Decision Support System for Agriculture Industry using Crowd Sourced Predictive Analytics

Remya S, Dr.R.Sasikala
Research Scholar, Department of CSE
VIT University, Vellore, Tamilnadu.

***Abstract*—It is really tough to manually examine the raw data. The Datamining strategies are used to detect the applicable information from uncooked data. The data mining algorithms are efficient for retrieving a specific pattern. In Datamining techniques decision trees are the most commonly used methods for predicting the outcome or behavior of a pattern because they can successfully and efficiently visualize the facts. Presently several decision tree algorithms are advanced for predictive analysis. Right here we gathered a dataset for rubberized mattress, from coir board CCRI, and applied the several decision tree algorithms on the data set and as compared every one. Every set of rules gives a completely unique choice tree from the input statistics. This paper focuses in particular on the Fuzzy c4.5 set of rules and compares one-of-a-kind choice tree algorithms for predictive analysis. Here by using predictive analytics, a decision can be made for each rubberized firms**.

Keywords—Predictive analytics; coir fiber; fuzzy-C4.5; crowdsourcing

I. Introduction

Agriculture is a department of technology within the exercise of farming, soil cultivation for growing of vegetation and breeding of animals to afford our primary necessities such as food, medicines, clothes, shelter and the other products. The Indian coir enterprise accomplishes an essential role in the economy of the foremost coconut growing states such as Kerala and Tamil Nadu. Coconut husk is the basic raw material for coir products. At least 50 per cent of the available coir husk is used to produce coir products and the rest is used as fuel in rural areas. Subsequently, coir industry can be grown up [1]. Numerous development programs are also undertaken by the Government authorities for the improvement in quality and production of coir industry. These efforts widen the export markets for coir and coir products. Practically, it's far not possible to accumulate all of the husks for coir production and to achieve most utilization. However, evaluation has been done on the availability of husk for manufacturing, coir manufacturing capacity, actual manufacturing of coir and the capacity utilization globally. Here the goal is to advise a predictive evaluation system with the assist of crowdsourcing [2]

Crowdsourcing is an online tool or model used for solving a problem with the aid of, shared intelligence of interacted communities for precise purposes. The predictive analysis can predict the trends in future and behavior, can understand the outcomes and thereby improving the performance in the industry. It can also help to drive the strategic decision making. The analysis suggests that management authorities can analyzes the husk availability, utilization of husk, coir production, percentage share of the states to coir production and the production potential in India. A comparative analysis of India's coir production with the other countries can also be done with the help of this predictive analysis. Here by using a crowd sourcing algorithm it is possible to predict the international business of coir in South India with the other parts of the country and outside India. This can analyze the expectation of the clients in the areas regarding Ayurveda, Textiles and Agriculture with coir industry[3]. Thus the main objective is to make a better policy development, better decision-making, and thus lead to a sustainable growth in the coir industry for better future.

Indian coir production also has now not unfolded along all the states because of diverse motives together with loss of natural retting facilities, skilled labor and low cost technology. a preceding analysis on the husk availability, utilization of husk, coir production are accomplished and it exhibits that only a small percentage of the coconut husk produced in India is being utilized for coir manufacturing and the remaining major part unutilized. Efforts are being made by means of the coir board to improve the utilization of husks; however they do not be successful due to the absence of natural retting facilities in lots of locations.

The production of value added and innovative products helps to satisfy the global demand and by growing the boom of Kerala's traditional industry products in coir industry [4]. Therefore the success and future development of coir industry in South India particularly relies upon at the ability and the efficiency of the industry to discover the export markets of the coir products. Coconut husk is the primary raw material for the coir industry and the coir manufacturing concentrate mainly in the areas where it is more available. Even though India has a long coastline growing coconut palms; coir industry is concentrated in Kerala. Kerala is the traditional home of Indian coir industry and ranked as the first in coconut production and in coir production and in coir manufacturing and the increase may be in our arms**.**

II. International Trade of Coir

Global trade is the trade of capital, items, and services throughout international borders or territories. In maximum international locations, such alternate represents a substantial share of gross domestic product(Amarnath & Brindha, n.d.).

Development and automation of advanced technologies in this superior era, inclusive of transportation, globalization, multinational organizations, and outsourcing are all having a primary impact on the international trade system. Increasing worldwide alternate is vital to the continuance of globalization.

India and Sri Lanka are the two top manufacturers of coconut fiber within the international wide. They account for approximately ninety % of global manufacturing of coir fiber. Sri Lanka was the most important exporter of various fibers inside the international while India exports value added coir fiber merchandise including yarn, mats and rugs. Now a days the request for coir fiber products and the Medicinal, Ayurveda, fabric and Agricultural properties of coconut and coir are growing within the world wide especially in China, Malaysia and Germany[5].However there are such a lot of challenges for the coir industry to expand the markets and get into sustainable.

Small scale industries play a dynamic role in exploiting the resources and giving employability. Indian coir industry is one among that and each unit requires upto 15 works per day and having cost about Rs 5 lakhs in average. At present coconut fiber and ropes are plentiful in India and it has a significant request in the global market. We can establish the coir industry by considering the choice of the market in assist with a decision support system. The coir fiber can be utilized in many forms such as coir yarn, mats, mattress, floor coverings, coir pith and geo textiles.

A. The Challenges Include

- The fiber value is unreliable and the productivity is very low at the refine level, because of the dearth of a worth based pricing system for coir fiber
- Most of the industry participants do no longer diagnose the necessities at the technical level and distribution systems of current and rising markets for natural fiber merchandise due to the lack of information, and also they do no longer have the competitive opportunities for coir products in domestic markets.
- Now a day's markets for coir fiber are declined, due to the competition from different artificial substances and foreign suppliers.
- Local coir mills process only a fraction of the available husks.
- Working conditions and productivity are generally reduced.

So we ought to improve the reliability of product, expand progressive products together with the assist of customers in importing countries.

B. Global Market Context

- The technical advantage of coir when compared to other natural fibers is due to the high lignin content and its related stiffness.
- Mild degradation and better flexibility are the properties of coir fibers as compared to other different natural fibers which are used for upholstery works , erosion control products and other applications

Value-added products such as Tawashi brushes, floor coverings and high end twine can have increase or hold the global marketplace volume of coir products. One of the coir product rubberized mattress is discussed here for making coir industry a good challenge. The coir mattress is made by combining the high density of rubberized coir with the core. It is shielded by the jacquard fabric to give a perfect look and feel to the mattress. The rubberized coir mattress is made by using natural fibers such as coconut fiber and unprocessed rubber. It makes the reticulated layers circulate the air properly inside the mattress and maintains an ideal temperature for sleeping [6]. This type of mattress is highly hygroscopic in nature. It has the ability to hold water molecules from the surrounding surface if water or any other liquid splits over it. A mattress has the ability to preserve the moisture and maintain a refreshing impression on the body. Rubberized Coir mattresses are a combination of natural materials like coconut fiber and natural rubber. The combination creates a sheet that provides natural comfort, firm support, allows for air circulation and is non-absorbent[7]. The best quality of coir mattress comes with features such as it has sufficient layers of coir sheets which help it to maintain the firmness and give accurate support to the body. It offers the exact firmness with little cushioning to make it durable. It should be reliable for a long time.

The crowd sourcing based predictive analysis helps to analyses [8]

- Factors limiting market growth and how to improve the growth
- Current market trends and structure
- Expectation of clients
- Up-to-date view of technological improvements
- In depth analysis of business developments
- A precise growth in the technology
- Helps to predict what type of coir product can be developed from a coir samples

III. Trade Analysis in Production and Development

A. Coir

Coir is mined from the outer shell of coconuts surrounding the seed of the coconut palm called Cocos nucifera, and it is the thickest among the commercial natural fibers and more resistive in nature. It is a coarse short fiber. It has a very low decomposition rate and this nature is the added advantage of coir in geotextiles. The average length of coir fiber is 35 cm and having diameter of 12 to 25 microns units. Usually a coconut is harvested once in 45 days and an average of 10 kg of coir can be extracted from 1000 coconuts. The lignin component found in the coir makes this one resilient. But it is less flexible than cotton and so unsuitable for dyeing. The tensile strength of coir is less when compared to other natural fibers, but its good resistive nature to micro bial actions keeps it more suitable for agricultural purposes [5]. There is no need of chemical treatment. Mainly two types of coir are widely used such as brown fiber and white fiber. The brown fiber is extracted from matured coconuts and the white fiber is extracted from the immature green coconuts. Mature coir fiber contains more lignin content and less cellulose.

B. Environmental Benefits

Erosion is the main problem in the environment which affects our health also. Coir can be used to overcome erosion problems. Coir geotextiles has the capability to preserve moisture naturally and it protects from the sun radiation .Unlike other geo-synthetic material, the coir fiber provides better soil support up to 4 years

C. Uses of Coir

Traditionally coir is extracted from the coconut after the retting process of several months, but now days it is mechanized. There is an increased use of coconut husk de fibering machines. White coir is mainly used in the manufacturing of fishing nets. Brown coir is stronger than white coir and is used for the applications of sacking, doormats, insulation panels, mattresses.

D. Applications of Coconut Coir

Upholstery mattresses: Coir is a lot of broadly familiarized in upholstery industry and it is an acceptable interim for synthetic rubber. It can as well be get used to as an aggregate with natural rubber and is used for cushioning and other mattresses [9]

There are several reasons for using coir mattresses. But go head on trying coir mattresses based on your usages and expectations

- **Hygroscopic quality:** Coir mattresses have the ability to provide refreshing sleeps and it is very comfortable. It can easily absorb moisture.
- **Perfect ventilation:** They provide maximum air flow in and out the mattresses and that keeps it cool. Thus can enjoy sleep better.
- **Added support:** The coir mattresses are made with natural fiber only and they are very springy in nature and provide support for the whole body. This can minimize back pain also and aligns the spinal code accordingly
- **They are recommended for people with allergies:** They do not have any dust and so very much useful for allergic
- **Cooling sensation:** Its moisturizing absorbed nature allows continuous air flow and thus maintains a smooth temperature to sleep.
- **Friendly to the environment:** They are 100 percent natural and do not contain any chemicals and so suitable for our health and environment also.
- **Great affordability:** A good coir mattress can buy with an affordable price ranges in 600 to 1000 dollars .It is ideal for children ,adults and elder people
- **They are slightly durable:** They can be used for a long period of time and also durable in nature.
- **Require no maintenance:** No maintenance can be required for coir mattresses unlike other spring or foam mattresses.

At present the coir industry is highly industrialized in India and Sri Lanka only. However it is highly economic in Brazil, Indonesia, and Philippines etc. It is a small scale industry in India and people use local mills for extracting the fiber. Annually around 650000 tons of coir is producing globally, mainly in India and Sri Lanka. They are the main exporters in the coir industry. About 80 percent of the coir produced is carry across as raw fiber and remaining are in the form of yarn, rugs and mats.

India has made a wonderful effort to promote the coir industry including the national and international coir festivals and exhibitions. This effort supports our markets in the coir industry and we can meet the demand. Research and development efforts are also contributed to this area particularly in geo textiles and other new applications. Thus we can promise our prospects in the coir industry. Coconut grown in more than 90 countries in the world and that’s why we can grow internationally through coir business [10].

The main production is done by the small holders and so the production of fiber is scattered at small volumes. Integrated farm level processing is a solution this problem and can facilitate greater availability of technology for the dehusking and fiber extraction. Many of the coconut growing countries are not utilizing the husk for making value added products. So India can grow by more concentrating into it and thus provide better employment. Thus we can increase the income of coconut farmers and thereby reducing the poverty also. Improved production with increased qualities and quantities in coir industry must be one of our ultimate aims to grow India faster.

E. Actions taken for improving global markets

- An international based forum can promote the development section of the coir products, marketing section. It can improve the quality. It can direct the correct way of production and will help to prevent unhealthy competitions. Thus we can improve the business intelligence through this forum
- Innovation can be promoted by getting close to the customers
- Assess Prospects and Needs in Selected Markets
- Concentrate on more Erosion Control Products

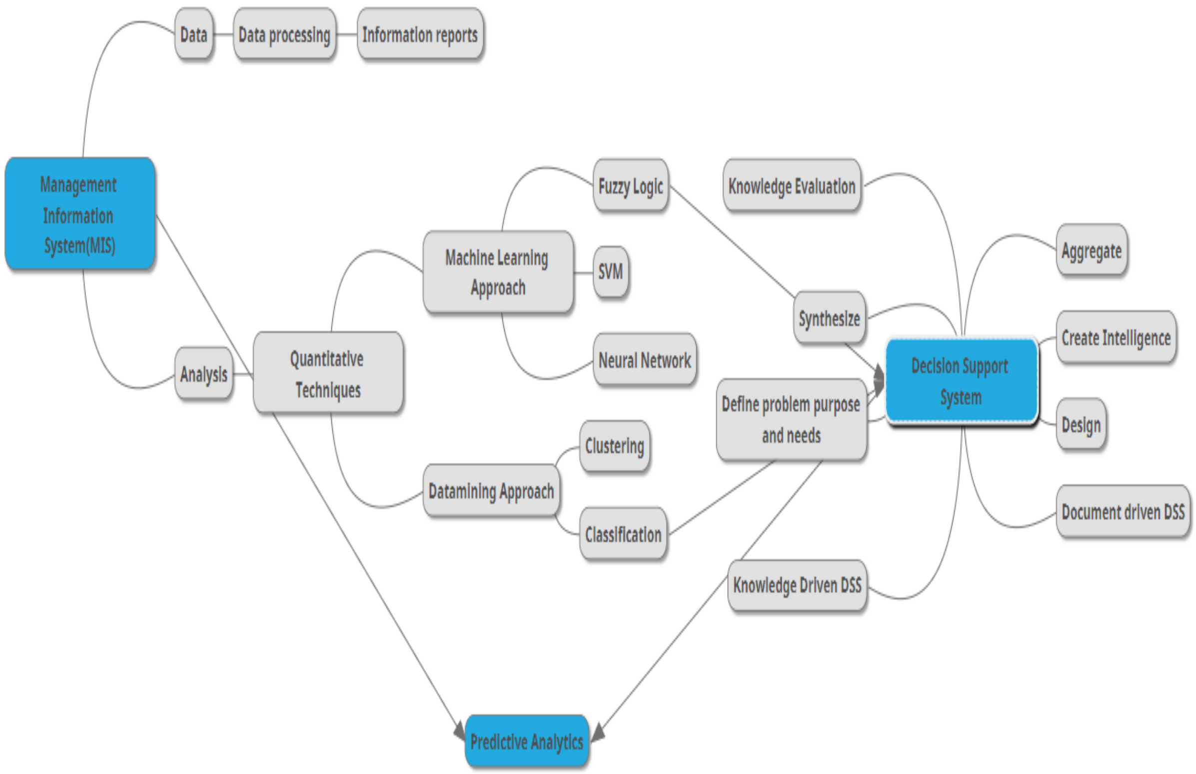


Fig. 1. Work flow of Predictive Analytics

IV. Decision Support System (DSS)

Decision support system is used for making decisions in ambiguous and uncertain conditions. Even though many techniques are available for predictive analysis, decision tree is the commonly used system for making decisions in a predicted manner. Decision Support System uses a machine learning algorithm for making decisions [11]. It provides a solution for non-structured problems. Several decision making algorithms are there such as ID3, CART, and C4.5. Here C4.5 algorithm is used for the decision making to establish the coir industry .Here the prediction is performed based on the classification label. In decision tree based method, the training data is categorized into different groups according to the classification label. Here the prediction is performed by using a top down approach known as fuzzy C4.5 algorithm. Statistical classification algorithm is used in this hybrid approach.

A decision support system (DSS) is an automated computerized statistics system used to aid in making decisions in a commercial enterprise or in business. A DSS lets users to examine big piles of data and assemble records that may be used to remedy issues and make better decisions.

The advantages of decision assist structures encompass greater informed choice making, well timed problem-fixing and advanced efficiency for coping with troubles with unexpectedly changing variables.

Decision making process is mostly based on the interactive decision support system with hierarchical structures. It can be categorized into two such as document driven DSS and Knowledge driven DSS. The document driven DSS manages retrieves and manipulate the unstructured information in a variety of electronic formats whereas the knowledge driven system provides specific problem solving expertise stored as facts, rules and procedures. The basic building blocks of DSS consist of the following units such as data bases, query, model creation and execution and statistics and forecasting. Here we combine the two decision support system such as machine learning approach and data mining approach.

Fig.1 indicates the working flow of the proposed predictive analytics. A database system acts as a bank for the decision support system. It fills the information in to the bank and stores the large amount of information that is relevant to the class for the Decision Support System. It provides a logical data structure that has been outlined rather than physical data structures and the user can interact with. The database should be capable of providing information to the user regarding the type of the available data and how to access the data.

We were provided with a training dataset consisting of information about rubberized coir mattress. This data was in the form of a CSV or spreadsheet and had details such as density, IHI, VariAging,flexing,compset aged, compset unaged,pH value, chloride content and sulphate content. For simplicity of performing data mining tasks, the information was filled into a MySQL database.

V. Data Preprocessing and Quality Issues

Formerly we consumed details; we then segmented the training dataset further, considering numerous viable splitting attributes, i.e. the attributes which could have a better impact at the performance. Over a time of 4 many years the commercial enterprise recorded an extensive alternate in best viewpoints.

Rubberized Coir is an organic product. Coconut fiber is the important and basic raw material that going into the manufacturing process of rubberized coir. In that capacity, this item is bio-degradable as well as ecofriendly. This is the primary factor to support its, particularly when contrasted with its competitors such as PVC and U Foam and other synthetic products. In the present day when condition and environmental conditions are the top most in each perceiving individual's mind. This fact increases its production and propagates it everywhere.

1) The raw material of rubberized mattress is a bye-product from coconut husk. It is discarded as a waste and also can be used as a material for door mats and floor coverings. Thus the use of rubberized coir is actually boosted as a utility product by which our body can rest and recline. In view of this the price for the raw fiber is increased substantially and an innovative technology has started the production of the rubberized coir products. This will become a benefit to the poor farmers.

2) There are a lot of troubles like excessive cost of the transportation of the raw substances and the coir merchandise throughout the international locations and nation. Additionally the enterprise face several issues within the maintenance of the in- house power plant because if the production as soon as commenced it cannot be stopped in among as the industry prospers moderately well

3) Our present efforts in the Microbiological and chemical Research & development area need to be focused to accomplish it lighter and adaptable to accomplish it complete absorptive or insulating, and to accomplish fire resistant. All R & D efforts desire abundant monetary inputs and the industry looks advanced and quickly to Coir Board for their assistance and guidance[12].

VI. Problem Statement

Generally, to establish an industry in a village, generally we need to consider several parameters such as the availability of raw material, power consumed by the industry, the water required for the processing. Out of these the availability and the quality of the raw material is the very important factor. There are several choices for identifying these requirements. One is the entrepreneur has to gather all the details about these parameters and has to visit all the villages. But it is time consuming and cost effective. He has to take few opinions from the domain experts and it may lead to confusion for making a decision or lead to a false decision.

VII. Proposed System

The existing system can be made better by replacing the domain experts. The total time taken can also be reduced by the industrialist by gathering and analyzing the records. It replaces the expert by means of decision support system. The decision support system deal with the selection of the raw material in view of the numerous attributes(Jayasekara & Amarasinghe, 2010). Additionally it makes which sort of the raw material is suitable for which kind of product. This decision is primarily based on C4.5 algorithm. The C4.5 algorithm can build a knowledge level from the industry facts bases and might proceed for the decision making. In fuzzy C4.5 algorithm15], at each node of the decision tree, one of the attribute which is the most effective one of the data set is chosen for splitting it into samples S. These samples are classified into one class or another based on the splitting condition. The splitting condition is based on the information gain of the attribute. This attribute has the highest information gain and that is considered as the criteria for making decisions [13]. Thus we can standardize the data. This process continues up to get the final class choice.

The data base contains information about the parameters of the raw material such as density, IHI, VariAging, flexing, compset aged, compset unaged, pH value, chloride content and sulphate content. By using FuzzyC4.5 algorithm we can predict which type of the raw material is suitable for which type of the rubberized mattress. This decision support system is a classification and prediction technique and has been applied to the training data set to build a suitable and accurate knowledge base.

While the industrialist needs a selection, at first he need to fill the input form, that is provided by the system. The Decision Support System tests the input data based on knowledge base and generates class label for that entry. Primarily based on that class label, it offers the output to industrialist as suitable or no longer appropriate

A. C4.5 Algorithm

C4.5 algorithm is used for prediction and the main working is based on the decision tree approach. Branch nodes and leaf nodes are the main concern in a decision tree approach. The branches in the tree represent the attribute choice and the class label is the leaf node and it makes a decision. The action can be start from the root node and the algorithm is applied recursively on each node for constructing the decision tree. By analyzing each branch we will get the idea of the scenario and the output. This is the basic idea of the C4.5 algorithm. It is a top down and greedy approach and can done a search through the given set of data and can make a classification. For classifying and predicting the class labels a statistical property called information gain is used here[14].

Decision trees are supervised learning method and the aim is to divide the data set into homogeneous groups in terms of the class label and the splitting attributes. The decision is done from root to the leaf.

B. Basic Terms

The C4.5 algorithm defines the amount of information provided and is known as information gain. For calculating the gain first we have to find out the entropy. The entropy and the information gain can be calculated as[15]

$$Entropy, Pi = \sum_{i=1}^{n} - Pi * \log(Pi) \quad (1)$$

$$Information\ gain(I) = Pi - \sum_{k=1}^{n} -Pk * \log(Pk) \quad (2)$$

Where Pi is the set of all possible values for the Attribute A. This information gain is used for ranking the attributes in the given data set and has to build the decision tree based on this ranking.

VIII. Results And Discussion

A. Fuzzy C4.5 Algorithm

In Fuzzy C4.5 algorithm, it selects the attribute to be tested based on the information gain which is designed by the likelihood of membership values for statistics in a node. The basic steps to generate a fuzzy decision tree are described here:

1) Make a node which represents the root of the tree and is having the fuzzy membership value as 1 for all training data.

2) The leaf node has to be created iteratively if the element which represents data in the fuzzy set fulfills any of the subsequent conditions:

a) Set a given threshold value and the data set having the number of is less than a given threshold.

b) The percentage of objects which represents the data set having a value greater than or equal to the given threshold.

c) If there are no other attributes are presented for classification.

3) The leaf node have to be assigned a name either the name which represents the class or the class names with value of the membership.

4) If the node does not satisfy any of the above conditions then do the following:

a) For all attributes, calculate the information gain; select the attribute that has greatest information gain as the test attribute.

b) Divide the fuzzy data set at the node into fuzzy subsets using the test attribute selected in a, with the membership value of an object in a subset set to the product of the membership value in parent node dataset and the value of the selected attribute fuzzy term in the parent node dataset.

c) For each of the subsets, generate new node with the branch labeled with the fuzzy term of the selected attribute.

5) Recursively repeats the steps from 2 for newly generated node.

The flowchart of the fuzzy-c4.5 algorithm is shown in fig.2

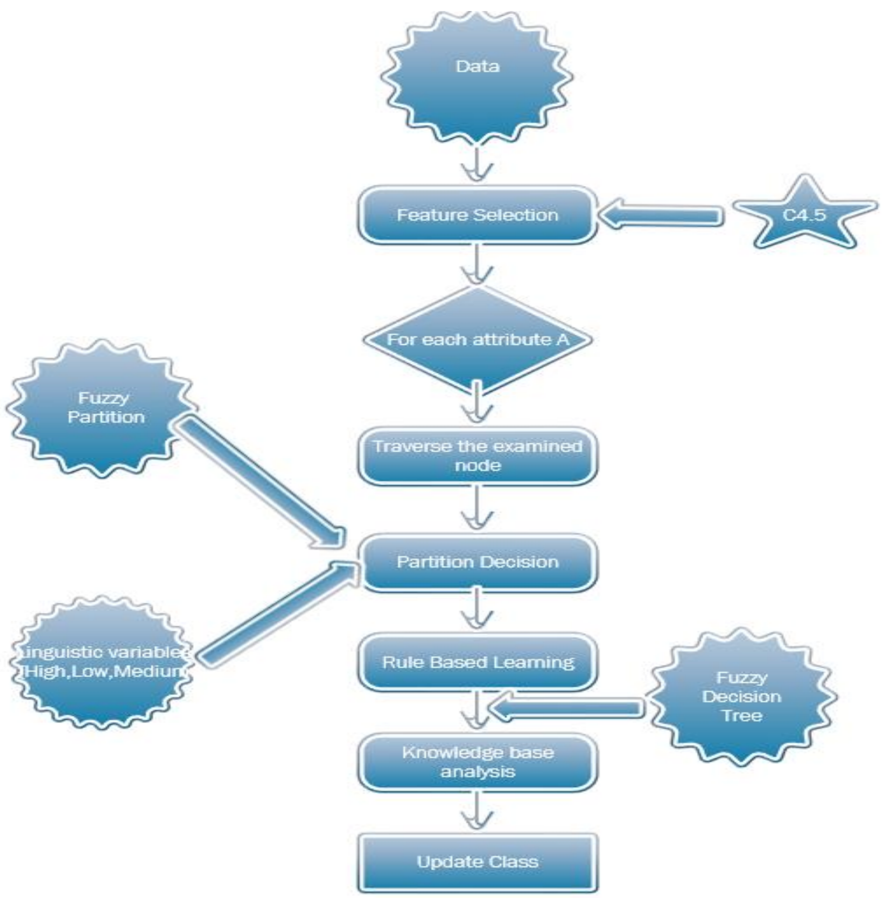


Fig. 2. Flow chart for the FuzzyC4.5

B. Comparison with Several Algorithms

Accuracy: "By the term 'accuracy', we mean the degree of compliance with the standard measurement, i.e. to which extent the actual measurement is close to the standard one". Here the Table 1 shows the overview of the decision tree algorithms ID3,C4.5 and Fuzzy-C4.5. The Table 2 presents a performance comparison of Fuzzy-C4.5 and Other algorithms in terms of accuracy with the data set.

TABLE I. OVERVIEW OF DECISION TREE ALGORITHMS

Algorithm	Splitting Criteria	Attribute Type	Pruning method	Missing Values	Outlier detection
ID3	Information Gain	Handles categorical value only	No pruning	Do not handle missing Values	Susceptible
C4.5	Gain Ratio	Handles categorical value and numerical value	Error based pruning	handle missing Values	Can handle
FuzzyC45	Membership value	Handles categorical value and numerical value	Cost based pruning	handle missing Values	Can handle

TABLE II. PERFORMANCE COMPARISON

Algorithm	Time taken to build the model	Accuracy
FuzzyC4.5	0.02s	96.7224
C4.5	0.02 s	94.6667
ID3	0.02 s	93.9128
REPtree(CART)	0.03 s	92.7083
Linear NN search	0.01 s	58.3333
Multilayer Perceptron	0.27 s	79.1667
NaiveBayes	0.28 s	80.2083

In Fuzzy C4.5 Algorithm the classification of the output is we can predict "which type of firm "based on the attributes density, IHI, VariAging, flexing, compset aged, compset unaged, pH value, chloride content and sulphate content. Here only a sample set of values is shown in the Table 3. The original data set contains 1000 values. The C4.5 algorithm differs in several respects from other prediction algorithms

The primary advantages of FuzzyC4.5 algorithm can be stated as

ID3 algorithm selects the best attribute based on the concept of entropy and information gain for developing the tree. FuzzyC4.5 algorithm acts similar to ID3 but improves a few of ID3 behaviors:

- It can deal with the training dataset with missing values of the attributes
- It could cope with the opposing cost attributes
- It prunes the selection tree virtually after its creation
- It could cope with the attributes with discrete and continuous value sets
- speed of the C4.5 algorithm is considerably quicker as compared with the alternative algorithms.
- The FuzzyC4.5 set of rules offers greatly improved scalability of each decision trees and the rule sets. The scalability is more suitable by means of using multi-threading and efficient pruning.
- The FuzzyC4.5 algorithm outfits the classifiers for prediction and thereby advances the accuracy.

IX. IMPLEMENTATION RESULT

The original dataset used for the testing and implementation result is from the Central Coir Research Institute, Alappuzha. Hence the data set is highly confidential here only a sample data set is provided. The dataset is tested with java platform and got the decision tree as shown in figure 3.

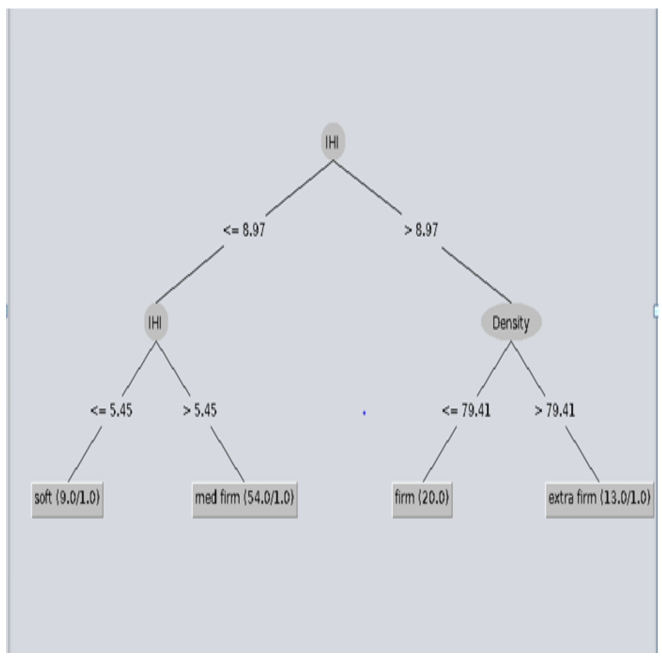


Fig. 3. Decision tree created using Fuzzy C4.5

X. CONCLUSION

At present, decision trees are extensively used in many applications inclusive of medical field, business analysis, diagnosis, cognitive technology, artificial intelligence, engineering and information mining. The class accuracy of selection trees has been a subject of numerous researches. In this paper, we advocate an efficient Fuzzy C4.5 algorithm for data mining. This method provides the performance and the comprehensibility that are crucial to data mining.

TABLE III. SAMPLE DATASET VALUES

SI NO:	Density	IHI	Var aging%	flexing%	compset%(aged)	compset unaged%	PH value	chloride content%	sulphate content%	Grade
1	78.65	10.55	-4.36	-2.75	14.44	5.56	6.4	0.08	0.04	Firm
2	84.76	12.58	-7.72	-4.69	15.84	7.14	6.4	0.07	0.09	extra firm
3	71.71	10.63	-9.54	-6.22	16.67	12.6	6.45	0.02	0.02	Firm
4	64.81	7.45	-16.43	15	17.87	12.5	6.5	0.08	0.04	med firm
5	68.23	7.99	-6.09	-8.7	8.18	4.55	6.5	0.02	0.02	med firm
6	77.71	11.8	-6.15	-6.33	13.57	5.5	6.5	0.01	0.01	Firm
7	94.75	14.73	-5.52	-4.44	7.73	5.84	6.45	0.08	0.04	extra firm
8	68.53	8.9	1.12	1.68	7.72	2.85	6.5	0.08	0.04	med firm
9	66.94	7.22	-14.9	-15.27	14.36	10.91	6.5	0.1	Nil	med firm
10	68.52	8.7	-15.49	16.36	16.47	13.52	6.5	0.1	Nil	med firm
11	76.9	10.65	-3.56	19.44	18.52	10.28	6.45	0.08	0.01	Firm
12	88.57	12.8	-4.51	13.06	19.9	9.81	6.5	0.04	0.02	extra firm
13	68.39	8.65	-15.88	11.58	19.81	14.15	6.53	0.18	0.08	med firm
14	68.07	8.63	-9.64	-1.12	19.52	10.9	6.4	0.16	0.095	med firm
15	68.49	8.7	15.7	-1.69	12.14	9.43	6.2	0.22	0.1	med firm
16	48.75	3.97	-18.69	-15.96	15.73	13.37	5.9	0.25	0.08	Soft

Here we used a data set from the CCRI for the classification and prediction purpose and got 96% accuracy in the results. In particular, the method permits us to correctly classify patterns of decision boundaries well, that is difficult to do the use of attribute based classification strategies. The rubberized mattresses are classified here based on the attribute values. The experiment outcomes show that our technique is greater efficient in overall performance and clarity of rules in comparison with other techniques which include C4.5. Our initial experiments have proven that the algorithm is strong for many real world applications. Future work is to expand a fuzzy decision tree pruning algorithm to generate greater compact deep learning methods. We also plan for improving the algorithm a good way to manage a large volume of data efficiently, which is often an important requirement to data mining.

REFERENCES

[1] Dasanayaka, S. W. S. B., & Sardana, G. D. (2010). Development of SMEs through clusters: a comparative study of India, Pakistan and Sri Lanka. *World Review of Entrepreneurship, Management and Sustainable Development*, *6*(1–2), 50–70

[2] Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., & Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, *17*(2), 76–81.

[3] Turban, E., Sharda, R., & Delen, D. (2011). *Decision support and business intelligence systems*. Pearson Education India

[4] Thomas, S., Woh, Y., Wang, R., & Lim, K. (2017). ScienceDirect ScienceDirect Probing the hydrophilicity of coir fibres : analysis of the mechanical properties of single coir fibres. Procedia Engineering, 200, 206–212. https://doi.org/10.1016/j.proeng.2017.07.030

[5] Amarnath, J. S., & Brindha, M. (n.d.). An Investment and Export Analysis of Coir Enterprise in Tamil Nadu, India.

[6] Dasanayaka, S., Sardana, D. G., & others. (2015). Development of Small and Medium Enterprises through Clusters and Networking: A Comparative Study of India, Pakistan and Sri Lanka. *International Journal of Economics and Business Administration*, *3*(2), 84–108.

[7] Fernandez, C. (2003). STRATEGIC ALLIANCE FOR THE DEVELOPMENT OF COIR INDUSTRY IN THE APCC COUNTRIES AND DEMAND PROSPECTS FOR COIR PRODUCTS IN ECOFRIENDLY APPLICATIONS.

[8] Abu-halaweh, N. M., & Harrison, R. W. (2009). Practical Fuzzy Decision Trees

[9] Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010

[10] Wilson, C. (2007). Flow resistance models for flexible submerged vegetation. *Journal of Hydrology*, *342*(3–4), 213–222

[11] Sierra-Pérez, J., Boschmonart-Rives, J., & Gabarrell, X. (2015). Production and trade analysis in the Iberian cork sector: Economic characterization of a forest industry. *Resources, Conservation and Recycling*, *98*, 55–66. https://doi.org/10.1016/j.resconrec.2015.02.011

[12] Turban, E., Sharda, R., & Delen, D. (2011). *Decision support and business intelligence systems*. Pearson Education India

[13] Prasad, M. (1996). Physical, chemical and biological properties of coir dust. In *International Symposium Growing Media and Plant Nutrition in Horticulture 450* (pp. 21–30)

[14] Jayasekara, C., & Amarasinghe, N. (2010). Coir--Coconut Cultivation, Extraction and Processing of Coir. *Industrial Applications of Natural Fibres: Structure, Properties and Technical Applications*, 197–217.

[15] Kim, M. W., Lee, J. G., & Min, C. (1999). Efficient Fuzzy Rule Generation Based on Fuzzy Decision Tree for Data Mining

Multivariate Copula Modeling with Application in Software Project Management and Information Systems

Syed Muhammad Aqil Burney[1]
Dept. of Actuarial Science Risk Management
Institute of Business Management,
Karachi, Pakistan

Osama Ajaz[2]
Dept. of Sciences
FAST-National University,
Karachi, Pakistan

Shamaila Burney[3]
Dept. of Mgmt. Science
Hamdard University,
Karachi, Pakistan

***Abstract*—This paper discusses application of copulas in software project management and information systems. Successful software projects depend on accurate estimation of software development schedule. In this research, three major risk factors and their impact on software development schedule are considered. Software development schedule is calculated by COCOMO-II model. Two models are simulated 100000 times, model-I considered dependence among risk factors by T-copula and model-II considered risk factors independent. The comparison of the two risk models revealed that model-II always underestimate the software development schedule while model-I evaluated the software schedule risk accurately. Therefore it is necessary for software development experts to consider dependence among various risk factors. R-package copula is employed to implement the algorithm for multivariate T-copula. Multiplier goodness-of-fit test shows that T-copula is good choice for characterization of dependence among three risk factors.**

KeyWords—T-Copula; COCOMO – II; software development schedule; risk analysis

I. Introduction

Software project management and information system consists of multiple activities under the umbrella of software engineering [1]. It involves planning, scheduling, budgeting and managing entire software development process. Each activity in a project management consumes time i.e. software project schedule. Further, entire software development life cycle depends on accurate estimation of software project schedule. Beside estimation of project schedule, it is the responsibility of a project manager to identify risk factors that results in project delays or failure. Therefore it is necessary to assess software project schedule accurately. Plethora of literatures is available on risk factors that results in schedule overrun or project failure. For further detail, see [2] [3] [4].

Generally software project manager just estimate project schedule and completely or partially ignore the impact of risks on estimated project schedule, even if project manager consider some risk factors; it is assumed to be independent. Positive and negative dependence may affect software risk severely. For example, it is possible that project manager inaccurately estimate the software schedule to 20 months and project ends in 26 months. This is also possible that one risk may cause another risk to happen i.e. increase in customer requirements during software development duration strongly cause schedule overrun and in turn may lead to loss of key employee. Further, project delays have negative impact on customer satisfaction which results in bad reputation of an organization. This research paper analyzes software project schedule and consider dependence among risk factors by means of copula.

The reason to consider copulas in this research is that most of literatures available on theory of copulas discuss application in Econometrics, Finance and in Insurance [5]. Further, very few articles available that discuses copulas in software project management or in information systems. See [6] [7] [8] as examples and the references therein.

The remaining paper is organize as follows: Section 2 gives an overview to the theory of copulas. Section 3 discusses software schedule and associated risks. Section 4 discusses application of copula model for software houses based in Karachi and finally this research end up with Section 5 that discusses overall conclusion.

II. Theory of Copulas

The discovery of copula is associated to the seminal research work of Sklar [9]. He derived the word “copula” for multivariate joint distribution that links to its marginal distribution. In this section, we present brief overview to the basic theory of copulas for higher dimensions. For further details and proofs about copulas and its historical development, please see [10] [11] [12] [13] and the references therein.

A. Multidimensional Copulas

A d-dimensional copula or d-copula is a function “C” from I^d to I where I ϵ [0, 1], if and only if it satisfies the following conditions:

(i) For every m in Id, C (m1, m2… md) = 0 if at least one of m is equal to zero.

(ii) If all m’s are set to one then for every k ϵ {1, 2… d}, C (m) = m_k.

(iii) C is d-increasing, for all ($a_1, a_2, \ldots, a_d$), ($b_1, b_2, \ldots, b_d$) in I^d such that $a_i \leq b_i$,

V_c ([**a**, **b**]) ≥ 0

$$\sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\cdots\sum_{i_d=1}^{2}(-1)^{i_1+i_2+\cdots+i_k}C(m_{1_{i_1}}, m_{2_{i_2}}, \ldots, m_{d_{i_d}}) \geq 0$$

All above properties of d-copula are multivariate version of bivariate copula. Property (i) defines grounded condition, property (ii) defines copula margin if (d – 1) variables are known, and property (iii) defines C volume of rectangle [a x b] for any d-dimension. Further, it is easy to show that, any convex linear combination of copulas is a copula i.e. $\sum_{i=1}^{d}\alpha_i C_i$ for all $\alpha_i > 0$ and $\Sigma\alpha_i=1$ [13]. Nelsen and Joe discusses some other important properties of copulas that we state as following theorems 1, 2, and 3 below [11] [13].

Theorem 1

For every d-copula, the Frechet-Hoeffding bound inequality is given by:

$$C_L^d \leq C(m_1, m_2, \ldots, m_d) \geq C_U^d \quad (1)$$

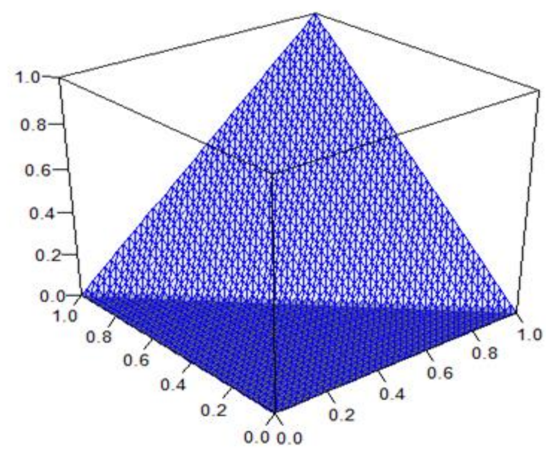


Fig. 1. Figure 1: Graphs of Frechet – Hoeffding Lower Bound Copula.

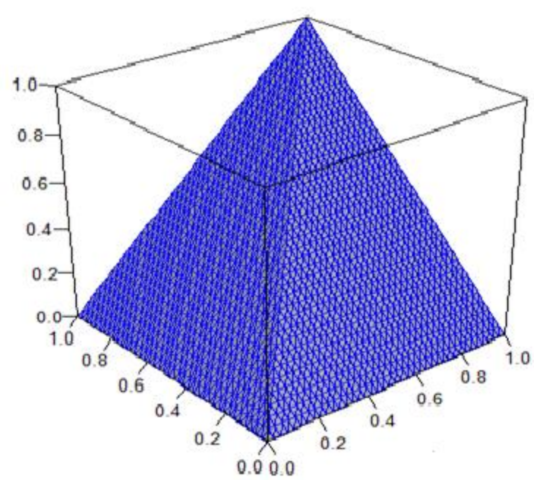


Fig. 2. Graph of Frechet-Hoeffding Upper Bound Copula.

Where C_L^d and C_U^d represents Frechet – Hoeffding lower and upper bounds, these bounds define as; $\max(\sum_{i=1}^{d} m_i - d + 1, 0)$ and $\min(m_1, m_2, \ldots, m_d)$ respectively. Notice that, for d > 2, the lower bound C_L^d is not a copula. For further details see theorem 3.6 in [13]. For d = 2, the graphs of C_L^d and C_U^d are given in figure 1 and figure 2 below respectively.

Theorem 2

For all m in I, the copula is independent if and only if,

$$C^d(m) = m_1 . m_2 . \ldots m_d \quad (2)$$

For all $d \geq 2$, the continuous random variables $X_1, X_2 \ldots X_d$ are independent if and only if their d-copula is C^d (m). For d = 2, the graph of independent copula is given as figure 3:

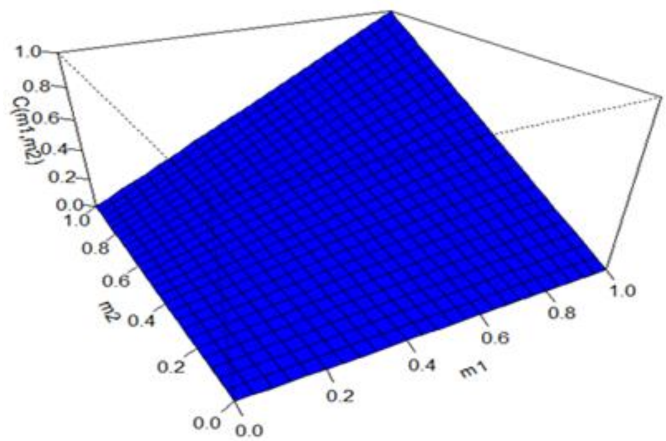


Fig. 3. Graph of Independent Copula.

Theorem 3

For all **m**, **n** in I^d, the d-copula satisfies Lipchitz condition,

$$|C(\mathbf{m}) - C(\mathbf{n})| \leq \sum_{i=1}^{d}|m_i - n_i| \quad (3)$$

Thus d-copula is a continuous function from I^d to I.

Example 1

As an illustration, consider the following extended FGM copula for d = 3,

$C^3(m_1, m_2, m_3) = m_1 m_2 m_3 [1 + \alpha_{12}(1 - m_1)(1 - m_2) + \alpha_{13}(1 - m_1)(1 - m_3)$

$+ \alpha_{23}(1 - m_2)(1 - m_3) + \alpha_{123}(1 - m_1)(1 - m_2)(1 - m_3)]$ (4)

It is easy to show that, the above three dimensional FGM copula satisfies the basic requirements of d-copula.

Hence (4) is a copula, theorem 2 hold for d = 3 if dependence parameters α_{ij} equals to zero. For further details about d-dimensional extended FGM copula, see Drouet and Kotz [14].

B. Sklar's Theorem

Copulas are important because of Sklar's theorem [9]. According to this theorem, any multivariate joint distribution can be represented by its marginal distribution. Consider $X_1, X_2 \ldots X_d$ be continuous random variables with their joint distribution function J and univariate marginal distribution $F_i(X_i) = P(X_i \leq x_i)$, $i \in \{1, 2\ldots d\}$. Then, there exists a d-copula C^d such that,

$$J(x_1, x_2\ldots x_d) = C^d(F_1(x_1), F_2(x_2)\ldots F_d(x_d)) \quad (5)$$

For all $x \in R^d$, the function C^d is uniquely determined, if $F_i(X_i)$ are continuous. Otherwise C^d is uniquely determined on (range of $F_1(x_1)$) x (range of $F_2(x_2)$) x … x (range of $F_d(x_d)$).

Let J, C^d and $F_i(X_i)$ be as in above (5), let $F_1^{(-1)}$, $F_2^{(-1)}$, …, $F_d^{(-1)}$ are quasi-inverses of $F_i(X_i)$ where $i \in \{1, 2, \ldots, d\}$, respectively. Then,

$$C(m_1, m_2\ldots m_d) = J(F_1^{(-1)}(m_1), F_2^{(-1)}(m_2)\ldots F_d^{(-1)}(m_d)) \quad (6)$$

for all $m \in I^d$.

For the sake of simplicity, let F_1, $F_2\ldots F_d$ be continuous and differentiable distribution functions. Then the corresponding density function to (6) is

$$j(x) = c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \prod_{i=1}^{d} j_d(x_d) \quad (7)$$

Where $j_d(x_d)$ is the marginal density of $F_i(X_i)$, $i \in \{1, 2\ldots d\}$, and

$$c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) = \frac{\partial C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d))}{\partial F_1(x_1), \partial F_2(x_2), \ldots, \partial F_d(x_d))} \quad (8)$$

According to (5), the copula function C^d is uniquely determined and can also be represented as,

$$C(m_1, m_2, \ldots, m_d) = \int_0^{m_1}\int_0^{m_2} \ldots \int_0^{m_d} c\,(n_1, n_2, \ldots, n_d)\,\partial n_1.\partial n_2 \ldots \partial n_d \quad (9)$$

Where $m_i = F_i(x_i)$ and corresponding density is c, then we have joint density function (7).

III. Software Development Schedule & Associated Risk Factors

A. Project Scheduling Risk

Risk is related to future happenings and it has two characteristics one is *Uncertainty* and other is *Losses* [1] [15]. If the risk is certain to occur then it has positive or negative impact on projects' objectives. In all, risk has two dimensions: probability of occurrence of event and its impact. If risk associated to software project schedule is certain to occur then estimated schedule exceeds deadline which results in financial losses and bad reputation of an organization. This research explores the relationship between risk and its impact on software project schedule. Scheduling risk is the probability of one or more events, if they occur has positive or negative impact on software development duration [16].

There are many uncertain risk factors that affect project schedule severely. However, for this research, three major risk factors are considered that every project manager must face [4]. These three major risks are defined below as:

a) Imprecise measurement of software effort:

Software effort is defined as number of working hours spend on the project. In this research, software effort is expressed as:

$$Software\ Effort = \left|\frac{Actual\ Effort - Estimated\ Effort}{Estimated\ Effort}\right| \times 10 \quad (10)$$

Incorrect measurement of software effort results in project delays. Incorrect estimation of effort is consequences of lack of experienced manager or inadequate knowledge about estimation tools.

b) Loss of Employees during project:

It is the usual turnover rate of employees during projects. It express simply in percentages. It includes resignations, death, medical leave, retirement or transfer of employees during project.

c) Change of Customer requirements:

Change of customer requirements includes Increment or decrement of customer requirements during software development duration and expressed as percentages.

B. Risk Assessment Model

Many risk assessment techniques exists. I will assess software development risk by risk assessment model. For this research, the following schedule risk model is considered:

$$\text{Schedule Risk} = (\text{Project duration}) \times R_1 \times R_2 \times R_3 \quad (11)$$

Where R_1 indicates impact of imprecise estimation of software effort, R_2 indicates impact of risk of Loss of Employees during project and R_3 indicates impact of Change of customer requirements during software development period on project schedule. As can be seen, the model is multiplicative in nature. For further details about this model, see [16] [17]. The probability distributions for these three risk factors are derived by using expert data from various software houses based in Karachi. Project cost and schedule is estimated by COCOMO-II model [18].

C. COCOMO – II

Boehm et al. [18] proposed a COCOMO-II model and it requires three stages to estimate software project cost, effort and schedule. For early stages of software project development, application composite model is used. When information packages, software architecture and infrastructure is finalized, early design model is used. Finally, post architecture model is used during software development duration.

We have calculated new software project schedule for 400000 SLOC using COCOMO – II. By setting variables to desire level, the COCOMO-II model estimated new software development schedule from 17 months to 40 months. Therefore, the new software development project can be completed at least in 17 months and at most 40 months while the actual schedule for this project was 30 months. This minimum and maximum duration represents best and worst case. The levels of COCOMO-II model for new software development schedule is set by consulting software development experts.

D. T – Copula Method to Model Dependence

Let n-dimensional random vector $Y = (Y_1, Y_2, \ldots, Y_d)^T$ has d-variate t-distribution with υ degrees of freedom, μ mean vector and correlation matrix R and its joint probability density function (PDF) is defined by:

$$f(y) = \int_{-\infty}^{y_1}\int_{-\infty}^{y_2}\cdots\int_{-\infty}^{y_d} \frac{\Gamma\left(\frac{\upsilon+d}{2}\right)}{(\pi\upsilon)^{\frac{d}{2}}\Gamma\left(\frac{\upsilon}{2}\right)|R|^{\frac{1}{2}}}\left[1+\frac{1}{\upsilon}(X-\mu)^T R^{-1}(X-\mu)\right]^{-\frac{\upsilon+d}{2}} \quad (12)$$

The multivariate t-copula is then defined as:

$$C(m_1, m_2, \dots, m_d) = T_{R,\upsilon}\left(T_\upsilon^{-1}(m_1), T_\upsilon^{-1}(m_2), \dots, T_\upsilon^{-1}(m_d)\right) \quad (13)$$

$$= \int_{-\infty}^{T_\upsilon^{-1}(m_1)}\int_{-\infty}^{T_\upsilon^{-1}(m_2)}\cdots\int_{-\infty}^{T_\upsilon^{-1}(m_d)} \frac{\Gamma\left(\frac{\upsilon+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)(\upsilon\pi)^{\frac{d}{2}}|R|^{\frac{1}{2}}}\left(1+\frac{1}{\upsilon}X^T R^{-1}X\right)^{-\frac{\upsilon+d}{2}} d_{x_1}\dots d_{x_d} \quad (14)$$

Where $T_\upsilon^{-1}(u_i)$ is the inverse of the univariate distribution functions of t-distribution with υ degrees of freedom. The corresponding canonical representation of multivariate T-copula density is given by:

$$c(m_1, m_2, \dots, m_d) = |R|^{-\frac{1}{2}}\frac{\Gamma\left(\frac{\upsilon+d}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right)}\left(\frac{\Gamma\left(\frac{\upsilon}{2}\right)}{\Gamma\left(\frac{\upsilon+1}{2}\right)}\right)^d \frac{\left(1+\frac{1}{\upsilon}\delta' R^{-1}\delta\right)^{-\frac{\upsilon+d}{2}}}{\prod_{i=1}^{n}\left(1+\frac{\delta_i^2}{\upsilon}\right)^{-\frac{\upsilon+1}{2}}} \quad (15)$$

Where $\delta_i = T_\upsilon^{-1}(m_i)$.

Lot of studies has shown that T-copula is superior to any other copula such as Gaussian copula [19] [20]. It is symmetric but has tail dependence. We have modeled dependence among three risk factors by T-copula. Inference function for margins (IFM) method is used to estimate the parameters of T-copula. In the figure 4 below, contour plot for multivariate T-copula in two dimensions is given.

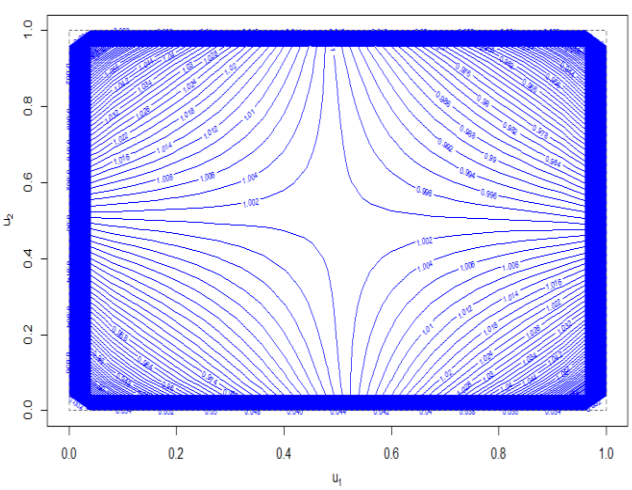


Fig. 4. Contour Plot for Multivariate T-Copula Density.

The contour plot shows moderate dependencies between variables. This is the beauty of copula that even if the correlation is zero nevertheless marginal distributions are related to joint distribution [9].

IV. Application of Copulas in Software Project Management

In this research, two risk models are considered for new software project schedule. Model-I considered dependence among risk factors and Model-II assumes that risks are independent. For model –I, we have used multivariate T-Copula to model dependence among risk factors. Monte Carlo method is employed to simulate the two models. Simulations for both model executed 100000 times.

A. Distributions of Three Risk Factors

The results of copulas are useful if and only if we have fitted best distributions to marginal distributions. These marginal distributions are for Software Effort, Loss of employees and change of customer requirements during project. The marginal distributions for the three risks are derived using data of several software houses based in Karachi. Several statistical hypothesis testing tools and graphs are employed to assess goodness of fit for three marginal distributions. The results for goodness-of-fit tests for sample size 200 are provided in the table 1, 2 and 3.

TABLE I. Probability Distribution for Software Effort (R_1), H_o: R_1 conforms to Normal Distribution

Test	Test Statistic	p-value
Chi-Square test	0.52	0.9144
K-S test	0.053495	0.6162
Anderson & Darling Test	0.50791	0.7386
Cramer-Von Mises Test	0.06606	0.7765

All goodness-of-fit tests listed in above table 1 provide high p-values at 1% level of significance. Therefore, the risk distribution for imprecise measurement of software effort conforms to normal with parameters (40.49525%, 20.79036%)

TABLE II. Probability Distribution for Loss of Employees (R_2), H_o: R_2 conforms to Weibull Distribution

Test	Test Statistic	p-value
Chi-Square test	11.41	0.1795323
K-S test	0.08365265	0.1217015
Anderson & Darling Test	1.6798	0.1389
Cramer-Von Mises Test	0.2426781	0.1980663

All goodness-of-fit tests listed in the above table provide high p-values at 1% level of significance. Therefore, risk distribution for loss of employees during project conforms to Weibull with parameters (1.478794%, 27.747935%).

TABLE III. PROBABILITY DISTRIBUTION FOR CHANGE OF REQUIREMENTS (R_3), H_0: R_3 CONFORMS TO WEIBULL DISTRIBUTION

Test	**Test Statistics**	**p-value**
Chi-Square	2.8	0.5918327
K-S Test	0.06114199	0.4432991
Anderson & Darling Test	0.65045	0.6014
Cramer-Von Mises Test	0.08189777	0.6810586

All goodness-of-fit tests listed in the above table3 provide high p-values at 1% level of significance. Therefore, risk distribution for change of customer requirements conforms to Weibull distribution with parameters (1.56949%, 27.27296%).

B. Simulation Results of the Two Models

As described above, two risks models are considered. Model – I considered dependence among three risk factors by T-copula and model – II considered risk factors are independent. Both the models are simulated 100000 times and their histograms are presented in the fig. 5 and fig. 6 respectively.

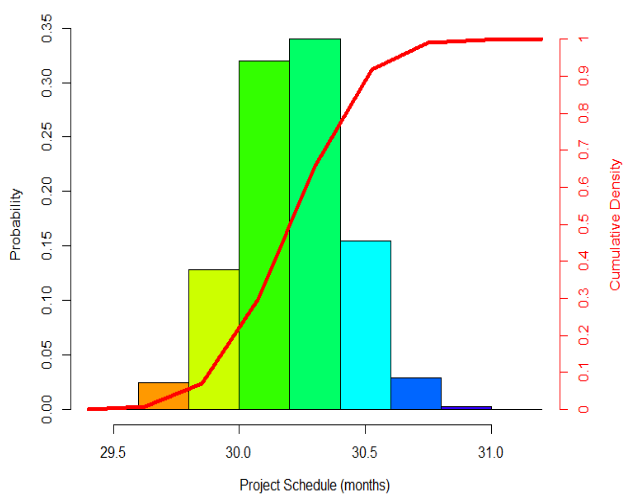


Fig. 5. Simulation Histogram for Model – I.

In the figure 5 above, the left y-axis represents probability density function and right y-axis represents cumulative density function for the simulated model-I. According to the above simulated histogram, the new software project schedule can vary from 29 months to 31 months. There is 100% chance that, the new software project can be completed in almost 30 months while there is approximately 38% chance that the new project can be completed in almost 29 months.

Multiplier goodness of fit test [21] is applied to the chosen T-copula. The goodness of fit provide test statistics S_n^k = 0.036957 with p-value 0.1004. Since p-value is high enough at 1% level of significance. Therefore the chosen T-copula model is appropriate for characterization dependence among three risk factors. Simulation histogram for model – II is shown below in the figure 6.

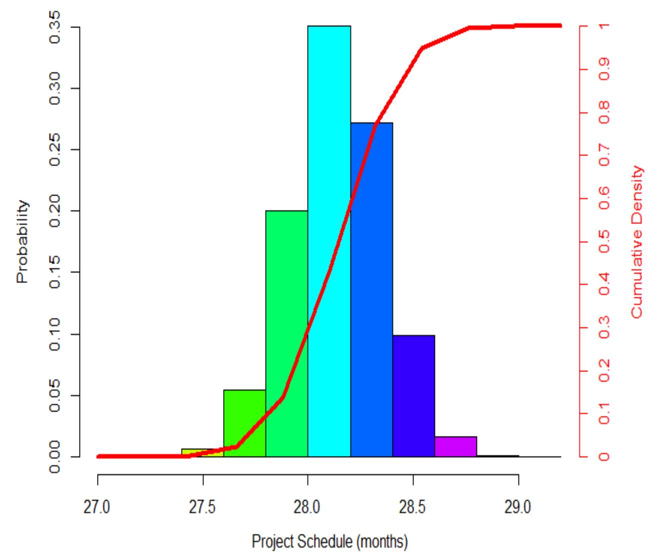


Fig. 6. Simulation Histogram for Model – II.

The simulated model-II shows that the new project schedule can vary from 27 months to 29 months. There is 100% chance that the new project can be completed in 28 months and less than 5% chance that the new project can be completed in 27 months.

The comparison of the two simulated risk model revealed that the new software schedule is underestimated without considering dependence among three risk factors. Further comparison of the simulated result revealed that, there is almost 100% chance for new project to be completed in 30 months in risk model-I and 100% chance to be completed in 28 months in risk model-II. There is almost 40% chance for the new software project to be completed in 29 months in risk model-I and almost 5% in risk model-II. The original duration for new project is 30 months. Its mean that the model-I which considered dependence among risk factors by T-copula evaluated software project duration accurately.

V. CONCLUSIONS

Project delays or failures are practicing routine in many software houses across Pakistan. In this research, we have considered three major risk factors that can negatively impact the estimated project schedule. The risk factors are evaluated by two models. Model-I assumed dependence among risk factors by multivariate T-copula and model-II assumed independence among risk factors. Both models implemented for some software houses based in Karachi and the analysis revealed that model – I which considered dependence among risk factors by T-copula, evaluated project schedule accurately. Multiplier goodness-of-fit test showed that, the chosen T-copula is appropriate for characterization of dependence among three risk factors. It is concluded that if a software manager do not consider dependence among risk factors then he may underestimate the software project schedule. Schedule overruns result in high budgeting cost, dissatisfaction of customers and sometimes failure of software project. Therefore copulas are important for characterization of dependence among several variables and software houses in Pakistan can utilize theory of copulas for better outcomes of project management. Further, theory of copulas can be implemented to other risk factors defined by software project manager.

REFERENCES

[1] Roger S. Pressman, “Software engineering a practitioner’s approach (7^{th} ed).” Dobuque, IA McGraw-Hill, 2009.

[2] Wen-Ming Han, Sun-Jen Huang, “An empirical analysis of risk components and performance on software projects,” Journal of Systems and Software, Volume 80, Issue 1, 2007, Pages 42-50.

[3] T. Christiansen, P. Wuttidittachotti, S. Prakancharoen and S. Arj-ong Vallipakorn “Prediction of risk factors of software development project by using multiple logistic regression” ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 3, Feb. 2015.

[4] Tharwon Arnuphaptrairong “Top Ten Lists of Software Project Risks: Evidence from the Literature Survey” Proceedings of IMECS, Vol. 1, March 16 – 18, 2011, Hong Kong.

[5] P. Jaworski et al. (eds.), Copula Theory and Its Applications, Lecture Notes in Statistics 198, DOI 10.1007/978-3-642-12465-5_1, © Springer-Verlag Berlin Heidelberg 2010.

[6] D. Wu, H. Song, M. Li, C. Cai and J. Li, "Modeling risk factors dependence using Copula method for assessing software schedule risk," *The 2nd International Conference on Software Engineering and Data Mining*, Chengdu, 2010, pp. 571-574.

[7] A. Burns, G. Bernat, I. Broster, “A probabilistic framework for schedulability analysis” *Proceedings of the Third International Conference on Embedded Software (EMSOFT 2003), pp. 1–15 (2003)*

[8] Guillem Bernat , Alan Burns , Martin Newby, Probabilistic timing analysis: An approach using copulas, Journal of Embedded Computing, v.1 n.2, p.179-194, April 2005.

[9] Syed Muhammad Aqil Burney “Fuzzy Logic & Copulas applications in Financial and Actuarial Science” Talk delivered at Centre for Advanced studies in Mathematics (CASM) at LUMS Lahore, Pakistan, 2013.

[10] A. Sklar “Fonctions de repartition a n dimensions et leurs marges” Publ. Inst. Statist. Univ. Paris vol. 8, 229-231 1959.

[11] Roger B. Nelsen, An Introduction to Copulas (Springer Series in Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, 2006.

[12] Durante, Fabrizio, Giovanni Puccetti, Matthias Scherer, et al. "My introduction to copulas" *Dependence Modeling*, 5.1 (2017): 88-98. Retrieved 20 Nov. 2018, from doi: 10.1515/demo-2017-0006.

[13] Joe H., “Multivariate Models and Dependence Concepts,” Chapmann & Hall, London, 1997.

[14] D. Drouet-Mari and S. Kotz, “Correlation and dependence,” Imperial College Press, London, 2001.

[15] Syed Muhammad Aqil Burney, “Risk Theory and Insurance: A stochastic approach,” KU Press, Karachi, (2003).

[16] B. W. Boehm “Software Risk Management Tutorial,” IEEE Computer Society Press, Washington, 1989.

[17] Paul L. Bannerman, “Risk and risk management in software projects: A reassessment,” The Journal of Systems and Software 81 (2008) 2118–2133.

[18] Barry W. Boehm, Clark, Horowitz, Brown, Reifer , Chulani , Ray Madachy , Bert Steece, “Software Cost Estimation with Cocomo II with Cdrom,,” Prentice Hall PTR, Upper Saddle River, NJ, 2000.

[19] S. Demarta and A. J. McNeil “The t copula and related copulas” Technical Report ETH Zurich, 2004.

[20] Breymann, W., A. Dias and P. Embrechts, “Dependence structures for multivariate high frequency data in finance,” Quantitative Finance **3**, 1–14 (2003).

[21] Kojadinovic, Ivan & Yan, Jun, "Modeling Multivariate Distributions with Continuous Margins Using the copula R Package," Journal of Statistical Software, Foundation for Open Access Statistics, vol. 34(i09), 2010.

The Fundamentals of Unimodal Palmprint Authentication based on a Biometric System: A Review

Inass Shahadha Hussein* 1, Shamsul Bin Sahibuddin2, Nilam Nur Amir Sjarif3
Faculty Razak of Technology and Informatics
UTM
Kuala Lumpur, Malaysia

***Abstract*—Biometric system can be defined as the automated method of identifying or authenticating the identity of a living person based on physiological or behavioral traits. Palmprint biometric-based authentication has gained considerable attention in recent years. Globally, enterprises have been exploring biometric authorization for some time, for the purpose of security, payment processing, law enforcement CCTV systems, and even access to offices, buildings, and gyms via the entry doors. Palmprint biometric system can be divided into unimodal and multimodal. This paper will investigate the biometric system and provide a detailed overview of the palmprint technology with existing recognition approaches. Finally, we introduce a review of previous works based on a unimodal palmprint system using different databases.**

***Keywords*—Biometric system; palmprint; palmprint features; unimodal**

I. Introduction

A biometric system is a technological system that uses information about a person (or another biological organism) to identify that person. Biometric systems rely on specific data about unique biological traits in order to work effectively [1]. Biometric systems can be applied to various applications such as criminal identification, car theft, laptop, and desktop authentication security systems, building access, locker protection, national security, military field, immigration, citizen identification, network access, attendance monitoring, and transaction processing system [2]. There are different types of biometric systems that have been employed by researchers. Examples of these types include biometric systems based on voice [3], DNA [4], keystroke [5], face [6], iris [7], hand geometry [8], fingerprint [9], gait [10] signature [11]. The development of palmprint recognition has helped humans perform tasks that were previously difficult or had to be carried out manually. Indirectly, these developments have helped humans by providing faster, more efficient and productive performance of tasks. Many researchers utilized the palmprint authentication system due to its reliable and confidential features [12], [13], [14].

The currently available literature shows that there have been a number of valuable reviews on the palmprint recognition system. However, finding a comprehensive study that is inclusive of various aspects of the palmprint recognition system is scarce. Therefore, this paper aims to investigate the biometric system and provide a review of the unimodal palmprint recognition system.

The paper is organized in the following sections. In Section 2, the biometric system will be introduced. Section 3 comprises an overview of the palmprint technology. Section 4 will review previous works based on the unimodal palmprint system with using different databases. Finally, the paper is concluded in Section 5.

II. Biometric System

Biometrics can be divided into physiological features and behavioral features. The physiological features include fingerprints, hand geometry, and palmprint. Fig. 1 shows the examples of physiological biometrics. The behavioral features comprise signatures, handwriting, and movements as illustrated in Fig. 2. These are unique features that exist in each individual and remain unchanged during a person's lifetime, hence providing a hopeful solution to the community.

Every trait has its metrics and limitations. For instance, the voice is less accurate, a keystroke needs a long observation time, and the face is affected by different poses, illumination, and aging factors. In addition, the iris sensor is very expensive, the hand geometry changes as children grow, the wearing of rings as well as rapid development of pregnant women in a short time, and the DNA is not user-friendly. Other limitations include the faint fingerprints of elderly persons and those involved in manual labors and even missing gait can be influenced by medical conditions, clothing, surface, footwear, and signature which are easy to forge [14, 15].

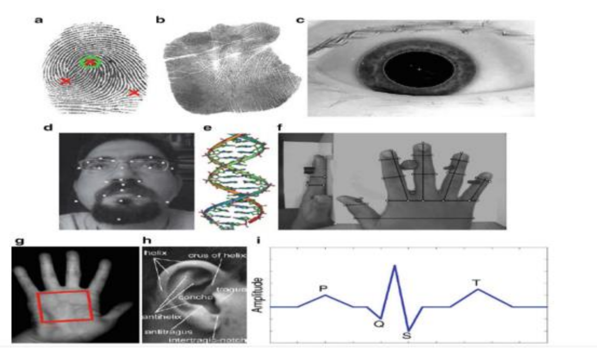


Fig. 1. Examples of Physiological Characteristics: a. Fingerprint, b. Palmprint, c. Iris, d. Face, e. DNA, g. Vein of hand, h. Ear and i. ECG.

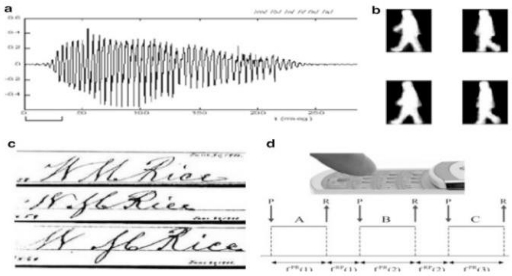


Fig. 2. Examples of Behavioral Characteristics: a. Voice, b. Gait, c. Signature and d. Keystroke.

A. The Components of the Biometric System

The main recognition components which are currently being used by the biometric system can be divided into five units [15-17]:

- Acquisitions unit: This unit depends on the biometric trait used. The sensor will be used to capture the trait which belongs to the user. The sensor captures an image, a voice signal or a frame sequence. The trait captured is called “sample.”
- Segmentation unit: The sample region that contains the biometric information is isolated to produce the region of interest which will be used in the next unit.
- Features extraction unit: The features of the segmented sample are extracted, and a feature vector of the biometric trait is calculated. This vector is so suitable for the storage and analysis of databases through an information processing system. The vectors can be strings of bits or images or coordinates of specific points in the image signals or algebraic functions.
- Matching unit: The vector is compared to one or more of the prior-vectors which are stored in the database. The result of the matching unit is the "degree of matching," and the "similarity between the two comparators."
- Decision unit: The degree of matching is used to produce the final decision output of the biometric system. A threshold for the degree of matching value is used to convert the degree of matching to a logical decision, whether the comparison vectors belong to the same individual or not.

B. Characteristics of the Biometric Traits

The main characteristics of the biometric trait that must be taken into account [15, 18] are as follows:

- Universality: Everyone should possess the biometric trait.
- Distinctiveness: The individuals should be distinguished by the biometric trait.
- Permanence: Over time, the biometric trait should not change.
- Collectability: How easy it is to obtain biometrics.
- Performance: The speed and accuracy that can be acquired by using the biometric trait.
- Acceptance: The willingness of people to supply their biometric attributes for the process of recognition.
- Circumvention: The difficulty of piracy in the biometric system to avoid unauthorized access.

Table 1 illustrates a comparison of the characteristics of different biometric technologies [18].

TABLE I. DIFFERENT BIOMETRIC TECHNOLOGIES CHARACTERISTICS

Biometrics	**Characteristics**						
	Universality	**Distinctiveness**	**Permanence**	**Collectability**	**Performance**	**Acceptability**	
Facial thermogram	H	H	L	H	M	H	L
Hand vein	M	M	M	M	M	M	L
Gait	M	L	L	H	L	H	M
Keystroke	L	L	L	M	L	M	M
Order	H	H	H	L	L	M	L
Ear	M	M	H	M	M	H	M
Hand geometry	M	M	M	H	M	M	M
Finger	M	H	H	M	H	M	M
Face	H	L	M	H	L	H	H
Retina	H	H	M	L	H	L	L
Iris	H	H	H	M	H	L	L
Palmprint	M	H	H	M	H	M	M
Voice	M	L	L	M	L	H	H
Signature	L	L	L	H	L	H	H
DNA	H	H	H	L	H	L	L

(H: high, M: medium, L: low)

Based on this comparison, we can see that the distinctiveness, permanence and performance characteristics for the palmprints are high. These are the main characteristics that make the palmprint a promising technology for authentication purposes.

III. PALMPRINT TECHNOLOGY OVERVIEW

The usefulness of palmprints has been progressively verified in the last fifteen years. A palmprint is defined as a small area on the surface of the palm which contains more information that can be very useful for a person's authentication system.

The images of palmprint contain unique characteristics of a reliable human identification making them a highly competitive feature in the palmprint recognition [19]. The studies of palmprint should utilize either the high resolution or low-resolution images. High-resolution images will be 500 dpi or less, and these images are appropriate for forensic applications such as criminal detection. In most of the previous researches on palmprint, the main focus of recognition is on low-resolution images below 150 (dpi s), and these images are more appropriately used for civil and commercial applications such as access control [20, 21] [22] [23].

Furthermore, the palmprint image has a unique feature as it will be unchanged for a long period of time [24, 25]. Palmprint offers many distinctive features that can be used for precise recognition. The features of palmprint will be discussed in detail in the subsection below.

A. Palmprint Image Features

In general, the features of an image can be divided into two categories, global features, and local features. Global features remain a topic of interest in the past decade. Global means extracting the initial features of the whole image. Each pixel is taken into consideration [26]. The advantages of global features are they compress the size of representation and increase the speed of computation. However, the disadvantages are the variance of geometric distortion. Therefore, using global features alone is not efficient. Local features are considered the efficient representation of the image especially in the field of pattern recognition and computer vision. The local features are invariant to scale, translation and rotation [27]. Global features describe the visual content of the entire image using the vector. They represent texture, color, and shape information. Meanwhile, the local features aim to detect the interest points (IPs) in an image and describe them by a set of vectors [28]. Some examples of global palmprint descriptors are invariant moments (Hu and Zernike) [29, 30] and histogram oriented gradients (HOC) [31]. For local palmprint descriptors, some of the examples are SIFT [32], SURF [33], LBP [34], BRISK [35] and MSER [36] . Fig. 3 represents the palmprint ink region while Fig. 4 represents the palmprint without ink region [37]. Meanwhile, the features of palmprint are shown in Table 2.

B. Importance of Palmprint Technology

Many researchers have highlighted the importance of palmprint system in their studies showing that palmprint has more distinctive features and availability of rich information such as lines, wrinkles, hills, dots, points, and texture. These unique characteristics offer different possibilities to represent the features of the palmprint and pattern recognition [19]. The sensors that are used to capture palmprint are low-cost with very low images of 75 dots per inch (dpi) [24]. The nature of palmprint which is anti-spoofing has made palmprint a reliable biometric feature. [24]. Among the various biometric features include the face, DNA, fingerprint, hand, iris, hand vein retina, sound and signature, palmprint has good potential as a personal identification method because it is fast and generally acceptable. In comparison to face recognition, palmprint is not affected by age and accessories. If compared to the recognition of a fingerprint, palm contains rich information and it only needs low-resolution image device capture that reduces the cost of the system. If compared to the recognition of the iris, images of the palmprint can be captured without intrusions as people may fear the side effects on their eyes.

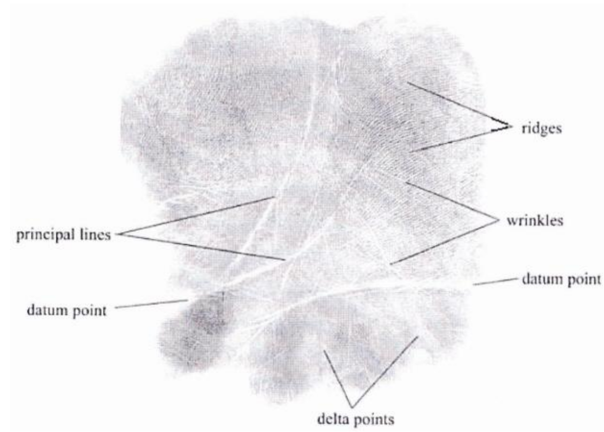


Fig. 3. Palmprint Features with ink.

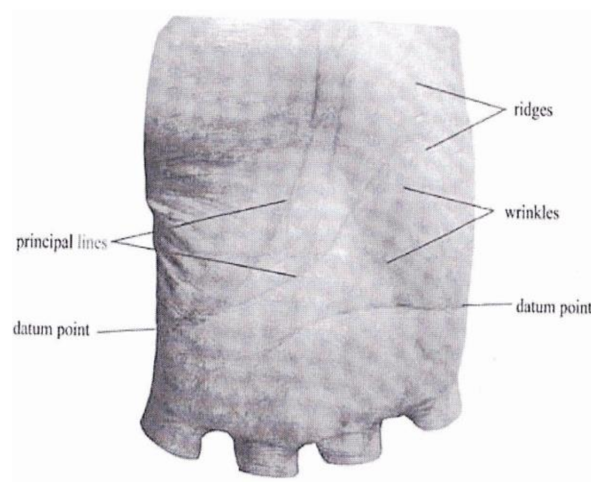


Fig. 4. Palmprint Features without ink.

TABLE II. MAIN FEATURES EXTRACTED FROM PALMPRINT

Name of Feature	Explanation
Geometry	Depending on the shape of the hand; such as height, width, and surface area
Principle lines	Includes the position and location
Datum points	Extracting the two endpoints of the principal lines provides a balance when recording the image. Moreover, by defining Euclidean distance, palmprint size will be obtained.
Wrinkles	Small and shallower
Minutiae	Small and shallow points with various measurements.
Delta points	Small regions which the center of the entire area. They are similar to the shape of a delta.

C. Palmprint System Authentication

Authentication can be defined as giving the authorized persons the proper authority and access in proper time. Authentication is divided into identification and verification. These two different ways the biometric systems can be used according to the needs of the application. The identification phase can be performed once to several times by comparing the biometric data with each modal of the system database to detect the non-personal identity of the specified partition. The identification phase is very complicated and costs time, but it is very useful for negative recognition of critical issues such as in the context of forensic and criminal issues. However, in the phase of verification, the system confirms the identity of the individual already known by comparing the biometric data with the modal stored over in the database with that identity [38].

Verification needs less time since it is necessary to reduce the number of comparisons. Applications such as the security system for laptops or telephones, the presence system, security of access in electronic banking and offices are examples of the verification procedure [14]. Fig. 5 illustrates a typical palmprint authentication system.

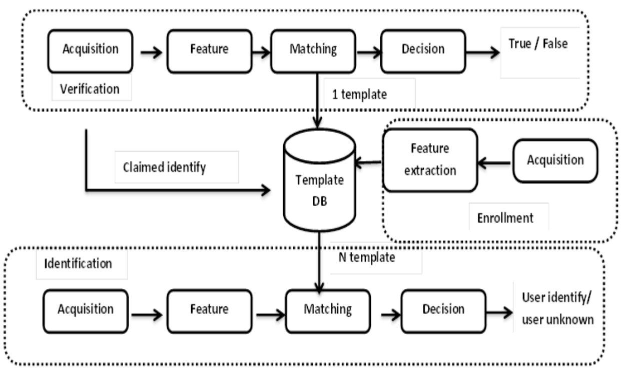


Fig. 5. A Typical Palmprint Authentication System.

As a result, palmprint authentication has the potential to achieve high accuracy for both identification and verification of humans [39]. The authentication of the palmprint of the hand is of great benefit to both civil and forensic applications since approximately 30% of the palmprint latent are raised from scenes of crime (i.e., wheels, steering, blades, etc.). For example, law implementation officials successfully use palmprint recognition to track down killers [38].

IV. Review of Recognition Approaches of Palmprint

The approaches of palmprint matching and feature extraction are divided into three categories called holistic, local and hybrid. All of these approaches are explained in detail below.

A. Holistic based Approaches

Here, the original image will be used as the input of a holistic extractor or matcher. Table 3 summarizes the holistic approaches.

B. Local based Approaches

Local features refer to a pattern or distinct structure found in an image such as a point, edge, or small image patch. They are usually associated with an image patch that differs from its immediate surroundings by texture, color, or intensity. Some examples of local features are blobs, corners, and edge pixels. There are two groups of local features for palmprint recognition; ridges and creases points which can be extracted from high-resolution and low-resolution images.

Table 4 shows a summary of these approaches for local features. SIFT algorithm is one of the accurate methods used to extract local features. SIFT extractor is used for both low-resolution palmprint images and high-resolution images [23].

C. Hybrid- Approach

Hybrid approaches combine two or more of the recognition approaches to get high accuracy. For example, it could be a hybrid of two or three palmprint representations; Gabor line and subspace features LBP + 2DLPP [13] or a combination of the local features with holistic features [40].

V. Review of Previous Works based Unimodal Palmprint System

The first system is called automated finger identification system (AFIS), built to support palmprint recognition in late 1990 by a Hungarian company. This is the first palmprint benchmark. Presently, a large number of researchers are investigating the palmprint technology by using different systems and various algorithms to improve this technology. The first system is unimodal which is then developed according to a multimodal with different criteria. The unimodal biometric system can be defined as using single trait for recognition. In the case of unimodal palmprint, it uses either the left or right palmprint. We will investigate the unimodal based on the following sections: a unimodal with various studies based contact based, contactless and high-resolution palmprint images.

TABLE III. Holistic based Approaches

The approaches	**Work representation**
Subspace method	
Unsupervised linear method	PCA applications [13], ICA [41].
Supervised linear method	PCA + LDA [42]
Kernel method	Kernel PCA and kernel Fisher discriminant applications [43, 44]
Tensor method	2DPCA [13] 2D LPP [25]
Transform the domain	Subspace methods in the transform domains[42]
Invariant moments:	Zernike moments [45] and Hu Invariant moments [46]
Spectral representation:	Global statistical signatures in the wavelet domain [47]
Wavelet signature	
Correlation filter	Advanced correlation filter [48]

TABLE IV. LOCAL BASED APPROACHES

The approaches	**Work representation**
Line based	
Gaussian derivatives	First and second class derivatives Gaussian derivatives [49]
Wide line detector	Extract the location and width from palmprint line [50]
Hausdorff Distance	Line segments Hausdorff Distance application [51]
Codec- Based	2D-Gabor code phase filter responses [23]
Code-Palm	
Code- Fusion	2D-Gabor code phase filter responses with the maximum magnitude [52]
Code- Orientation	Orientation information coding of palm lines [53] RLOC [54]
Texture – descriptor	Applications of LBP [55], DCT Coefficients coding [56].
Scale-invariant feature	Application of SIFT [57]

A. Review based Contact based Palmprint Database

The popular poly U database and Jiaotong Beijing University optical scanner are examples of contact based palmprint images. Fig. 6 illustrates the Poly U database image. There are two main approaches to extract features from the contact based images namely the line feature-based and orientation code based [23]. To extract the line feature either line or edge detectors such as DOG [58], Gabor filter [59], and radon filter [60] can be used. The DOG is sensitive to noise, illumination and it is difficult to distinguish the principal lines from wrinkles. On the other hand, to detect a line, we can detect radon transform (RT) which detects the intensity along potential lines that exist in the small local area. By developing RT and using a modified finite radon transform (MFRAT) [54], the summation of the image pixels over the lines is calculated. In general, the line based image does not achieve a very good result. Fig. 7 shows an example of the DOG line features and MFRAT line features.

Meanwhile, the orientation code based includes palm code [61], competitive code [62], fusion code [63], robust line orientation code (RLOC) [54], double orientation code (DOC) [64] and binary orientation co-occurrence vector (BOCV) [63]. Based on these approaches, a local orientation descriptor is used to represent palmprint images. In general, the orientation based method uses one or multiple orientations. Fig. 8 shows the examples of competitive code.

Furthermore, many algorithms for recognition are being developed over the years. Shang et al. [41] employed the artificial neural network in the palmprint classifier using the radial basis probabilistic neural network (RBPNN) and achieved a good result. Meanwhile, Raut and Humbe [65] suggested the extraction lines from palmprint biometric system by processing the morphological processes. This study discusses the importance of the exact location where the palmprint and shape of the hand images were extracted from a one-hand image acquired by a sensor.

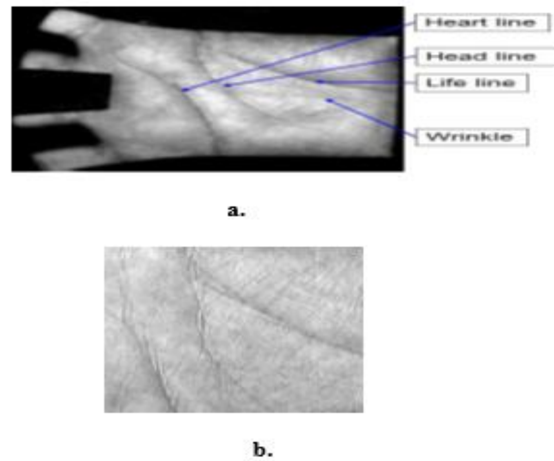


Fig. 6. Poly U Database: a. Contact based Image, b. ROI.

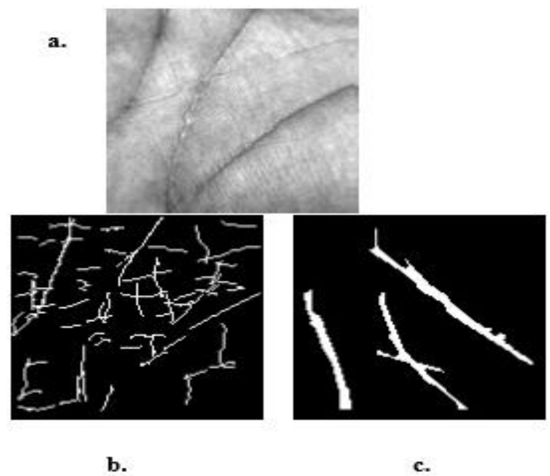


Fig. 7. Line Feature Detector a. ROI, b. DOG Line Features c. MFRAT Line Features.

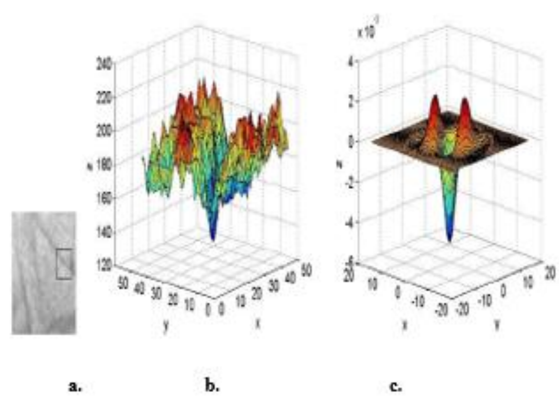


Fig. 8. Examples of Competition: a. Sub-Image, b. The Intensity of Image, c. Gabor Filter with a Small Orientation.

TABLE V. PREVIOUS WORKS BASED CONTACT BASED DATABASE

Refrence	Recognition Approach	Parameter	Database
[41]	Fixed ICA + RBPNN	ACC= 96.76%	PolyU palmprint/ database 214 images
[70]	Traditional SIFT	ACC= 97.9%, FAR=2.1 FRR=2.1	PolyU-online-palm print 600 images
	SIFT + weighted sub region.	99% FAR=2.1 FRR=2.1	
[66]	PCA + NN	ACC= 82.23%	PolyU-online-palm print 600 images
	PCA + CS	ACC=87.33%	
	Gabor + NN	ACC=94.33%	
	Gabor + CS	ACC=96.67%	
	DR + NN	ACC=95.67%	
	DRs + CS	ACC=97%%	
[65]	morphological image processing	ACC= 100% FAR=0 FRR=0	PolyU palmprint_2D/ small size
[24]	(B-BSIF) with SRC based on ICA.	ERR= 4.06	PolyU Palmprint contact less/ 356 images
		ERR= 0	multispectral PolyU / 500 images
[12]	DBN	ACC=85.13**%**	Beijing Jiaotong University
[13]	PCA	ACC= 95.6%	Jiaotong Beijing University optical scanner /500 image
	2D 2PCA	ACC= 96.4%	
	B (2D) 2PCA	ACC= 97.2%	
[56]	DCT and AR signal	ACC=99.79%	Poly U-online-palmprint different size.
[67]	CNN	EER=8.94	The Poly U/193 user
[53]	half-Gabor" filters	EER= 0.0204	The Poly U
[69]	WACSLBP and WSRC	ACC=99.14%	PolyU

In a study conducted by Li et al.[66], they found that the solution to the problem of the illumination and noise of the image of palmprint recognition can be solved by proposing an approach to palmprint recognition through directional representations in low-quality conditions. In addition, they claimed that the Gabor algorithms are not robust for image recognition under uneven illumination and noise. Xin et al. [12] introduced deep learning into palmprint recognition involving three steps. First, a deep belief net (DBN) is built by top-to-down - unsupervised training. Second, the optimum parameters are chosen to build a robust performance. Finally, the testing sample is labeled using the DBN learning models. They proposed that deep learning should be considered as a robust method of palmprint recognition. Raghavendra and Busch [24] proposed a sparse representation of the features scheme for the recognition of palmprint obtained from the Bank of Binary Features of the Statistics Images or B-BSIF. Meanwhile, Zhai et al. [13] proposed that the fusion between the block-wise bi-directional two-dimensional principal components analysis and the classification of grouping spares classification is effective. Ergen [56] introduced a method that integrates discrete cosine transformation (DCT) and an auto-regressive (AR) modeling for biometric identification. In a study by Kumar and Wang[67], a convolutional neural network (CNN) was presented as a matcher left palmprint with right palmprint images of the same hand. They improved the similarity between the left palm and right palm of the same person. Fei et al. [68] performed a Gabor half-filter bank for the extraction of the double half of the palm recognition. In addition, they evaluated this work by comparing this method with a single dominant orientation. It was found that the double half-orientations could characterize the global orientation feature of a palmprint with greater precision. Zhang et al. [69] conducted a hybrid approach by combining the weighted adaptive center-symmetric local binary pattern (WACSLBP) with the weighted sparse representation (WSRC) to improve the classification. This approach consisted of coarse and fine steps. In the coarse step, the similarity between the test sample and one sample of each training class are being used. A small number of candidate classes of the test sample is reserved, and most of the training classes could be excluded. Whereas, in the fine step, the rotation invariant weighted histogram feature vector was extracted from each candidate's sample. Moreover, using WACS-LBP for the testing sample, the weighted sparse representation optimal problem was constructed using the similarity between the test sample and each candidate training sample. Table 5 summarizes previous works founded on the contact based database images.

B. *Review based Contactless Palmprint Database*

IITD and CASIA databases are examples of contactless images, which are collected by using a commercial camera under free environment using a non-user peg. Fig. 9 illustrates a contactless image. So the problems in this database are variants to translation, scaling, rotation, illumination, and noise. The contactless images are under low-resolution. Therefore, the line and texture are the most important features. A number of researchers use subspace learning approaches [68] and collaborative approach [71]. Robust features such as SIFT [72], LBP [73] LLDP [74], and OLOF [75] show excellent robustness to rotation and scaling, illumination and translation changes. These descriptors have achieved a good accuracy in contactless images. Fig. 10 and Fig. 11 illustrate SIFT and LBP features respectively. However, many algorithms for contactless recognition have been developed over the years**.** Kumar and Zhang [76] introduced the user's identity characterization through the simultaneous use of three main representations (hybrid approach) of the palmprint, namely Gabor features, line features, and principal component. Chu et al. [52] proposed a famous Gabor magnitude features. Kumar and Kumar [77] investigated the possibility of using the palmprint as a cryptic vault to develop a reliable and easy-to-use encryption scheme that suggests the use of symmetric and asymmetric encryption approaches. Palanikumar et al. [78] introduced the contrast enhancement scheme called adaptive Increasing Value Histograms equalization (AIVHE) which can be used to adaptively match the Histograms equalization (HE) based on the palmprint lines, textures and hand geometry properties. Zhao et al.[79] used SIFT features and I-RANSAC algorithm for palmprint verification for the contactless dataset. Kanchana and Balakrishnan [80] proposed a method of Rabin-Karp Palm-Print Pattern Matching (RPPM) to improve the matching accuracy of the palmprint features by using the double hash in RPPM. Ali et al. [25] used a hybrid approach for palmprint identification, by using various algorithms such as local binary pattern (LBP), two-dimensional localized preserving projection (2DLPP) and the fusion of LBP + 2DLPP. Table 6 summarizes previous works based on the contactless database.

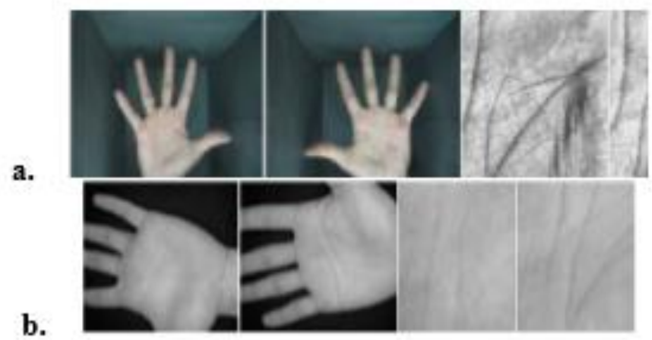


Fig. 9. Contactless Image a. IITD Images b. CASIA Images.

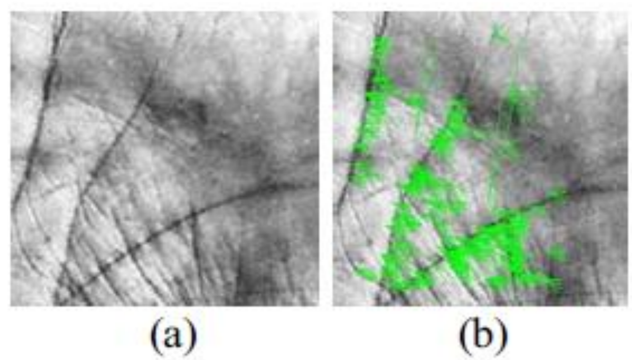


Fig. 10. Extract SIFT Features a. ROI b. SIFT Features.

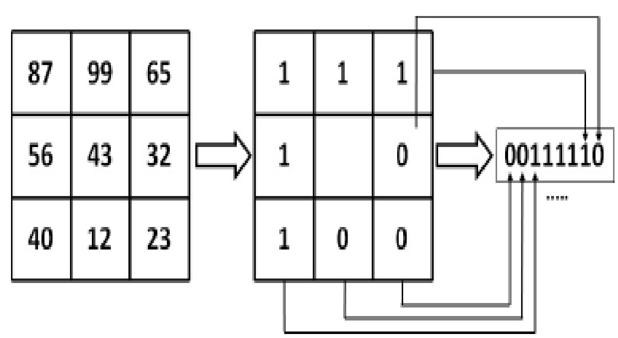


Fig. 11. Examples of LBP.

TABLE VII. PREVIOUS WORKS BASED CONTACTLESS DATABASE

Reference	Recognition Approach	Parameter	Database
[76]	Gabor features	FAR=0.82 FRR=7.20 ERR=4.89	Unknown database of 100 users/. 1000 image, acquired by a digital camera using peg-free setup
	line Features	FAR=2.95 FRR= 7.60 ERR= 6.19	
	PCA features	FAR=2.74 FRR= 7.20 ERR= 5.83	
[52]	Gabor + LDA	ERR= 0.35	UST database / 5560images
		ERR= 0.17	CASIA / 4796 images
[77]	DCT	FAR=0.35, ERR=0.3	Unknown database captured by a digital camera without pinout in indoor environments / 85images
[78].	AIVHE		Unknown database captured by Nokia 2700 with 1200 × 1700 resolution /100 images
[79].	SIFT + I-RANSAC.	ERR=0.5134 left palm ERR= 0.5524Right palm	IIT Delhi's version 1.0
[80]	RPPM	ACC=73.88%	CASIA/35 user
[53]	half-Gabor" filters	EER= 0.0633	IITD database
[25]	LBP	ACC=95.7% FAR=0.043 FRR=0.043 ERR=0.043	CASIA / 1200 image
	2DLPP	ACC=97.33% FAR=0.030 FRR=0.023 ERR=0.0265	
	LBP + 2DLPP	ACC=98.55% FAR=0.0221 FRR=0.0145 ERR=0.0183	
[69]	WACSLBP and WSRC	ACC=99.14%	CASIA

C. *Review based High-Resolution Palmprint Images*

The high-resolution images are usually greater than 500dp [81]. These images include ridge and creaser features whereby the ridges are divided into local ridge direction (LRD) and minutia, while the creaser includes principle lines. Fig. 12 shows the features of high-resolution images from THUPALMLAB.

LRD is the angle between the ridge across the small area and the horizontal axis. (LRD) represents (Ɵij, rij), where Ɵij is the ridge direction at (i, j) pixel and the (rij) represents the crosses ponding direction. [82].

To extract the LRD features, three methods have are widely used; gradient based, discrete Fourier transform and Gabor filter. In these methods, (rij) is calculated by deriving the direction vector in the local window. In a gradient based method, the orthogonal direction which represents the maximum changes in the intensity is calculated as the LRD [83].

In the DFT using the sine wave to represent the ridges in the palmprint image, DFT is considered as the most effective tool to extract the LRD feature [81]. Though, in the Gabor filter based, the LRD area is calculated by choosing the direction of the Gabor filter which performs the max-filtering response after applying multiple frequency spectrum in the local area [83]. However, the Gabor filter is considered as a complicated method that requires more time to execute.

Minutiae points are considered the most significant feature in high-resolution images [82]. Minutiae points are represented by a vector (x, y, ϴ) where x, y are the coordinates of a point and ϴ is the LRD direction. To extract the minutia features, the following three steps are involved [23]:

1) Finding ϴ of LRD by using Gabor filter;

2) Obtaining the Skelton ridge image by binarized and thinned palmprint image;

3) Calculating the center of the ridge of each pixel to detect minutiae points.

Principle lines feature can be seen in both low resolution and high-resolution images [84]. The principles line involves three long creases. The MFART and Hough transform is applied to detect these lines [84]. Fig. 13 shows these three features.

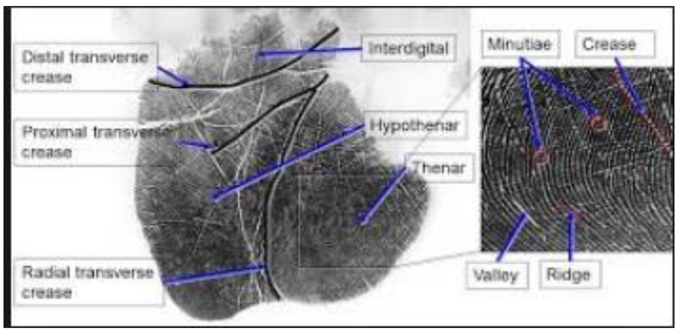


Fig. 12. High-Resolution Feature from THUPALMLAB Database.

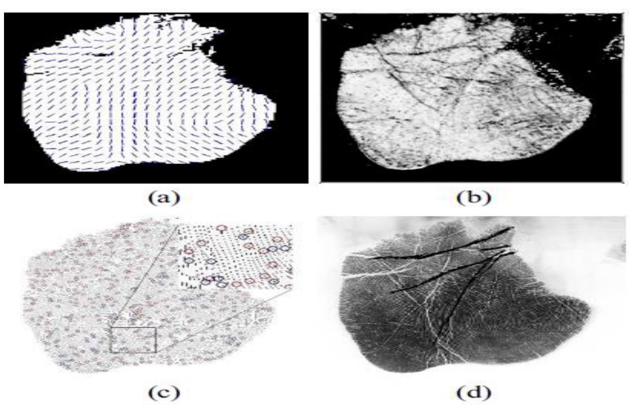


Fig. 13. Line Features: a. LRD Map b. Local Ridge Density Image, c. Minutiae Point Map, D. Principal Line Image.

On the other hand, Carreira et al. [85] used SIFT algorithm with the RGB-mapping algorithm. Dai and Zhou [84] used a multi feature-based matching technique that took 5s to achieve a single match. Jain and Feng [81] used a minutiae code while Wang et al. [86] and Wang et al. [87] used the minutiae based method which performed local matching before global similarity. Cappelli et al. [88] extended the fingerprint encoding for palmprint images. These studies had achieved excellent results but have suffered from high computational complexity for a large number of minutiae points to be compared during matching. To solve this problem, Tariq et al.[22] conducted GPU to decrease the time of execution. The GPU is a special processor used in the graphic to perform true general-purpose. In addition, the processor requires recombining various software. They have a rudimentary programming tool as well as are poor in programming languages. Table 7 summarizes the previous works based on the high-resolution database.

TABLE VIII. PREVIOUS WORKS BASED HIGH-RESOLUTION DATABASE

Reference	Recognition Approach	Parameter	Database
[85]	SIFT	EER=22.5%	THUPALMLAB
[22]	Minutiae encoding	EER=0.38	THUPALMLAB
[20]	RGB-mapping algorithm	---	THUPALMLAB
[84]	Composite algorithm	FRR=16%	THUPALMLAB

(ACC= accuracy, FAR=false acceptance rate, ERR=error recognition rate, FRR= false rejection rate, EER= equal error rate)

VI. CONCLUSION

This paper provides a comprehensive review of palmprint based biometric. We begin with investigating the biometric system. Then, we introduce an overview of palmprint technology; we summarize tables of the previous works based unimodal palmprint based on different databases. To sum up, palmprint recognition is still an open problem and not completely solved. For the importance of contactless images, accuracy needs improvement. Furthermore, the high-resolution images need to be investigated more as these images serve as an important source for high-security application for forensic usages. However, the unimodal biometric system has a variety of problems, such as; the noise, variance within the class; discrimination; nonuniversality and spoofing. To overcome the limitations of the unimodal system, a multimodal palmprint can be used by combining right and left palmprints at different levels of fusion. In the forthcoming paper, we will introduce an overview of multimodal palmprint biometric system.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Higher Education (MOHE) and University Technology Malaysia (UTM) for their educational and financial support. This work is conducted at the Razak Faculty of Technology and Informatics under Cyber-Physical Systems Research Group (CPS RG) and funded by Universiti Teknologi Malaysia (GUP Tier 1: Q. K130000.2538.18H2).

REFERENCES

[1] Gomez-Barrero, M., et al., General Framework to Evaluate Unlinkability in Biometric Template Protection Systems. IEEE Transactions on Information Forensics and Security, 2018. 13(6): p. 1406-1420.

[2] Naderi, H., B.H. Soleimani, and S. Matwin. Manifold Learning of Overcomplete Feature Spaces in a Multimodal Biometric Recognition

System of Iris and Palmprint. in 2017 14th Conference on Computer and Robot Vision (CRV). 2017. IEEE.

[3] Deligne, S., et al., Methods and apparatus for use in speech recognition systems for identifying unknown words and for adding previously unknown words to vocabularies and grammars of speech recognition systems. 2017, Google Patents.

[4] Singh, D., et al., Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. Nature communications, 2016. 7: p. 12778.

[5] Fouad, K.M., B.M. Hassan, and M.F. Hassan, User authentication based on dynamic keystroke recognition. International Journal of Ambient Computing and Intelligence (IJACI), 2016. 7(2): p. 1-32.

[6] Biswas, S. and J. Sil, An efficient face recognition method using contourlet and curvelet transform. Journal of King Saud University-Computer and Information Sciences, 2017.

[7] Pillai, J.K., M. Puertas, and R. Chellappa, Cross-sensor iris recognition through kernel learning. IEEE transactions on pattern analysis and machine intelligence, 2014. 36(1): p. 73-85.

[8] Chan, A., T. Halevi, and N. Memon. Leap motion controller for authentication via hand geometry and gestures. in International Conference on Human Aspects of Information Security, Privacy, and Trust. 2015. Springer.

[9] Yuan, C., X. Sun, and R. Lv, Fingerprint liveness detection based on multi-scale LPQ and PCA. China Communications, 2016. 13(7): p. 60-65.

[10] Makihara, Y., et al., Gait recognition: Databases, representations, and applications. Wiley Encyclopedia of Electrical and Electronics Engineering, 2015.

[11] Vera-Rodriguez, R., et al., DYNAMIC SIGNATURE RECOGNITION FOR AUTOMATIC STUDENT AUTHENTICATION. Departamento de Tecnologia Electronica y de las Comunicaciones, EPS, Universidad Autonoma de Madrid (SPAIN), 2015.

[12] Xin, Z.D.P., et al., Palmprint recognition based on deep learning. ICWMMN2015 Proceedings, 2015.

[13] Zhai, L., et al., An efficient classification method based on principal component and sparse representation. SpringerPlus, 2016. 5(1): p. 832.

[14] Jain, Y. and M. Juneja, A Novel Approach for Multimodal Biometric System Using Iris and PalmPrint, in Progress in Advanced Computing and Intelligent Engineering. 2018, Springer. p. 79-88, springer.

[15] Genovese, A., V. Piuri, and F. Scotti, Touchless palmprint recognition systems. Vol. 60. 2014: Springer.

[16] Gudavalli, M., S.V. Raju, and B. Chandrasekhar. An Optimal Content Matcher for Accurate Human Recognition in Multimodal Biometric Systems. in Advanced Computing (IACC), 2016 IEEE 6th International Conference on. 2016. IEEE.

[17] Srivastava, S., Accurate Human Recognition by Score-Level and Feature-Level Fusion Using Palm–Phalanges Print. Arabian Journal for Science and Engineering, 2018. 43(2): p. 543-554.

[18] Barra, S., Design of a Multi-biometric Platform, based on physical traits and physiological measures: Face, Iris, Ear, ECG and EEG. 2016, Universita'degli Studi di Cagliari.

[19] Zhang, D., W. Zuo, and F. Yue, A comparative study of palmprint recognition algorithms. ACM computing surveys (CSUR), 2012. 44(1): p. 2.

[20] Rajeev, S., et al. Forensic print extraction using 3D technology and its processing. in Mobile Multimedia/Image Processing, Security, and Applications 2017. 2017. International Society for Optics and Photonics.

[21] Iula, A. and D. Nardiello, A Method for Biometric Recognition of Ultrasound Palmprint Based on Principal Lines. International Journal of Future Computer and Communication, 2016. 5(1): p. 13-17.

[22] Tariq, S.A., et al., Massively parallel palmprint identification system using GPU. Cluster Computing, 2017: p. 1-16,Springer Science+Business Media, LLC 2017.

[23] Fei, L., et al., Feature Extraction Methods for Palmprint Recognition: A Survey and Evaluation. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018.

[24] Raghavendra, R. and C. Busch, Texture based features for robust palmprint recognition: a comparative study. EURASIP Journal on Information Security, 2015. 2015(1): p. 5.

[25] Ali, M.M.H., P.L. Yannawar, and A.T. Gaikwad, Multi-Algorithm of Palmprint Recognition System Based on Fusion of Local Binary Pattern and Two-Dimensional Locality Preserving Projection. Procedia Computer Science, 2017. 115(Supplement C): p. 482-492.

[26] ping Tian, D., A review on image feature extraction and representation techniques. International Journal of Multimedia and Ubiquitous Engineering, 2013. 8(4): p. 385-396.

[27] Hassaballah, M., A.A. Abdelmgeid, and H.A. Alshazly, Image Features Detection, Description and Matching, in Image Feature Detectors and Descriptors. 2016. p. 11-45.

[28] Loncomilla, P., J. Ruiz-del-Solar, and L. Martínez, Object recognition using local invariant features for robotic applications: A survey. Pattern Recognition, 2016. 60: p. 499-514.

[29] Li, Y., K. Wang, and D. Zhang. Palmprint recognition based on translation invariant Zernike moments and modular neural network. in International Symposium on Neural Networks. 2005. Springer.

[30] Noh, J.S. and K.H. Rhee. Palmprint identification algorithm using Hu invariant moments and Otsu binarization. in Computer and Information Science, 2005. Fourth Annual ACIS International Conference on. 2005. IEEE.

[31] Jia, W., et al., Histogram of oriented lines for palmprint recognition. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014. 44(3): p. 385-395.

[32] Morales, A., M.A. Ferrer, and A. Kumar. Improved palmprint authentication using contactless imaging. in Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on. 2010. IEEE.

[33] Badrinath, G. and P. Gupta. Robust biometric system using palmprint for personal verification. in International Conference on Biometrics. 2009. Springer.

[34] Guo, Z., et al. Hierarchical multiscale LBP for face and palmprint recognition. in Image Processing (ICIP), 2010 17th IEEE International Conference on. 2010. IEEE.

[35] ZHANGH, L. and Y. LIU, Improved binary robust invariant scalable key-points algorithm fusing depth information. Journal of Computer Applications, 2015. 35(8): p. 2285-2290.

[36] Kang, W. and X. Chen, Fast representation based on a double orientation histogram for local image descriptors. IEEE Transactions on Image Processing, 2015. 24(10): p. 2915-2927.

[37] Kong, A., D. Zhang, and M. Kamel, A survey of palmprint recognition. Pattern Recognit, 2009. 42.

[38] Lumini, A. and L. Nanni, Overview of the combination of biometric matchers. Information Fusion, 2017. 33: p. 71-85.

[39] Wu, X., D. Zhang, and K. Wang, Fisherpalms based palmprint recognition. Pattern recognition letters, 2003. 24(15): p. 2829-2838.

[40] Sun, X., M. Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. in Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on. 2009. IEEE.

[41] Shang, L., et al., Palmprint recognition using FastICA algorithm and radial basis probabilistic neural network. Neurocomputing, 2006. 69(13-15): p. 1782-1786.

[42] Jaswal, G., R. Nath, and A. Kaul. Texture based palm Print recognition using 2-D Gabor filter and sub space approaches. in Signal Processing, Computing and Control (ISPCC), 2015 International Conference on. 2015. IEEE.

[43] Ekinci, M. and M. Aykut, Kernel Fisher discriminant analysis of Gabor features for online palmprint verification. Turkish Journal of Electrical Engineering & Computer Sciences, 2016. 24(2).

[44] Kusban, M., A. Susanto, and O. Wahyunggoro, Combination a Skeleton Filter and Reduction Dimension of Kernel PCA Based on Palmprint Recognition. International Journal of Electrical and Computer Engineering, 2016. 6(6): p. 3255.

[45] Alsubari, A. and R. Ramteke, Extraction of Face and Palmprint Features based on LBP, HOG and Zernike Moments. Extraction, 2017. 172(5).

[46] Liu, Y., Y. Yin, and S. Zhang. Hand Gesture Recognition Based on HU Moments in Interaction of Virtual Reality. in 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics. 2012.

[47] Saranraj, S., et al. Palm print biometric recognition based on Scattering Wavelet Transform. in 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). 2016.

[48] Hennings, P., M. Savvides, and B.V.K.V. Kumar. Palmprint Recognition with Multiple Correlation Filters Using Edge Detection for Class-Specific Segmentation. in 2007 IEEE Workshop on Automatic Identification Advanced Technologies. 2007.

[49] Mu, M., Q. Ruan, and Y. Ming. Shape parameters of Gaussian as descriptor for palmprint recognition based on Dual-tree Complex Wavelet Transform. in IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS. 2010.

[50] Pudzs, M., et al. Complex matched filter for line detection. in 2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA). 2015.

[51] Fang, L. and M.K.H. Leung. Palmprint matching using pairwise relative angle based Hausdorff distance. in 2009 IEEE 13th International Symposium on Consumer Electronics. 2009.

[52] Chu, R., et al. Learning Gabor magnitude features for palmprint recognition. in Asian Conference on Computer Vision. 2007. Springer.

[53] Fei, L., Y. Xu, and D. Zhang, Half-orientation extraction of palmprint features. Pattern Recognition Letters, 2016. 69: p. 35-41.

[54] Jia, W., D.-S. Huang, and D. Zhang, Palmprint verification based on robust line orientation code. Pattern Recognition, 2008. 41(5): p. 1504-1513.

[55] ALMEZOGHY, H.H., et al., HUMAN IDENTIFICATION WITH PALM PRINT BASED ON LOCAL BINARY PATTERN. 2017.

[56] Ergen, B., Scale invariant and fixed-length feature extraction by integrating discrete cosine transform and autoregressive signal modeling for palmprint identification. Turkish Journal of Electrical Engineering & Computer Sciences, 2016. 24: p. 1768-1781.

[57] Doush, I.A. and A.-B. Sahar, Currency recognition using a smartphone: comparison between color SIFT and gray scale SIFT algorithms. Journal of King Saud University-Computer and Information Sciences, 2017. 29(4): p. 484-492.

[58] Zhang, L., H. Li, and J. Niu, Fragile bits in palmprint recognition. IEEE Signal processing letters, 2012. 19(10): p. 663-666.

[59] Bounneche, M.D., et al., Multi-spectral palmprint recognition based on oriented multiscale log-Gabor filters. Neurocomputing, 2016. 205: p. 274-286.

[60] Malik, J., et al., Accuracy improvement in palmprint authentication system. International Journal of Image, Graphics and Signal Processing, 2015. 7(4): p. 51.

[61] Zhang, D., et al., Online palmprint identification. IEEE Transactions on pattern analysis and machine intelligence, 2003. 25(9): p. 1041-1050.

[62] Zuo, W., et al. The multiscale competitive code via sparse representation for palmprint verification. in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. 2010. IEEE.

[63] Guo, Z., et al., Palmprint verification using binary orientation co-occurrence vector. Pattern Recognition Letters, 2009. 30(13): p. 1219-1227.

[64] Fei, L., et al., Double-orientation code and nonlinear matching scheme for palmprint recognition. Pattern Recognition, 2016. 49: p. 89-101.

[65] Raut, S.D. and V.T. Humbe, Biometric Palm Prints Feature Matching for Person Identification. International Journal of Modern Education and Computer Science, 2012. 4(11): p. 61-69.

[66] Li, H., L. Wang, and Z. Zhang, Robust Palmprint Recognition Based on Directional Representations. Intelligent Information Processing VI, 2012: p. 372-381.

[67] Kumar, A. and K. Wang, Identifying humans by matching their left palmprint with right palmprint images using convolutional neural network. Proc. DLPR, 2016.

[68] Fei, L., et al., Low-rank representation integrated with principal line distance for contactless palmprint recognition. Neurocomputing, 2016. 218: p. 264-275.

[69] Zhang, S., et al., Combining Modified LBP and Weighted SRC for Palmprint Recognition. Signal, Image and Video Processing, 2018: p. 1-8.

[70] Qu, Z. and Z.Y. Wang, The Improved Algorithm of Scale Invariant Feature Transform on Palmprint Recognition. Advanced Materials Research, 2011. 186: p. 565-569.

[71] Zhang, L., et al., Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach. Pattern Recognition, 2017. 69: p. 199-212.

[72] Morales, A., M.A. Ferrer, and A. Kumar, Towards contactless palmprint authentication. IET computer vision, 2011. 5(6): p. 407-416.

[73] Ojala, T., M. Pietikainen, and T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, 2002. 24(7): p. 971-987.

[74] Luo, Y.-T., et al., Local line directional pattern for palmprint recognition. Pattern Recognition, 2016. 50: p. 26-44.

[75] Sun, Z., et al. Ordinal palmprint represention for personal identification [represention read representation]. in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. 2005. IEEE.

[76] Kumar, A. and D. Zhang, Personal authentication using multiple palmprint representation. Pattern Recognition, 2005. 38(10): p. 1695-1704.

[77] Kumar, A. and A. Kumar, Development of a New Cryptographic Construct Using Palmprint-Based Fuzzy Vault. EURASIP Journal on Advances in Signal Processing, 2009. 2009(1): p. 967046.

[78] Palanikumar, S., M. Sasikumar, and J. Rajeesh, Entropy Optimized Palmprint Enhancement Using Genetic Algorithm and Histogram Equalization. International Journal of Genetic Engineering, 2012. 2(2): p. 12-18.

[79] Zhao, Q., X. Wu, and W. Bu. Contactless palmprint verification based on SIFT and iterative RANSAC. in Image Processing (ICIP), 2013 20th IEEE International Conference on. 2013. IEEE.

[80] Kanchana, S. and G. Balakrishnan, Palm-Print Pattern Matching Based on Features Using Rabin-Karp for Person Identification. The Scientific World Journal, 2015. 2015: p. 8, Hindawi Publishing Corporation.

[81] Jain, A.K. and J. Feng, Latent palmprint matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009. 31(6): p. 1032-1047.

[82] Maltoni, D., et al., Handbook of fingerprint recognition. 2009: Springer Science & Business Media.

[83] Jain, A.K., S. Prabhakar, and L. Hong, A multichannel approach to fingerprint classification. IEEE transactions on pattern analysis and machine intelligence, 1999. 21(4): p. 348-359.

[84] Dai, J. and J. Zhou, Multifeature-based high-resolution palmprint recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011. 33(5): p. 945-957.

[85] Carreira, L., et al., Personal identification from degraded and incomplete high resolution palmprints. IET Biometrics, 2015. 4(2): p. 53-61.

[86] Wang, R., et al. Towards regional fusion for high-resolution palmprint recognition. in Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on. 2013. IEEE.

[87] Wang, R., et al., Regional fusion for high-resolution palmprint recognition using spectral minutiae representation. IET biometrics, 2014. 3(2): p. 94-100.

[88] Cappelli, R., M. Ferrara, and D. Maio, A fast and accurate palmprint recognition system based on minutiae. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012. 42(3): p. 956-962.

Control of Grid Connected Three-Phase Inverter for Hybrid Renewable Systems using Sliding Mode Controller

Sami Younsi
Technical College at Dammam – KSA
Technical and Vocational Training Corporation

Nejib Hamrouni
Laboratory of Analysis and Treatment of Energetic and Electric Systems (ATEES)
Science Faculty of Tunis-University of Tunis El Manar

***Abstract*—This paper presents a power control approach of a grid connected 3-phase inverter for hybrid renewable energy systems that consists of wind generator, flywheel energy storage system and diesel generator. A sliding controller is developed around the grid connected inverter to control the injected currents which leads to control the active and reactive powers requested by grid and/or isolated loads. In series with the controller, a Space Vector Pulse width Modulation method is used to drive the six inverter switches to generate 3-phase voltages and currents for transferring the desired powers requested by the alternative side. Simulations under Matlab-Simulink software of the hybrid renewable energy systems are made to show performances given by the developed sliding mode controller.**

Keywords—Grid connected systems; sliding controller; hybrid renewable systems; SVPWM

I. INTRODUCTION

The international renewable energy program is a strategic initiative for the world that aims to substantially increase the share of renewable energy in the total energy mix. In 2017, worldwide, renewable power capacity had a 31% share of total power capacity, edging out coal [1]. In terms of total power output, renewable energy is predicted to make 28% of global electricity output by 2021[1]. This strategy is followed after the reduction of fossil fuel sources such as oil and gas. There are many renewable sources of energy such us wind energy, solar energy, and biomass. They appear clearly to complete the used energy. Their production fluctuates and it is not predictable [2,3,4]. These problems affect directly the quality of the active and reactive power transferred to the grid which are very fluctuating and which pose unbalances consumption-production. This limits their rate of penetration into the grid [5,6].

These problems of the grid connection and the quality of the energy transmitted can be solved by the good control of the active and reactive powers transferred from the renewable sources. This control is possible by the action on the inverter which is the principle element of connection between renewable sources and the grid. There are many researches who worked on the control of active and reactive powers generated by renewable wind, solar and hybrid energy systems and they have used various methods [7,8].

In this paper we present the active and reactive powers control for a 3-phase grid connected inverter of hybrid renewable energy system (HRES) that consist of wind system, flywheel energy storage system and diesel generator. A sliding mode controller is developed from the mathematic line model to control the injected currents which leads to control the active and reactive powers requested by grid and/or isolated loads. This method has been used in many works with wind and solar renewable energy systems generally to extract the maximum power (MPPT)[9], to control the grid connected photovoltaic system [10] and to control the electrical machines in wind energy systems[2,3,11]. In series with the sliding controller a Space Vector Pulse Width Modulation (SVPWM) method is used to drive the six inverter switches which generate consequently the necessary 3-phases voltages and currents for transferring the powers requested by the grid. This method is developed in this paper because it is easy to implement on the calculator and it can eliminate the low order harmonics generated by the inverter. The experiment results of the SVPWM show the high reliability of this method. Simulations under Matlab-Simulink software are made to show the results given by the developed sliding mode controller. These results show the advantages of this controller that offers a compromise between good speed of system response and reliability, especially since the system is very fluctuating in very short time depending on the wind speed.

This paper is organized as follows. Section II describes the HRES configuration. Section III presents the control approach within the sliding controller. Section IV presents the simulation results and discussions.

II. SYSTEM COMPONENTS AND DESCRIPTION

The simplest structure of the hybrid renewable energy system that studied is shown on figure 1. It consists of power electronic converters which increase the system efficiency and simplify the control methods. This configuration based on permanent magnetic synchronous machine (PMSM) offers a good efficiency for low power [2,3,4]. The AC/DC converters are used to connect hybrid system with the DC-link. It is a bi-directional converter, to allow the energy transfer between the DC-link and the flywheel in the two directions. The DC-link and the AC grid are connected by capacitor and 3 phase inverter. An RL filter can be also added to smooth the current. Salient parameters of the system are presented in table 1.

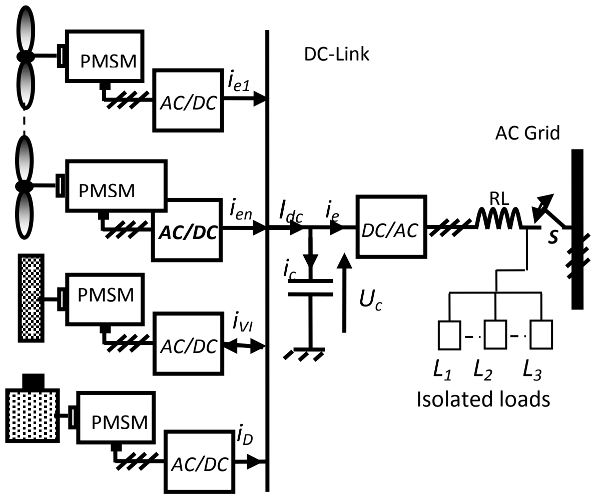


Fig. 1. The Grid Connected Hybrid System Configuration.

III. Control Strategy

In our previous works [2,3] we have developed local methods to control subsystems of the overall HRES and a global supervisor able to generate the reference powers requested by the grid. The aim of this paper is to control, by sliding controller, the 3-phase inverter which connects the DC-link bus to the grid.

The control of active and reactive powers transferred to the grid is carried out by the control of grid injected currents. The control vector $U=[V_{od\_ref} \quad V_{oq\_ref}]^T$ that represents the output reference voltages of the inverter in Park-dq axis is calculated by sliding mode method for current control. The reference values of currents are calculated by using the active and reactive powers reference values P_{req_ref} and Q_{req_ref} that requested by grid and/or isolated loads. These reference powers are calculated by using the powers generated by hybrid renewable energy system and by considering the power in the capacitor for regulation of DC link voltage U_c and losses in the electronic converters.

The PWM signals are determined by using SVM method to drive the six inverter IGBT`s of inverter. They generate the real inverter output voltages from their reference values calculated by the sliding mode controller and transfer the necessary active and reactive powers to the alternative side. Figure 2 shows the control strategy of a three-phase-inverter for HRES.

A. DC- Link Voltage Control

The main aim of the voltage control loop is to regulate the DC-link voltage at a specified value and to provide the reference current. The DC-link voltage is set to 515V while the grid line voltage (V_{rms}) is set to 220V. The closed loop control of the DC-bus voltage is needed because the output power of the renewable sources is variable with climatic conditions. The inverter input reference current $i_{e\text{-}ref}$ is described by the following relation:

$$i_{e-ref} = i_{dc} - PI(U_{c-ref} - U_c) \tag{1}$$

For the control of the DC-link voltage, a proportional integral corrector (PI) has been used. It is parameterized according to the capacitor and the dynamic of the regulation loop. By neglecting losses in the inverter switcher, the power balancing yields:

$$P_{req} \approx i_e . U_c \tag{2}$$

The reference active power injected into the grid is given by:

$$P_{req\_ref} \approx i_{e-ref} . U_c \tag{3}$$

B. Sliding Mode Control of 3-Phase Inverter

The sliding controller requires the knowing of the sliding surface, the condition of convergence and the control law. The general form of the sliding surface is given in[12]:

$$S(x) = \left(\frac{\partial}{\partial t} + \lambda_x\right)^{r-1} e(x) \tag{4}$$

With:

$e(x) = x_{ref} - x$, is the error of the variable to be regulated

λ_x is a positive constant which interprets the band-width of the desired control.

r: is a number of times that it is necessary to derive the output variables to reveal the command.

S(X)=0: A linear differential equation whose single solution is $e(x) = 0$.

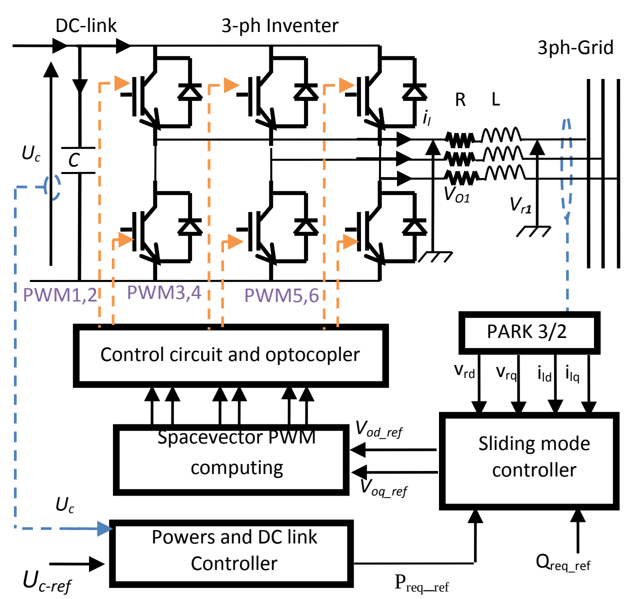


Fig. 2. Control Strategy Diagram.

The general form of the sliding controller is $U=U_{eq}+U_{nl}$. U_{eq} and U_{nl} are respectively the equivalent controller that is the solution of the equation $\dot{S}(X) = 0$ and the nonlinear part that can be interpreted as the overage value of the controller which permits to maintain the state of the system on the sliding surface [13,14]. The nonlinear command (*U_{nl}*) is given to guarantee the attractively of the variable to be controlled towards the sliding surface and to satisfy the condition of convergence. The simplest function is in the form of relay. It is given by U_{nl} =K.sign (S(X)) with *K* is a positive constant [13,14].

C. Application to the inverter

The control vector is carried out by the model of the RL-line filter. This model is given, in Park dq- axis, by:

$$\begin{cases} V_{Od} = Ri_{ld} + L\dfrac{di_{ld}}{i_{ld}} - L\omega_s i_{lq} + V_{rd} \\ V_{Oq} = Ri_{lq} + L\dfrac{di_{lq}}{i_{lq}} + L\omega_s i_{ld} + V_{rq} \end{cases} \tag{5}$$

With V_{od} and V_{oq} are the inverter output voltages in *dq*-axis ; V_{rd} and V_{rq} are the grid voltages in *dq*-axis ; *R* and *L* are respectively the resistance and the inductance of the filter; ω_s is the frequency of injected currents. The control vector $U = \begin{bmatrix} V_{od\_ref} & V_{oq\_ref} \end{bmatrix}^T$ is given by the direct inversing of Eq. 5. It's expressed as follows:

$$\begin{cases} V_{od\_ref} = V_{ld\_ref} - L\omega_s i_{lq} + V_{rd} \\ V_{oq\_ref} = V_{lq\_ref} + L\ \omega_s i_{ld} + V_{rq} \end{cases} \tag{6}$$

The output is the currents vector $I_l = \begin{bmatrix} i_{ld} & i_{lq} \end{bmatrix}^T$. $S_1$and S_2 constitute respectively the sliding surfaces of the exit variables i_{ld} and i_{lq} . They represent the errors of the direct and quadratic currents. They are expressed as follows:

$$\begin{cases} S_1 = I_{ld\_ref} - i_{ld} \\ S_2 = I_{lq\_ref} - i_{lq} \end{cases} \tag{7}$$

By using the equations of system (6), the first derivation of Eq. 7 gives:

$$\begin{cases} \dot{S}1 = \dot{I}_{ld\_ref} + \dfrac{1}{L}\left[Ri_{ld} - L\omega_s i_{lq} + V_{rd}\right] - \dfrac{1}{L}V_{od} \\ \dot{S}2 = \dot{I}_{lq\_ref} + \dfrac{1}{L}\left[Ri_{lq} + L\omega_s i_{ld} + V_{rq}\right] - \dfrac{1}{L}V_{oq} \end{cases} \tag{8}$$

The relative degree of the system is equal to 1 because the command appears in the first derivation of the variables to be controlled. Equation 8 can be replaced by:

$$\begin{cases} \dot{S}1 = B_1 + A_1 V_{ld} \\ \dot{S}2 = B_2 + A_2 V_{lq} \end{cases} \tag{9}$$

With :

$$\begin{cases} B_1 = \dot{I}_{ld\_ref} + \dfrac{1}{L}\left[Ri_{ld} - L\omega_s i_{lq} + V_{rd}\right] \\ A_1 = -\dfrac{1}{L} \\ B_2 = \dot{I}_{lq\_ref} + \dfrac{1}{L}\left[Ri_{lq} + L\omega_s i_{ld} + V_{rq}\right] \\ A_2 = -\dfrac{1}{L} \end{cases} \tag{10}$$

The solution of equation $\dot{S} = 0$ leads to the equivalent controller (U_{eq}). Its expression is given by:

$$\begin{cases} V_{ld\_eq} = -\dfrac{B_1}{A_1} \\ V_{lq\_eq} = -\dfrac{B_2}{A_2} \end{cases} \tag{11}$$

The nonlinear part of the sliding controller is given by:

$$\begin{cases} V_{ld\_nl} = K_d\, signS_1 \\ V_{lq\_nl} = K_q\, signS_2 \end{cases} \tag{12}$$

With:

$K_d = K_q = K = |V_{max}|$.

The global command $U = \begin{bmatrix} V_{od\_ref} & V_{oq\_ref} \end{bmatrix}^T$ is the sum of the equivalent and the nonlinear parts given by equations 11 and 12. It's expressed as follows:

$$\begin{cases} V_{od\_ref} = L\ \dot{I}_{ld\_ref} + R\ i_{ld} - Lw_s i_{lq} + V_{rd} + K_d\, signS_1 \\ V_{oq\_ref} = L\dot{I}_{lq\_ref} + Ri_{lq} + Lw_s i_{ld} + V_{rq} + K_q\, signS_2 \end{cases} \tag{13}$$

The reference currents, given in 13, are expressed as functions of the active and reactive reference powers requested by the grid and / or isolated loads. They are given by:

$$\begin{cases} I_{ld\_ref} = \dfrac{P_{req\_ref}.V_{rd} + Q_{req\_ref}.V_{rq}}{V_{rd}^2 + V_{rq}^2} \\ I_{lq\_ref} = \dfrac{P_{req\_ref}.V_{rq} - Q_{req\_ref}.V_{rd}}{V_{rd}^2 + V_{rq}^2} \end{cases} \tag{14}$$

D. Three-Phase Inverter SVPWM

The SVPWM (Fig.3) is a technique used to drive the inverter switches to generate equilibrate three-phase voltages with desired amplitude and frequency; V_{oa}, V_{ob} and V_{oc}, from the reference voltages V_{oa_ref}, V_{ob_ref} and V_{oc_ref}. These voltages are calculated by the Park transformation (dq-abc) of the inverter voltages V_{od_ref}, V_{oq_ref} provided by the sliding controller bloc [15,16].

As indicated by figure 4, the algorithm of SVPWM is based on the knowledge of the reference vector $\overrightarrow{V_{ref}}$ which rotates in six sectors in the switching hexagon. Its coordinates are the components $V_{\alpha ref}$ and $V_{\beta ref}$, and the change values for each angle θ_s which allowing the calculation of the switching times.

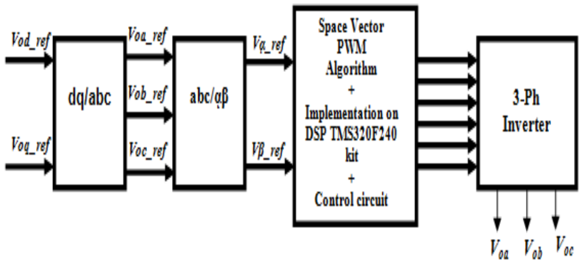


Fig. 3. SVPWM Signals Generation.

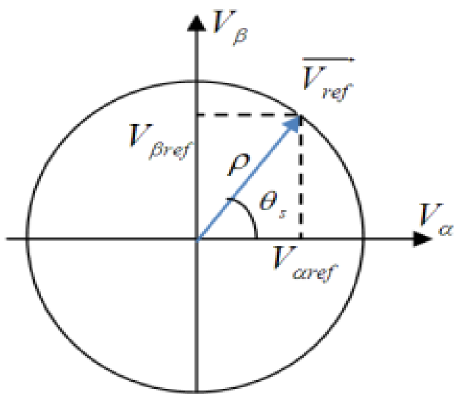


Fig. 4. Reference Vector Variation in αβ-axis.

- Amplitude Variation of Output Voltages

The components $V_{\alpha ref}$ and $V_{\beta ref}$ are expressed as follows:

$$\begin{cases} V_{\alpha ref} = \rho \cos\Theta_s \\ V_{\beta ref} = \rho \sin\Theta_s \end{cases} \quad (15)$$

The maximum radius of the circle within the switching polygon is $\rho_{max} = \sqrt{\frac{3}{2}} V_{max} = \frac{1}{\sqrt{2}} U_c$.

The 3-phase inverter controlled by SVPWM can generate three phase- voltages with maximum value equal to $\frac{U_c}{\sqrt{3}}$. Therefore, by selecting the variable ρ, we can set the amplitude of the inverter output signals.

- Frequency Variation of Output Voltages

The discretization of the reference vector requires a good choice of the number of samples per revolution, and therefore the sampling period on which to build it. The maximum number of samples per period is limited by the speed of calculation of the DSP TMS320F240.In general, for a number faster than 72 (5 ° for a sample) the reference vector is continuously variable over time. To fix the period of the reference vector and the period of the inverter output voltages, it is necessary to fix the modulation period given by [17]:

$$T_m = 2.T_e.\frac{n}{F_h} \quad (16)$$

With

- Te is the period corresponds to the count number of the (T1CNT) of DSP (TMS320F240). It is used to compare the corresponding values to the inverter switching times.
- F_h is the clock frequency of the DSP fixed at 20 MHz.
- n is the clock frequency of the DSP divider.
- The period of the inverter output variables is given by:

$$T = kT_m = 2kT_e \frac{n}{F_h} \quad (17)$$

The frequency of the inverter three-phase output voltages is 50Hz (T=0.02s). To set this frequency value the following parameters are selected:

- k = 360.
- T_e= 138.
- n = 4.

Figure5 shows the IGBT driving signals PWM_i for i={1, 2…6}.

IV. Simulation Study

Simulations by Matlab-Simulink software were performed to test and verify the 3-phase inverter control presented and developed in the previous section and to study the output inverter voltages, the injected currents, and the active and reactive powers transferred from the hybrid renewable energy system to the grid.

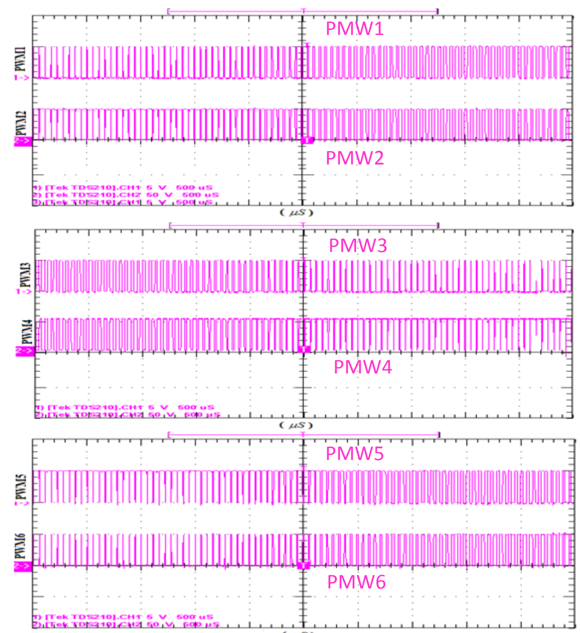


Fig. 5. Driving Signals.

These simulations have been obtained under various values of power requested by grid and/or loads. They are selected to demonstrate the most significant performances of the control approach. The reference and measure are respectively represented by dotted and continuous line. The interaction between the renewable energy sources is not discussed in this paper. Only the inverter control is considered in this paper.

TABLE I. PARAMETERS OF THE GRID CONNECTED SYSTEM

value	symbols	Parameters
0.01H	L	Inductance of the filter
0.92µF	C	Capacitance of the DC-link
0.05Ω	R	Resistance of the filter
314rd/s	w_s	Pulsation of the grid voltage
311V	K_d=K_q	Constant of the sliding control
515V	U_{cref}	DC-link voltage reference

A variable wind signal, shown on figure 6, is applied to the turbines. It varied between 7 and 11m/s regarding the maximal power point tracking zone.

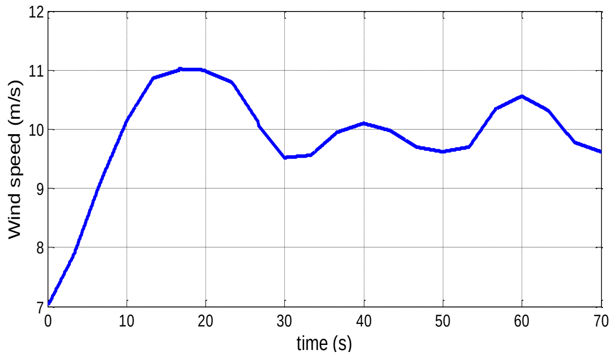
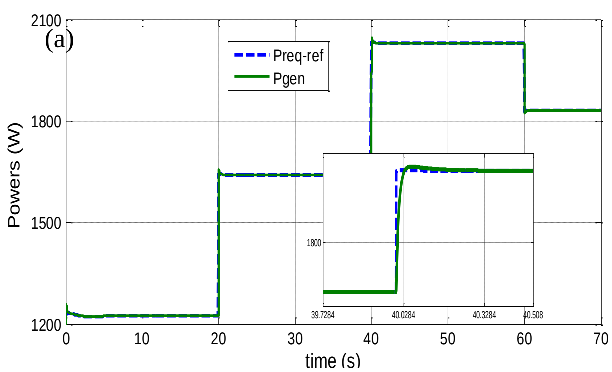


Fig. 6. Wind Speed Variation.

As indicated by figure 7, the reference active power requested by the grid and/or loads (P_{req_ref}) is considered variable within the time while the reference reactive power should be zero in order to obtain the line current in phase with the grid voltage. On figure 7a, it is shown that the active power generated by the hybrid system satisfies the requested power however the reactive power transferred to the grid oscillates around its reference fixed at zero (Fig. 7b). According to those figures, the inverter output reactive power remains equal to zero while the active power varied with the demand of the load in the alternative side.



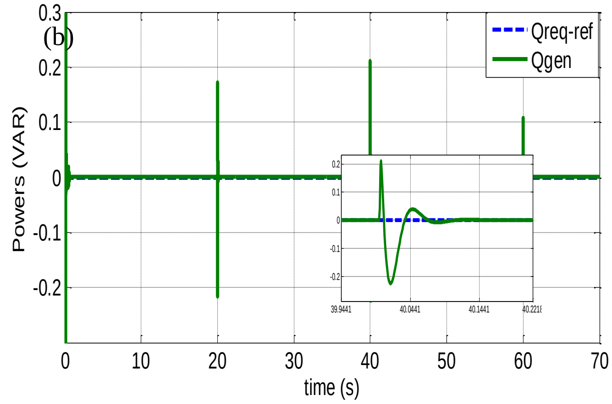


Fig. 7. Inverter output powers (measure and reference): (a): active powers, (b): reactive powers.

Figure 8a presents the DC-current provided by the renewable energy system to satisfy the power requested by the load. The current varied proportionally with the power values injected into the alternative side while the DC-link voltage is constant and follows its reference (U_{c_ref}) fixed at 515V. The DC-link voltage presents some instantaneously overshoots in each change of requested power load (Fig. 8b).

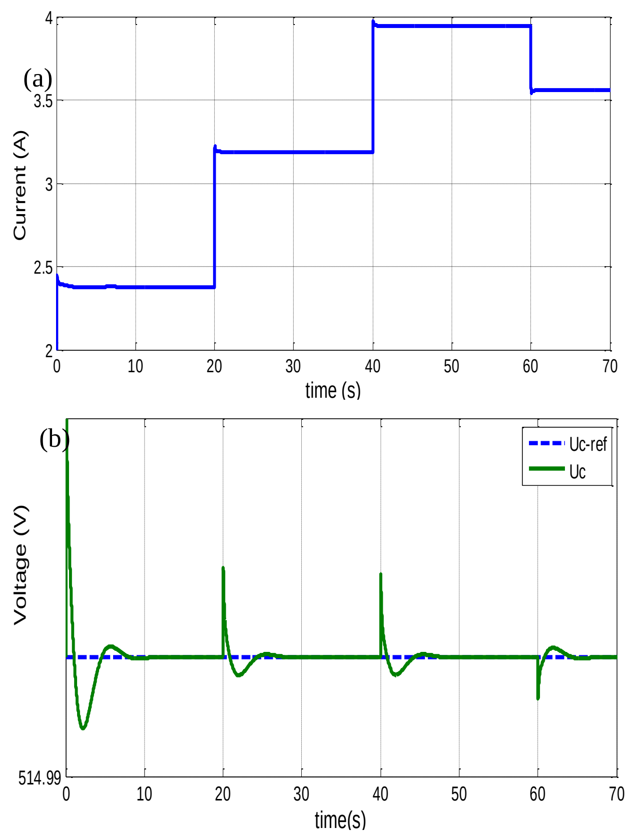


Fig. 8. Input characteristics of the inverter; (a): DC-current provided by the renewable energy system, (b): DC-link voltage (measure and reference).

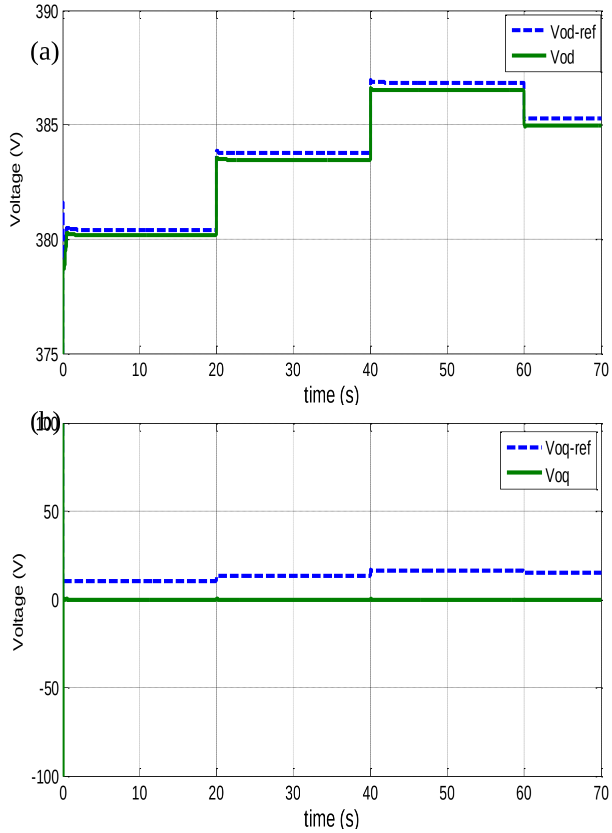
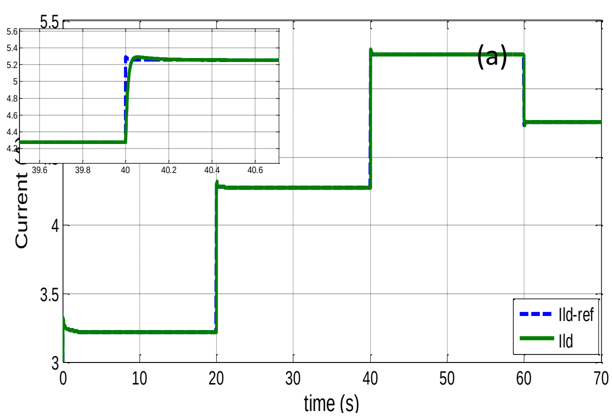


Fig. 9. Inverter output voltage (measures and references); (a): direct components; (b): quadratic components.

Figure 9 illustrates the dq-components of the inverter output voltage under variable load powers. As it can be seen by this figure, the sliding mode controller presents good performances in terms of stability and response time. The inverter output voltages reach instantaneously their reference values provided by the controller within the changes of the requested load power. The direct and the quadratic components of the inverter output current i_{Idq} are represented respectively by figures 10a and 10b. As it can be seen by those figures, the inverter output currents reach instantaneously their reference values provided by the sliding controller.



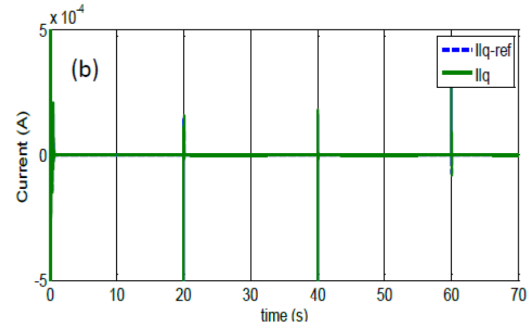


Fig. 10. Iinverter output current (measures and references); (a): direct components; (b): quadratic components.

Finally, the control unit composed by the sliding controller, the DC-link voltage regulator and the SVPWM block, achieves the control of the 3-phase grid connected inverter of hybrid renewable energy system. It permits to regulate instantaneously, the DC-link voltage, the inverter output powers (active and reactive powers) and the currents injected into the grid. The previously figures show the reliability and the short time response of the global system with the developed controller, even with the wind system that is very fluctuating and the diesel generator which has a very slow dynamics since it is a mechanical system.

V. Conclusion

In this paper, a control strategy based sliding mode controller, has been proposed. The command approach ensures the control of the 3-phase inverter output currents which leads the control of the active and reactive powers transferred into the grid and/or the isolated loads. In series with the sliding controller a SVPWM was developed to drive the six inverter switches to generate 3-phase voltages necessary to transfer the desired powers. This command approach (sliding controller with SVPWM) has given very acceptable results in terms of response time and accuracy of the controlled variables, especially since this system is very variable and has disturbances according to the wind speed.

Acknowledgment

This work is part of the project “Projets Jeunes chercheurs” funded by the Tunisian Ministry of High Education and Scientific Research. The support of the ministry is kindly acknowledged.

References

[1] E. Gosden, "Global renewable power capacity overtakes coal as 500,000 solar panels installed every day”, The Telegraph, London,2016.

[2] S. Younsi, M. Jraidi, N. Hamrouni and A.Cherif, "Modelling and control of hybrid renewable energy system connected to AC grid", International journal of computer sciences and engineering (IJCSE), Vol. 3 No. 12, 2011.

[3] S. Younsi, M. Jraidi, N. Hamrouni and A. Cherif, "Artificial neural network control of hybrid renewable energy system connected to AC grid", International journal of computational intelligence technique (IJCIT) Vol. 2, no. 2, pp. 44-52, 2011.

[4] S. Younsi, M. Jraidi and A. Cherif, "Modelling and control of wind farm associated with a flywheel energy storage system and a diesel generator connected to the network", Journées Tunisiennes d’Electrotechnique et d’Automatique (JTEA’2010), 26-28 Mars, Tunisia, 2010.

[5] S. El AIMANI, "Modélisation de différentes technologies d'éoliennes intégrées dans un réseau de moyen tension", Thèse de doctorat de l'école centrale de Lille, 2003.

[6] Gabriel-Octavian CIMUCA, "Système de stockage d'énergie associé à des générateurs éoliennes", Thèse de Doctorat de l'ENSAM, 2004.

[7] H. Ozbaki and M.Kesler, "Active and reactive power control of grid-tied three phase inverter for PV system", International journal of hydrogen energy, August 2016.

[8] M. Binh and M. T. Dat, "Active and reactive power controller of grid-connected single phase photovoltaic system", International journal of hydrogen energy.

[9] D. Rekioua and A. Y. Achour, "Tracking Power Photovoltaic system with sliding mode control strategy", Energy Procedia, Vol. 36, pp. 219-230, 2013.

[10] F. Pires,"Dual-inverter for grid connected photovoltaic system: modeling and sliding control", solar energy, vol. 86, no. 7, pp. 2106-2115, 2012.

[11] H. Amimeur and D. Aouzelleg, "Sliding Mode Control of a dual-Stator induction generator for wind energy conversion systems", International Journal of Electrical Power & Energy Systems.vol. 42 no. 1, pp. 60-70, 2012.

[12] V.I. Utkin, " Sliding mode in control optimization", springer-verlag, Berlin 1992.

[13] H. Buhler, "Réglage par mode de glissement", presse polytechnique romande, 1986.

[14] V.I. Utkin, "Sliding mode control design principles and application to electric drives", IEEE Trans on Elect, vol. 40 pp. 23-36, 1993.

[15] H. djaghloud and H. Benallah, "Space Vector Pulse Width Modulation Applied to Three-level Voltage Inverter", 5th International Conference on Technology and Automation ICTA'05, Thessaloniki, Grece Oct. 2010.

[16] J. H. Seo, C. H. Choi and D. S. Hyun " A new Simplified Space Vector PWM Method for Three-level Inverter " IEEE Transactions on Power Electronics, vol. 16, no. 4, pp. 545- 550, 2010.

[17] K.B. Priya, J. B. V. Subrahmanyam, C. Sikanth and M. Ayoub, "Space vector PWM Technique for 3phase voltage source inverter using Artificial Neural Network", International Journal of Engineering and Innovative Technology, Vol. 1, 2012.

AUTHOR'S PROFILE

Nejib Hamrouni received his engineering degree from the National Engineering School of Sfax, in 2000 and the PHD from the National Engineering School of Tunis, in 2009, both in electrical engineering. He is an Assistant professor at National Engineering School of Gabés from 2010 to 2015. Since September 2015 he is an assistant professor at ISSAT of Mateur. He has participated in several research and cooperation projects, and is the author of more than 20 international communications and publications.

Sami Younsi obtained his engineering degree from the National Engineering School of Sfax and his PHD in electrical engineering in 2013 from the Science Faculty of Tunis. He is an Assistant professor at the Institute of Technologies of Tunis.

Security Issues in Cloud Computing and their Solutions: A Review

Sabiyyah Sabir
Department of Computer Science & Engineering
University of Engineering & Technology (UET)
Lahore, Pakistan

***Abstract*—Cloud computing is an internet-based, emerging technology, tends to be prevailing in our environment especially computer science and information technology fields which require network computing on large scale. Cloud computing is a shared pool of services which is gaining popularity due to its cost effectiveness, availability and great production. Along with its numerous benefits, cloud computing brings much more challenging situation regarding data privacy, data protection, authenticated access etc. Due to these issues, adoption of cloud computing is becoming difficult in today's era. In this research, various security issues regarding data privacy and reliability, key factors which are affecting the cloud computing, have been addressed and also suggestions on particular areas have been discussed.**

Keywords—Cloud computing data protection; encryption; digital signature; security issues

I. Introduction

Cloud computing or cloud-based environment is a service that is internet based and that gives the facility of sharing computer resources along with other devices on demand. It is a mechanism to enable on demand shared resources. For example, server, data center, networks storage applications which can store data. That can be generated with minimum effort. Cloud computing provides the facility to the organizations and users to keep their data on private or third-party storage location and these location/data centers may be located far away from user may be in some other city or country in the world.

National institute of Science and Technology (NIST), gives the cloud computing's definition as "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" [1].

Figure 1 shows the characteristics of cloud services which help others to understand and comprehend the cloud computing in a better way. These characteristics are explained as under [2]:

A. On Demand Self-Service:

It refers to the service which enables provisioning of cloud resources to vendors on demand or whenever they are required such as network storage, service time without the interaction of human.

B. Broad Network Access:

Services are accessible over the network which are retrieved through some standardized mechanism which promotes the usage of heterogeneous platforms (workstations tablets, laptops, mobile phones).

C. Resource Pooling:

Resources of cloud Provider are pooled over server. Consumers are assigned different resources which are either physical or virtual one. Generally, consumer have no idea of exact location the resources provided to them except at the abstraction level like; state, country or data center.

D. Rapid Elasticity:

Services can be elastically released and monitored, for consumers services available to them can often appear as unlimited which can be scaled in quantity anytime.

E. Measured Services:

Cloud system are so designed that they can monitor the resources usage; for example, processing, bandwidth and active user accounts, storage to deliver transparency to provider as well as consumer. At some level of abstraction, they can optimize the resource usage by keeping a check through metering capability.

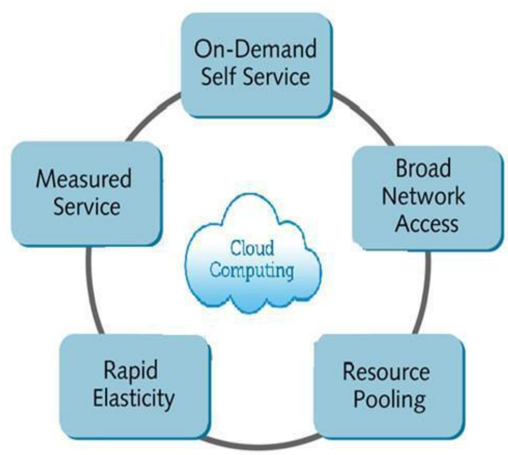


Fig. 1. Cloud Computing Characteristics.

Paper is divided into following sections; section tells about introduction of cloud computing, section II tells about the cloud computing models, section III is related work, section IV is factors affecting cloud computing, section V possible threats regarding cloud computing, section VI is about solutions to the security issues and section VII concludes the paper.

II. Cloud Service Models

Following service models are defined by

NIST which includes three categories [3]:

- Infrastructure as a Service (IaaS)
- Software as a Service (SaaS)
- Platform as a Service (PaaS)

Figure 2 explains the overall three models of cloud computing which are served to the clients according to their needs. These models are explained as under:

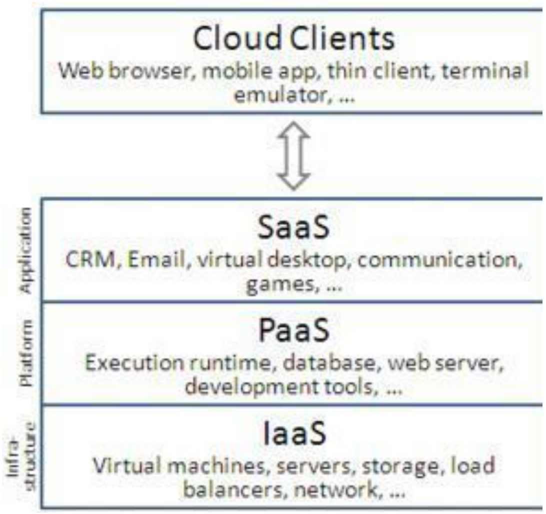


Fig. 2. Different Cloud Service Models.

A. Infrastructure as a Service (IaaS):

Iaas is all about providing the virtual machine, operating system or networks to the end users. Some other computing resources are also supported in IaaS, where the customer or client can run arbitrary operating system on virtual machines or any other software. Clients can control only the operating system or the software which he is running but he loses his control on the infrastructure which is providing him all these services.

B. Software as a Service (SaaS):

In this kind of scenario, user is only using the applications which are being provided by the vendor and those applications run on the cloud services. Same application is accessible by many other clients as well through some common mechanism, for example by using web browser, or email. Again, the clients or users have no control over the application or underlying infrastructures, network server or operating system upon which these applications run.

C. Platform as a Service (PaaS):

In PaaS, the client is able to create their own desired application by using some programming language, linked libraries. These languages or libraries are supported by the vendor. After creating the user desired application, it is deployed on the server provided by the vendor. User has also the authority to configure its application or can change the configuration settings later on.

The benefits of cloud computing might be very appealing but it has got huge number of risks and security issues like data leakage, data loss, intruders attacks, malicious insiders etc.

III. Related Work

Ayush Agarwal et al. (2016) highlight the emergence of cloud computing along with its security concerns like data loss, data breaches, insecure API's, account hijacking, denial of service [4]. Prachi Garg et al. (2017) have worked on different cloud security aspects like basic security which includes Cross site scripting attacks, Sql injection attacks, Man in the middle attacks [5]. Pradeep Kumar Sharma et al. (2017) security concerns for cloud like cost model charge model [6], service level agreements and issue of migration should be dealt. Naseer Amara et al. (2017) highlighted the security threats, architectural principles and cloud security attacks with their techniques that can minimize the effects of malicious attacks (mitigation techniques) [7]. Sh. Ajoudanian et al, (2012) said that following four parameters were the most crucial. (a) Data Confidentiality, used to avoid leakage of information to any unauthorized individual or system [8].

IV. Factors Affecting Cloud Security

There are numerous key factors which may affect cloud computing performance because it is surrounded by many technologies e.g load balancing, network, concurrency control, virtualization, operating system, database, memory management etc [9]. Figure 3 shows these concerns which are discussed as above.

Fig. 3. Factors Affecting Cloud Security.

The security factors of these technologies affecting the cloud computing are appropriate e.g. network which connects the cloud computing to the outer world has to be secured. Virtualization concept has to be carried out securely when mapping with the physical systems. Load balancing involves the handling the incoming requests traffic which sometimes overloads the server. Data mining algorithms can be applied to cope with malicious attacks.

V. Possible Threats Regarding Cloud Computing

Nowadays cloud computing is getting so much popularity that it is in the limelight of today's era. Along with its huge benefits cloud computing is facing much security issues which need considerable attention to resolve them for the betterment of this service. Following are the major concerns as described below [10];

- **Outsourcing:** in outsourcing the data, consumer might get lose the control. Some kind of appropriate mechanism is needed to prevent the cloud service provider (CSPs) to use the data against the consent of their clients.
- **Multi tenancy:** cloud is a shared pool of resources. Protection of data must be taken into account while providing the multi-tenant environment.
- **Service Level Agreements:** a clear contract between the consumer and provider is needed. The main goal of agreements is to build the trust.
- **Heterogeneity:** different cloud providers have different mechanism of data protection which leads to integration challenges.
- **Server Downtime:** Downtime is the time in which the system starts responding to the client after some service failure. Downtime should be kept minimized and power backups must be installed to keep downtime minimum.
- **Backup:** Data uploaded by the clients, should be backed up in case of any service failure. Cloud Seller should mention in SLAs that in case of any disaster, what should be the remedy or solutions to such problems. There are very rare chances of whole system failure like flood etc.
- **Data Redundancy:** Data redundancy is a situation in which same data is being kept on two different places. In case of cloud computing, it can be understood as to provide copies of same data, systems or equipment to the clients. cloud seller should try to keep data redundancy minimum.

VI. Solutions To Security Challenges In Cloud Computing

Security challenges in cloud computing need to be addressed properly. If appropriate solutions are not being provided adoption of cloud environment becomes more difficult. Apart of adoption, data transmission and operation tend to become more tedious. Figure no.4 elaborates that data protection and privacy is the most crucial factor among all [11].

Figure 4 elaborates the overall impacts of security concerns. The major security challenge is about data leakage and data segregation because cloud is a shared pool of resources. The next bigger challenge is to prevent the data leakage.

Fig. 4. Data Security Challenges.

To cope with the above challenges, following are some solutions which needs to be considered while considering about cloud computing security challenges;

VII. Data Encryption

Encryption is said to be a better approach regarding data security. Data should be encrypted before sending it to cloud. Data owner can permit some particular members to have access to that data [11]. The file or data being sent to cloud should be encrypted first then before storing it on cloud it should be again encrypted by the cloud provider; the process is known as multistage encryption. It has been observed that combination of different encryption algorithms provides better encryption on data. Experimental results show that RSA+IDEA gives the higher performance of encryption in securing the data [12].

VIII. Legal Jurisdiction

When it comes to understand and analyze the legal jurisdiction of cloud computing, the very basic aspects of cloud environment complicate the data protection. e.g presence of internet, virtualization, dynamically distributed data, multinational elements. Consumers, normally do not know that where their data resides in cloud. For example, a client from india may be using a server deployed in US, using an application which has been developed in japan and storing his crucial data at a data center which is physically located in Switzerland [13]. So, the resource allocated to the consumers should be marked to make sure that data is segregated.

IX. Distributed Denial Of Service (DDoS)

Distributed Denial of service is a kind of attack in which attacker creates some zombie machine by infecting the machine over the internet. Then these infected machines are used to attack on victim. When attacks/traffic from so many infected machines are directed towards one victim, its resources like CPU, bandwidth and memory starts getting exhausted and that particular resource becomes unavailable for consumers. To cope with this Deepali [14] has introduced a layer named as fog layer which sits in between cloud server and user. All the requests made to server are filtered through this fog layer and DDOS attacks get minimized.

X. Digital Signature

Digital signature is powerful tool for securing data in cloud computing. Mr. Prashant Rewagad [15] has proposed a solution using digital signature to secure data along with Diffie Hellman key exchange with AES encryption algorithm. Diffie Hellman key exchange facility marks it useless if the key is hacked in transmission because it is useless without private key of user, which is confined to legitimate user only. This three way mechanism which is proposed in this paper makes it harder to hack security system, therefore, protecting the data that resides in cloud.

XI. Conclusion

This paper gave the overview of cloud computing, its various security aspects and keys factors which are affecting the cloud security. Cloud consumer and provider should be sure that their cloud is fully protected. Cloud computing is growing in every industry but it suffers from certain issues regarding security and protection which are a hurdle in its adoption widely. Solutions to these problems have been suggested which can be used for better performance of cloud service.

References

[1] Cloud Computing Definition. 2011; Available from: https://www.nist.gov/newsevents/news/2011/10/final-version-nist-cloud-computing-definition-published.

[2] Mell, P., Grance, T, The NIST definition of Cloud Computing, version 15 National Institute of Standards and Technology (NIST), Information Technology Laboratory. October, 2009.

[3] what is cloud computing? how it works.

[4] Agarwal, A., S. Siddharth, and P. Bansal. Evolution of cloud computing and related security concerns. in 2016 Symposium on Colossal Data Analysis and Networking (CDAN). 2016.

[5] Garg, P., S. Goel, and A. Sharma. Security techniques for cloud computing environment. in 2017. International Conference on Computing, Communication and Automation (ICCCA). 2017.

[6] Sharma, P.K., et al. Issues and challenges of data security in a cloud computing environment. in 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON). 2017.

[7] Amara, N., H. Zhiqui, and A. Ali. Cloud Computing Security Threats and Attacks with Their Mitigation Techniques. in 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). 2017.

[8] Ahmadi, S.A.a.M.R., A Novel Data Security Model for Cloud Computing. International Journal of Engineering and Technology, 2012. 4(3).

[9] Jahangeer Qadiree, M.I.M., Solutions of Cloud Computing Security Issues. International Journal of Computer Science Trends and Technology (IJCST), April 2016. 4(2).

[10] Shahzad, F., State-of-the-art Survey on Cloud Computing Security Challenges, Approaches and Solutions. Procedia Computer Science, 2014. 37: p. 357-362.

[11] Rao, R.V. and K. Selvamani, Data Security Challenges and Its Solutions in Cloud Computing. Procedia Computer Science, 2015. 48: p. 204-209.

[12] Chennam, K.K., L. Muddana, and R.K. Aluvalu. Performance analysis of various encryption algorithms for usage in multistage encryption for securing data in cloud. in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). 2017.

[13] Sony, R., K. Sri, and D. Bhukya, Data Protection and Cloud Computing: a Jurisdictional Aspect. 2013. 81-91.

[14] Deepali and K. Bhushan. DDoS attack defense framework for cloud using fog computing. in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). 2017.

[15] Rewagad, P. and Y. Pawar. Use of Digital Signature with Diffie Hellman Key Exchange and AES Encryption Algorithm to Enhance Data Security in Cloud Computing. in 2013 International Conference on Communication Systems and Network Technologies. 2013.

Self Interference Cancellation in Co-Time-Co-Frequency Full Duplex Cellular Communication

Sajjad Ali Memon[1]
Assistant Professor in the Department of Telecommunication Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan

Faisal Ahmed Dahri[2]
Postgraduate student in Department of Telecommunication Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan

Farzana Rauf Abro[3]
Associate Professor in the Department of Electronics Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan

Faisal Karim Shaikh[4]
Professor in the Department of Telecommunication Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan

***Abstract*—The performance of co-time co-frequency full duplex (CCDF) communication systems is limited by the self-interference (SI), which is the result of using the same frequency for transmission and reception. However, current communication systems use separate frequencies for transmission and reception, respectively. Therefore, SI is an important issue to be fixed for future-generation systems. As the radio frequency (RF) spectrum is very scared and a CCDF system has the potential to reduce the current spectrum use by half. In this paper, a CCDF communication system is modeled and a combination of RF and digital cancellations is used to mitigate the SI. The simulation results reveal that the proposed combination of RF and digital cancellation achieve the bit-error-rate of 10^{-11} at an interference-to-signal ratio of 10 dB, which is satisfying value for CCDF communication. The achieved efficiency of the proposed system is 13 bits/sec/Hz at a signal-to-noise ratio of 50 dB. The antenna separation of 35 dB is considered for the proposed model to keep the data loss as minimum as possible. The performance can be improved further by increasing digital-to-analog converter bits but with added complexity.**

Keywords—Cellular; Co-Time Co-Frequency full duplex (CCFD); self-Interference cancellation (SIC); communication system

I. Introduction

Currently, wireless frequency spectrum turns out to be crowded and costly because of the development of bandwidth-demanding applications used in telecommunication. Thus, it is needed to present new advancements that can improve both spectrum efficiency and transmission rates. Full-duplex (FD) communication is a potential candidate technology for enhancing the spectral efficiency of next-generation wireless communication networks; such as fifth generation (5G) networks. FD cellular communication systems are also known as co-frequency co-time full duplex (CCFD), which transmit and receive signals simultaneously at the same frequency and time. It improves the system performance in terms of throughput and spectral efficiency. The signal strength of self-interference (SI) is stronger than that of the desired signal from the estimated node from 90 to 110 dB. Therefore, with a specific goal to make FD practically implementable, canceling or dropping the interference signal to the noise floor is necessary. In the present time, there is no available electromagnetic method which can bring down SI under the noise floor. Therefore, it is necessary to further suppress the SI at the baseband stage. With the use of separate antennas, significantly increases the self-interference cancellation (SIC) level, however it contains primary disadvantage, such as multiple antennas avoid from the dense assimilation of in-band FD systems because of the required physical distance between antennas. Therefore, the system performance is declined and causes an analog to digital conversion (ADC) saturation problem [1]. The authors have demonstrated a Lab view based algorithm to deal with SIC. The video stream was used as a type of data at the central frequency of 2.2 to 2.5 GHz with a bandwidth of 20 MHz. The ADC saturation was decreased but the overall performance of the system is not up to the mark in long-term evolution. The SIC is initiated and can be employed only on relay based node for communication. The mutual coupling model is used to cancel SIC. The power of interference was minimized to improve system performance [2]. Authors in [3] provide an overview of various concepts to decrease SIC. The measurements had been done to characterize the nature of SIC with a variety of techniques. The authors proposed the analog and digital cancellation methods but failed to provide the concluding technique. The comprehensive study is to assist in understanding and developing an appropriate cancellation technique. The researchers have mitigated the computational requirement with removing digital cancellation parameters and achieved 65% efficiency [4]. Ref. [5] personifies the potentiality, drawbacks and gaining of estimated methods. The error control coding plays a primary role in SIC to mitigate the interference. Selective filtering methods were designed to cater to the problem of SI [6]. The SIC level was controlled at some stage but not completely

This work was supported by the **Start-Up Research Grant Program** project number **1718** of the **Higher Education Commission** of **Pakistan**.

removed in all available techniques. Least mean square (LMS) algorithm and Rician channel were introduced as an interference channel with high powered in the line of sight direction. Simulations were carried out to confirm through Lab view software [7]. The receiver amplifier saturation and dynamic range were overlooked; these are considered as main aspects of canceling technique in passband signal. The researchers have extended the range of full duplex systems [8-11]. In present work, two antennas are required for full duplex communication; one side works as a transceiver and vice versa but the main challenge is generated through transmit antenna at the receiving antenna is addressed in such systems is a self-interference. After 2020, higher data rate requirement with the fastest growth in data communication devices and development in mobile communication turn the idea of researchers for developing new generation networks for communication; for instance 5G. Therefore, CCFD became a latent technology and prevalent research topic in a 5G cellular communication system. CCFD will play a significant part in spectrum resources and efficiency. Additionally, the SIC is a core problem of CCFD so that it carries great weight and a means to do research on SIC for next-generation wireless networks and propose amendments in current cellular communication. It is always done through half-duplex; the transmission and reception have been done on distinct slots when simultaneous communication is done through FD. To boost up the spectral efficiency of cellular communication is a vital piece of research in this domain.

In this work, we have proposed the nonlinear components and antenna separations with directive antennas which enable the SIC approach. The objective of this prominent work is to make an optimal gap between transmitting and receiving antenna so that we will be able to overcome the SIC.

II. Literature Review

In recent times, lots of researchers are wasting their energies on improving the management of the spectrum efficiently. There have been a sizzling issue and many research exercises on the FD have been done for the spectral efficiency enhancement in cellular communication systems. The demand for high-speed wireless communication is burgeoning due to the rise in various multimedia services and applications. The authors have investigated the FD technology transducer along with resource management solution for cellular communication. The progress is founded on SIC technique to trim down its impact and work commercially communication transceivers. The physical spacing is considered a simple way to achieve interference cancellation [12]. In mobile communication, the frequency offset allows the distortion in orthogonality of subcarriers which results in inter-carrier interference. This entailed and proposed the coefficient which reduces the effect of interfering carriers due to the channel frequency error [13]. The oscillator phase noise limits the possibility of SIC scheme for FD. The proposed methods are applicable to estimate and cancellation phase [14]. The proliferation of wireless communication links has greatly improved to acquire high-speed broadband communication services. The femto relay node based approaches were studied based on SIC. The results draw that during the transmission of wavelength code division multiple access signals, drastically decrease in 5 MHz including path loss and passive interference along with proper functioning of FD systems [15]. Double feed network approach is applied to mitigate the SIC and achieved 45-47 dB of SIC [16]. LMS algorithm used to overcome the echo cancellation and presented that the bandwidth of the desired data affects the system cancellation [17]. The investigators have studied simultaneous wireless information and power transfer for FD communication on the downlink. Multiple antennas were enabled for uplink and downlink communication. The presented form of results shows that their work is suitable only under restricted conditions when low data communication is needed [18]. The discrepancy between perceived spectrum and actual availability of spectrum by higher authorities of various approaches would be entertained to decrease the shortage of spectrum. Random signal architecture is persuaded to pick up the efficiency of networks due to digital domain signal cancellation [19]. The adaptive digital filters are utilized and it explores more about the low to medium ADC resolution in mixed digital cancellation. Cognitive radio needs more flexibility in spectrum than present radio systems. The authors claimed that we have exploited unutilized channel over multiple GHz of bandwidth spectrum. Antenna separation is presented to diminish the impact of SI [20]. Antenna design was modeled on monopole antennas in computer simulation software. The achieved results show the 50 dB of SI [21]. The base station and relays employ beamforming through an array of the antenna are considered for the broadband multi-cell system. The gain could be exploited for making a reduction in SI. It is experienced that with an increase in a number of antennas in picky separation may effect in SIC value [22]. The real-time and scalable digital cancellation methods were used to mitigate the SI. Video streamed with two different platforms at a central frequency of 20 MHz of bandwidth in long-term evolution infrastructure [1]. The compatibility and capability issues had been minimized for FD communication [23]. The energy harvested from the SIC model for FD communication. During the process of downlink transmission, energy harvested for FD communication [24]. Outcomes of paper [25] present the significant improvement in energy efficiency to allow recycling of the self-energy approach. Stochastic geometry was employed to analyze the FD wireless network throughput. The performance of half-duplex was optimized based on throughput and a mathematical model has been studied which highlight the suppression of SI and suitability of FD communication.

III. System Model

The simultaneous transmission of FD communication can be achieved through cancellation of SI. The SI signal is billion times stronger than the received signal. The initial concept of wireless communication is that radio frequency signals cannot be transmitted and received at the same time and frequency. The two nodes are attempting to transceiver simultaneously with co-time co-frequency. The interference signal is shown in the Fig. 1. Both the nodes have stronger signal at their own side from the transmitter. Such signals create interference which is known as SI. Theoretically, it is easy to solve but practically the receiver should have information of the transmitted signal and subtract it from the received signal after that decode the rest of the information signal. The model is a complete FD transceiver approach in terms of the component which enables to scrutinize the viability of communication channel. The block diagram of the FD is illustrated in Fig. 2. The simple architecture is considered due to its wider use in cellular communication. The digital cancellation and radio frequency cancellation module are used to process the linear operations. The ADC interface is deployed in the system to analyze the ADC bits impacts and detector input response was considered in this scenario for validation of system design. The assumption of noise figure at the receiver was discussed in form of the equation which is denoted by Frii's formula [26].

$$F_{rx} = F_{LNA} + \frac{F_{mixer} - 1}{G_{LNA}} + \frac{F_{VGA} - 1}{G_{LNA} G_{mixer}} \quad (1)$$

Where F denotes the noise factors of different components and G shows the gain of those factors. Antenna separation is a factor of RF cancellation because path losses were estimated based on the distance within the transceiver. The proper set of values of phase and amplitude would be taken for fixed delays may impact the multipath signals through coefficients and subtract interference signals from the received signal. There is a need for bulky attenuator with a fixed value to cancel required power levels for full duplex communication. The RF cancellation block requires the estimated path loss between antenna to ensure the power of reference and interference signals. The ADC interference and quantization noise can be estimated from equation 2. The signal to quantization noise ratio can be formulated as

$$SNR_{ADC} = 6.02b + 4.74 - PAPR \quad (2)$$

Numbers of bits are denoted by b and *peak-to-average power ratio* (*PAPR*) is the fixed value considered. The clipping of the signal should be avoided based on the full range and capabilities are used. The linear processing is enabled through radio frequency and digital cancellation stages. The transmitted signal from antenna attenuated from the power amplifier. The input ADC and data input are considered as an interface in the system. The ADC input is controlled by the automatic gain control component at a constant level. Cancellation signal has high power than the desired signal contains a reduced power level. The dynamic range of ADC bits is preserved would be actuated through SI to impact at the minimum level of the desired signal.

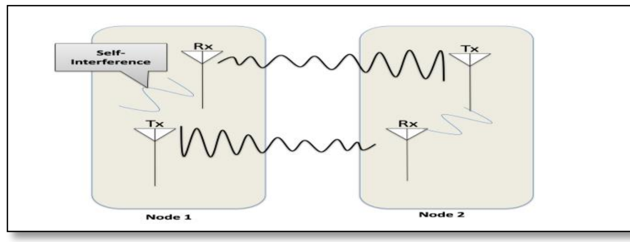


Fig. 1. The Diagram of SI from the Own Transmitter Device at the Receiver.

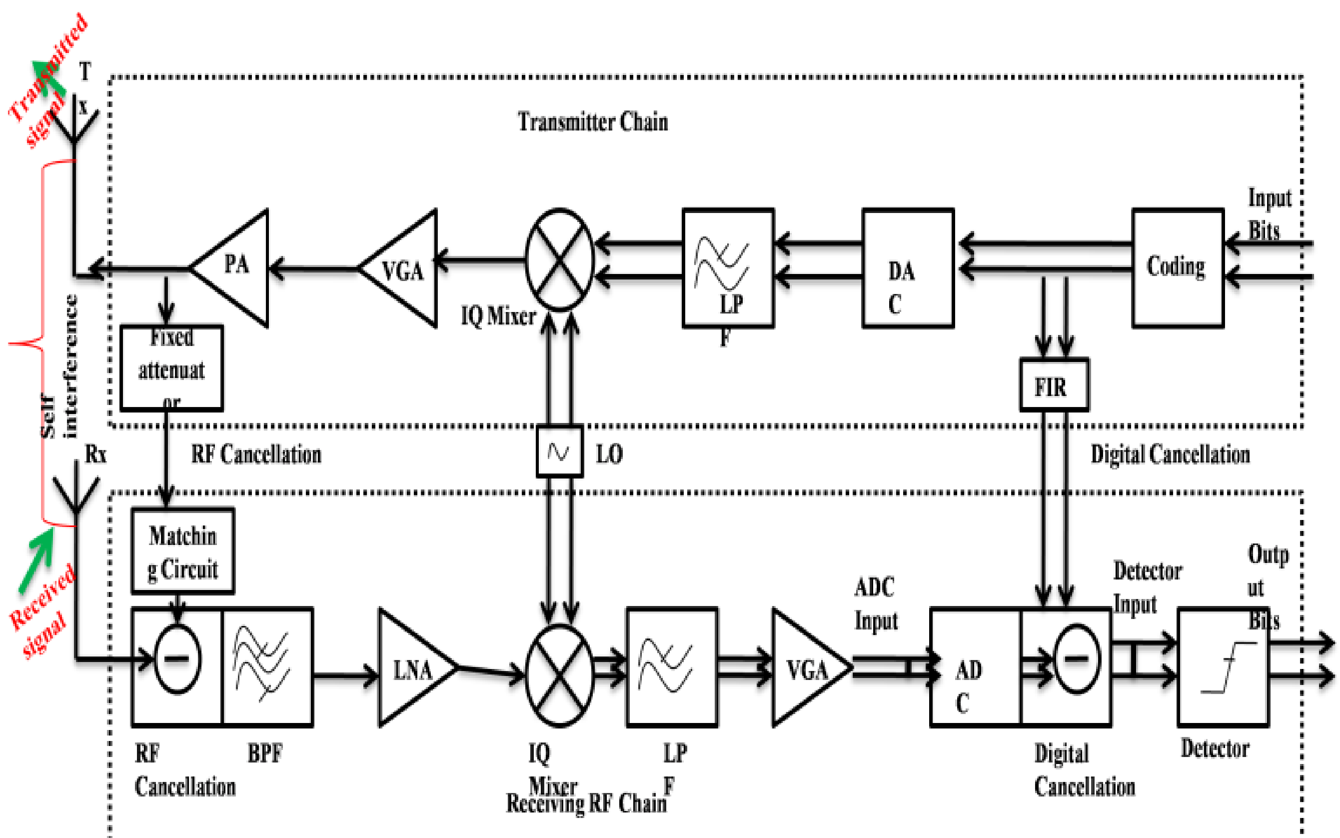


Fig. 2. The Simulated Design for FD Transceiver Framework Subjects to Desired and SI Signals [27].

IV. Simulation Results & Discussion

The section provides the details regarding the performance of simulated and proposed SIC scheme under different conditions. Matlab 2017a is used to simulate the system. The system is operated in FD mode, where simultaneous transmission and reception take place at the same time and frequency. The entire communication chain of transmitter and receiver is modeled to implement analog and digital cancellation. Fig. 3 presents the performance of the proposed design in terms of symbol error rate with the number of symbol to noise rate and theoretical symbol error rate (SER). The digital to analog conversion (DAC) impact on a different number of bits was analyzed for the SI system. Fig. 4 shows the analog to the digital impact of bits at the different range was investigated. In this system, SER values are optimal to reduce the SI. The increase in bits can provide better outcome as illustrated in the results. The presented forms of results are under assumptions that radio cancellation is considered as reference signal after the power amplifier. The power amplifier introduces nonlinear distortion in the RF cancellation of the transmitter chain. It is shown in the above figure that the number of DAC bits increased to enhance the performance while ADC bits are taken as fixed. It is noted that the result does not infer that more accuracy can be added to the system by increasing only the DAC bits. The word length of the system is dependent on the DAC bits and floating points are trimmed to set estimated constant quantization levels. The SI and desired channels were estimated based on fixed DAC bits with five training symbols. It is visible in Fig. 4 that performance is irrelevant as a contrast with theoretical. The performance improves with increasing the number of ADC bits and recommended that to quantify the bits in ADC is not a bottleneck. The overlapping is founded in an ADC bit sequence from 12 to 16. The lost bits increased as shown in Fig. 5 because of degradation in radio cancellation ability, it implies the SI power is high at the interface of ADC. The need for ADC is very much increased for achieving SI cancellation at the analog domain. It is observed that roughly 6 bits lost because of SI with greater transmit power of 25 dBm. This accentuates the way that in this situation SI is restricting factor for power. This show higher power is used which requires a number of ADC bits to reduce the loss of bits. The performance of bit error rate (BER) of the proposed scheme is presented in Fig. 6 in which nonlinear distortion beats the conventional schemes. The 10 dB of the gain of interference signal rate was accomplished with the proposed method. The *K* is a Rician channel factor at *K=3*, achieved better results and suitable for cellular communication. Fig. 7 depicts the spectral efficiency with no distortion suppression along half and full duplex results based on SNR values. The half-duplex system is restricted from the noises such as Gaussian; the SNR is characterized through the desired signal power divided by receiver Gaussian noise power. The better results of half and full duplex cellular communication can be achieved with high SNR values.

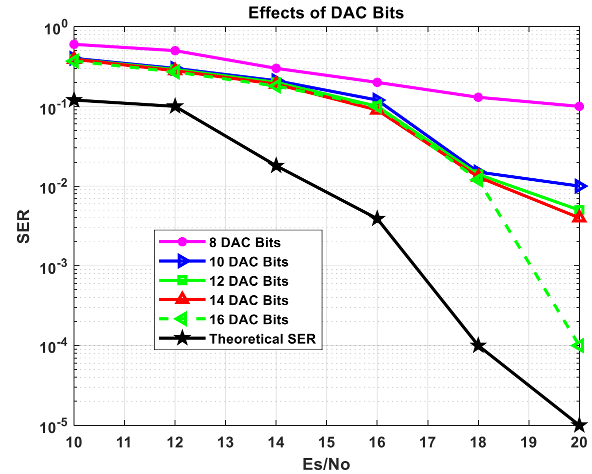


Fig. 3. Effects of Different DAC Bits on Symbol Error Rate for FD Communication.

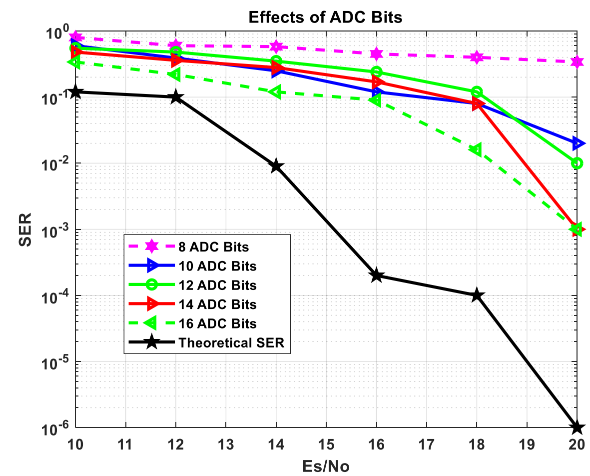


Fig. 4. Effects of Different ADC Bits on SER with a Fixed Value of DAC for FD Communication.

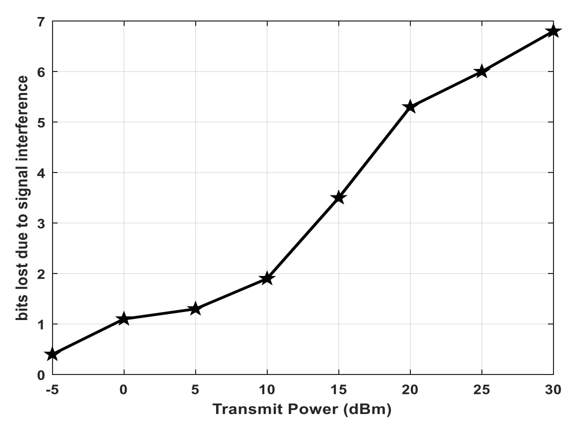


Fig. 5. Bits Lost Due to the SI Versus Transmit Power At Fixed Values of ADC with 35 dB Antenna Separation.

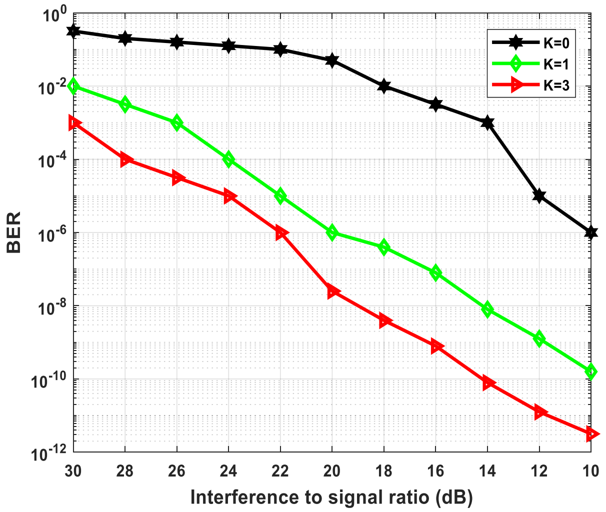


Fig. 6. BER Versus ISR with Different Rician Channel Order (*K=0, 1, 2*) for FD Communication.

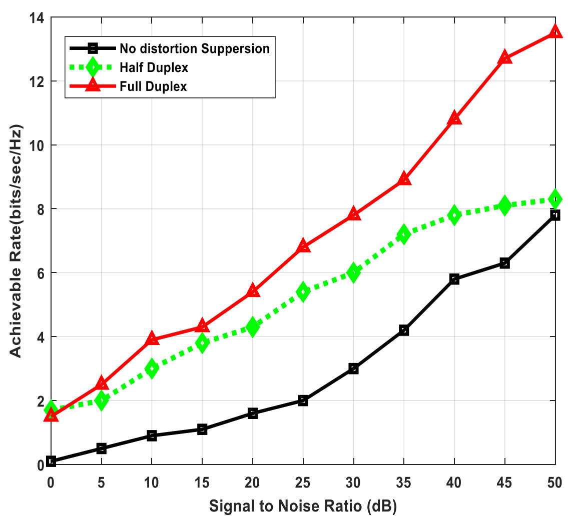


Fig. 7. Outcomes of Achievable Rate Versus SNR for a Half, No Distortion Suppression and FD Communication at Specific SI Signal Strength Values.

V. CONCLUSION

In this work, the SIC scheme is proposed for the CCFD communication system. This approach mitigates the SI and enables FD communication based on optimal and effective results. It is observed that high SNR would improve the performance of FD. The non-linearity impact is reduced and accumulated good outcomes from the simulated design. The system reliability is validated by comparing the spectral efficiency of FD and half duplex communication system. Quantization noise limits the tolerance in terms of transmitting power for the receiver ADC converter and digital cancellation. This issue is mitigated and achieved good results. The achieved results are improved as compared to existing works. The presented results signify that proposed method mitigate interference signal and compared to a conventional half-duplex system. To tune the SNR value, the performance of FD is maximized. The achievable rates of no distortion suppression, half duplex, and FD are 7, 8 and 13 bits/sec/Hz attained, respectively. The minimum BER is accomplished through the proposed method and performance is suitable for CCFD cellular communication.

REFERENCES

[1] Chu, Jian-Ya, Yuan-Te Liao, Chun-Yi Hung, and Terng-Yin Hsu. "Implementation of digital self-interference cancellation in LTE-based CCFD transmission." In Dependable and Secure Computing, 2017 IEEE Conference on, pp. 475-476. IEEE, 2017.

[2] Pawinee Meerasri, Peerapong Uthansakul, and Monthippa Uthansakul, "Self-Interference Cancellation-Based Mutual-Coupling Model for Full-Duplex Single-Channel MIMO Systems," International Journal of Antennas and Propagation, vol. 2014, Article ID 405487, 10 pages, 2014. https://doi.org/10.1155/2014/405487. .

[3] Le-Ngoc, Tho, and Ahmed Masmoudi. "Self-Interference-Cancellation in Full-Duplex Systems." In Full-Duplex Wireless Communications Systems, pp. 13-25. Springer, Cham, 2017.

[4] Korpi, Dani, Lauri Anttila, and Mikko Valkama. "Nonlinear self-interference cancellation in MIMO full-duplex transceivers under crosstalk." EURASIP Journal on Wireless Communications and Networking 2017.1 (2017): 24.

[5] Fu, Tong. Full-duplex Wireless Communications: Capacity Analysis, Low-density Parity-check Coding and Feasibility. Diss. McGill University Libraries, 2015.

[6] Öncel, Aksay Fatih. On Self-Interference Cancellation in Wireless Full duplex, Diss. Middle East Technical University, 2016.

[7] Koirala, Kumud. Adaptive Self-Interference Cancellation in Full-Duplex Radio. Tampere University of Technology (2016).

[8] Krier, John R., and Ian F. Akyildiz. "Active self-interference cancellation of passband signals using gradient descent." In Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on, pp. 1212-1216. IEEE, 2013.

[9] Li, Na, Weihong Zhu, and Haihua Han. "Digital interference cancellation in single channel, full duplex wireless communication." In Wireless Communications, Networking and Mobile Computing (WiCOM), 2012 8th International Conference on, pp. 1-4. IEEE, 2012.

[10] Shen, Ying, Juan Zhou, and Youxi Tang. "Digital Self-Interference Cancellation in Wireless Co-time and Co-frequency Full-Duplex System." Wireless Personal Communications 82, no. 4 (2015): 2557-2565.

[11] Li, Jiong, Hang Zhang, and Menglan Fan. "Digital self-interference cancellation based on independent component analysis for co-time co-frequency full-duplex communication systems." IEEE Access 5 (2017): 10222-10231.

[12] Kari Rikkinen, Alok Sethi, Visa Tapio , Laura González ,Cristina Lavín, Reinel Marante ,Björn Debaillie, Mina Mikhael, Hans Suys,Barend van Liempd, Mir Ghoraishi, Hassan Malik . "Full-Duplex Radios for Local Access – DUPLO" Ref. Ares(2015)2277009 - 2015.

[13] Zhao, Yuping, and S-G. Haggman. "Intercarrier interference self-cancellation scheme for OFDM mobile communication systems." IEEE Transactions on Communications 49, no. 7 (2001): 1185-1191.

[14] Ahmed, Elsayed, Ahmed M. Eltawil, and Ashutosh Sabharwal. "Self-interference cancellation with phase noise induced ICI suppression for full-duplex systems." In Global Communications Conference (GLOBECOM), 2013 IEEE, pp. 3384-3388. IEEE, 2013.

[15] Akyildiz, Ian F., Elias Chavarria-Reyes, David M. Gutierrez-Estevez, Ravikumar Balakrishnan, and John R. Krier. "Enabling next generation small cells through femtorelays." Physical Communication 9 (2013): pp. 1-15.

[16] An, Changyoung, and Heung-Gyoon Ryu. "Double Balanced Feed Network for the Self-Interference Cancellation in Full Duplex Communication System." Wireless Personal Communications 92, no. 4 (2017): 1599-1610.

[17] Avasarala, Sanjay. "Adaptive Baseband Interference Cancellation for Full Duplex Wireless Communication Systems". Diss. Arizona State University, 2016.

[18] Chen, Dong-Hua, and Yu-Cheng He. "Full-Duplex Secure Communications in Cellular Networks with Downlink Wireless Power Transfer." IEEE Transactions on Communications 66.1 (2018): 265-277.

[19] Yang, Jing. "Time domain interference cancellation for cognitive radios and future wireless systems". Diss. UC Berkeley, 2010.

[20] Jasim, A. A., K. M. Younus, A. Ali, K. H. Sayidmarie, A. Alhaddad, and R. A. Abd-Alhameed. "A simple self-interference cancellation technique for full duplex communication." In Internet Technologies and Applications (ITA), 2017, pp. 224-229. IEEE, 2017.

[21] Larsson, Peter, and Mikael Prytz. "MIMO on-frequency repeater with self-interference cancellation and mitigation." In Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th, pp. 1-5. IEEE, 2009.

[22] Zarbouti, Dimitra, George Tsoulos, and Georgia Athanasiadou. "Effects of antenna array characteristics on in-band full-duplex relays for broadband cellular communications." ICT Express 1, no. 3 (2015): 121-126.

[23] Sim, Min Soo, Kwang Soon Kim, and Chan-Byoung Chae. "Self-Interference Cancellation for LTE-Compatible Full-Duplex Systems." In 2018 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1-6. IEEE, 2018.

[24] Yadav, Animesh, Octavia A. Dobre, and H. Vincent Poor. "Is Self-Interference in Full-Duplex Communications a Foe or a Friend?." IEEE Signal Processing Letters 25, no. 7 (2018): 951-955.

[25] Z. Tong and M. Haenggi, "Throughput Analysis for Full-Duplex Wireless Networks With Imperfect Self-Interference Cancellation," in IEEE Transactions on Communications, vol. 63, no. 11, pp. 4490-4500, Nov. 2015.

[26] Gu, Qizheng. RF system design of transceivers for wireless communications. Springer Science & Business Media, 2005.

[27] Korpi, Dani, et al. "Full-duplex transceiver system calculations: Analysis of ADC and linearity challenges." IEEE Transactions on Wireless Communications 13.7 (2014): 3821-3836.

Semi Supervised Method for Detection of Ambiguous Word and Creation of Sense: Using WordNet

Sheikh Muhammad Saqib[1], Fazal Masud Kundi[2]
Institute of Computing and Information Technology Gomal University. Pakistan.

Asif Hassan Syed[3], Shakeel Ahmad[4]
Faculty of Computing and Information Technology in Rabigh (FCITR) King Abdul Aziz University (KAU) Jeddah Saudi Arabia

***Abstract*—Machine Translation, Information Retrieval and Knowledge Acquisition are the three main applications of Word Sense Disambiguation (WSD). The sense of a target word can be identified from a dictionary using a 'bag of words', i.e. neighbours of the target word. A target word has the same spelling of the word but with a different meaning, i.e. chair, light etc. In WSD, the key input sources are sentences and target words. But, instead of providing a target word, this should automatically be detected. If a sentence has more than one target word, then the filtration process will require further processing. In this study, the proposed framework, consisting of buzz words and query words has been developed to detect target words using the WordNet dictionary. Buzz words are defined as a 'bag-of-words' using POS-Tags, and query words are those words having multiple meanings. The proposed framework will endeavor to find the sense of the detected target word using its gloss and with examples containing buzz words. This is a semi-supervised approach because 266 words of multiple meanings have been labelled from various sources and used based on an unsupervised approach to detect the target word and sense (meaning). After experimenting on a dataset consisting of 300 hotel reviews, 100 % of the target words for each sentence were detected with 84 % related to the sense of each sentence or phrase.**

Keywords—Word sense disambiguation; machine translation; information retrieval and knowledge acquisition; target word; WordNet; bag of words

I. Introduction

Choosing the correct sense in a context is related to Word Sense Disambiguation (DWS) because most words have multiple meanings, i.e. the word "run" has 179 meanings of the word while the word "take" has 127 different definitions of the word [1]. WSD methods are usually classified into two types: knowledge-based and machine learning [2], [3]. Knowledge-based WSD systems exploit the information in a lexical knowledge base, such as WordNet and Wikipedia, to perform WSD. These approaches usually choose the sense with the definition most like the context of the ambiguous word, using textual overlap or using graph-based measures [4]. Machine learning approaches, also called corpus-based approaches, do not make use of any knowledge resources for disambiguation. These approaches range from supervised learning [5], in which a classifier is trained for each distinct word in a corpus of manually sense-annotated examples, to entirely unsupervised methods that cluster the occurrence of words, thereby inducing senses. Recent advances in WSD have significantly benefited from the availability of corpora annotated with word senses. Most accurate WSD systems to date exploit supervised methods which automatically learn cues useful for disambiguation from manually sense-annotated data [6], [7], [8].

In this study, WSD is categorized into two approaches:

- WSD-1: can be used to determine a summary of a sentence. However, in a sentence, there may be a word with more than one meaning, i.e. "date", "bass" where the sense of these words will be considered in a sentence by a device or application.
- WSD-2: WSD can be used to detect the semantics of a word in a sentence concerning the polarity, i.e. "his work is unpredictable". Here the word "unpredictable" is a negative word, but in this instance, it will be considered as positive.

In this study, work is focused on WSD-1. There has been quite a lot of work conducted on WSD-1 by other researchers. However, in this study the target word has already been provided, i.e. "Sit on a chair", "Take a seat on this chair", "The chair of the Math Department". These phrases reflect the meaning of the chair, as the word has multiple senses?. Here, the target word "chair" is used to determine that the word chair means furniture or person [9]. And "I find a switch for the light" or, "I do like to eat something light", where the target word is portrayed as "light". Therefore, the sense may be viewed as "shine" or "weight" respectively [10]. Word detection has also been conducted in the work of [11], but these words could be considered as aspect or entity words on which an opinion has been given. "An electric guitar and bass player stand off to one side, not part of the scene". What could be the sense of "bass" in this sentence, [12] where the target word is given as "bass?". Therefore, there is need to filter the single word from obscure words or words with multiple-meanings. In this paper, we investigate, what could be the target word.

A manually created multiple meaning words list (MMWL) was developed through:

- The union of words taken from the English language [1];
- Multiple Meaning Words (100) grouped by the word "Grad" [13];
- Multiple meaning word list of 200 words [14], [15];

- Easy vocabulary words [16];
- Speech therapy ideas [15]; and
- Read words in context [17].

The manually developed MMWL contained 266 words with multiple meanings. The work has only used candidate definition/gloss [18], [19] but sense could also be detected from the examples to improve the accuracy because sometimes a ‘bag of words’ is not present in the definition/gloss of the target word.

The following contributions in this study are as follows:

- We propose to develop a method to filter a target word (sense required) from multiple ambiguous words with the help of using buzz words and query words using a lexicon of multiple meaning words list MMWL; and
- Generate a correct sense of target words with the help of buzz words using gloss and examples of target words from the lexicon of WordNet.

II. Related Work

A word in a sentence can be expanded by relating it to other words in a sentence to determine the actual meaning of the word. This is essential because the majority of research studies to date, have investigated opinion words [20], [21] or examined text mining through the creation of summaries of a given document [22], [23] in different languages such as Arabic [24], [24] and Chinese [25], [26]. This is so that a document or sentence can easily be understood by users as well as by intelligent machines. Automatic summary generation procedures have faced many problems including WSD words sense disambiguation. WSD also involves natural language processing applications [10]. For example, human intelligence can automatically sense and detect the meaning of a word from examining a sentence. However, in the field of artificial intelligence, efforts are continuing to be made to understand a sentence from the correct dimension or aspect given that a single word may have multiple meanings.

The MeSH-based disambiguation method, considers the meaning of a target word as the same throughout a document and the word tends to have the same meaning when used in the same collocation using MeSH which consists of words from different domains with the precision of 0.5841 [27]. Automatic disambiguated words on Wikipedia have several limitations due to the small sample size and a large number of fine senses found in WordNet [18]. In a study, supervised WSD [28] determined the sense value 2 and 3 from 57 target words. In a separate study, in [29] the determined sense of the target word was found by using three left, and three right words from the target word. This performed well only where the supporting words were present at the front and at the back of the target word. Instead of taking left, right words, the authors of [30] used a ‘bag of words’ from the sentence which were neighbours of a target word to identify or determine a binary vector. Also, the sense of the word can be detected if there are dependent words in the sentence near the target word. Many aspects of evaluating sense have been standardised through the efforts of SENSEVAL and SEMEVAL. This framework provides a shared task along with training and testing materials with sense inventories for all-words and lexical sample tasks in a variety of languages [31]. A relatively small set of training examples (seed sets) are identified in the framework to represent sense. Sense clusters are then generated through the addition of most similar words to the seed set elements. The most similar sense cluster to the input text context are then considered as the sense of the target word [32]. To address the limitation of the failed supervised scenario, studies have progressed on the kernel methods for automatic WSD using four target words: interest, line, hard and serve [6].The original algorithm based on glosses was found in traditional dictionaries such as the Oxford dictionary where the definition, or gloss, of each sense of a word in a phrase, is compared to the glosses of every other word in the phrase. A word is assigned the sense whose gloss shares the largest number of words in common with the glosses of the other words. The authors of [33] did not use examples of the word in the WordNet dictionary, but instead, used Lesk’s basic approach to take advantage of the highly interconnected set of relations among synonyms that WordNet offers by providing a target word. Besides the confusions in WSD, there many difficulties in handling these using the supervised and unsupervised methods. Work in [34] determined that supervised methods are the optimal predictors of WSD difficulties, but are limited by their dependence on labelled training data in different domain types such as bionadical [35], [36]. The unsupervised method performed well in some situations and can be applied more broadly [37], [38]. The accuracy of the unsupervised WSD algorithm is lower than its alternative supervised algorithm [39]. Word sense can also be detected from different sentences using latent semantic indexing by providing a query as the target word [40]. WSD is not only used in document clustering [19] but is also used in many applications that are based on artificial intelligence of a natural language. Work on WSD using the English language is progressing and is also being used in other languages such as Hindi, Hebrew, Russian and Tatar [41], [42], [43]. The application of WSD has not only been applied to text but also to images for determining the correct sense from a picture [44].

III. Proposed Methodology

In this work, there are two major tasks performed. First, the target word will be detected from within the sentence, and secondly, the sense of that word will be generated.

A. Detection of the Target Word

Filtered chunks (without) stop words will first be compared with MMWL to locate target words. Target words are defined as those words having multiple meanings. In Figure 1, tokens (T1, T2, T3, T4, T5, T6, T7, T8, T9, T10) are filtered as a filtered word (T2, T4, T5, T6, T7, T8, T9, T10) because T1 and T3 are stop words. These can be identified using Equations (1), (2) and (3).

$$S = \bigcup_{x=1}^{n} S_x \tag{1}$$

$$T(x) = \bigcup_{i=1}^{n} T_i \tag{2}$$

$$FT(x) = \bigcup_{i=1}^{n} \{T_i\ , \quad if\ T_i \notin SW \tag{3}$$

where x = 1, 2, 3…n, SW means stop words, S represents the total number of sentences, T(x) represents the tokens of the x^{th} sentence, and FT(x) represents the filtered tokens of the x^{th} sentence.

1) Buzz words: Buzz words are words that are adjectives, nouns, verbs and adverbs because their occurrence in sentences relate to the concept/explanation of the target word (identified in the next phases). Buzz words can either come from a list named 'BuzzTagList' containing: JJ (Adjective), JJR (Adjective, comparative), JJS (Adjective, superlative), NN (Noun, singular or mass), NNS (Noun, plural), NNP (Proper noun, singular), NNPS (Proper noun, plural), RB (Adverb), RBR (Adverb, comparative), RBS (Adverb, superlative), RP (Particle), VB (Verb, base form), VBD (Verb, past tense), VBG (Verb, gerund or present participle), VBN (Verb, past participle), VBP (Verb, non-3rd person singular present), VBZ (Verb, 3rd person singular present), or the buzz words can be extracted using Equation 4:

$$B_{words}(x) = \bigcup_{i=1}^{n} \begin{cases} FT(x)_i\ , & if\ (FT(x)_i) \in BuzzTagList \\ \end{cases} \quad (4)$$

Where x = 1, 2, 3…n, $FT(x)_i$ means the i^{th} token of the x^{th} sentence considered as B_{words} (buzz words) if it belongs to a noun, adjective or verb.

2) Query words: Query words are words having multiple meanings and can be obtained by comparing each filtered token with the manually created multiple meaning words list (MMWL). In Figure 1, suppose T2 and T8 are present in the MMWL, using Equation 5, we can find gloss and examples of each query word:

$$Q_{words}(x) = \bigcup_{i=1}^{n} \begin{cases} FT(x)_i\ , & if\ (FT(x)_i) \in (MMWL) \\ \end{cases} \quad (5)$$

Where x = 1, 2, 3…n, $FT(x)_i$ means ith token of the x^{th} sentence considered as Q_{words} (query words), if it belongs to a multiple meaning words List (MMWL).

3) Query strings: All query words have been created; so now we can easily locate the gloss/definitions of query words using the WordNet dictionary. As shown in the WordNet dictionary, a word can have multiple definitions, with each definition having multiple examples [45], [46], [47], [48]. By concatenating all definitions and examples (from each definition) this can be considered as a query string. In Figure 1, string-1, string-2 are query strings of T2 and T8 because there are two query words. All query strings from all query words can be created using Equation 6.

$$Q_{string}(x)_i = \bigcup_{i=1}^{n} \ Gloss_{WND}(Q_{words}(x)_i\ U\ \bigcup_{j=1}^{n}\ Examples_{WND}(Q_{words}(x)_{ij})) \quad (6)$$

where x = 1, 2, 3…n, will determine a complete string of each query word $Q_{string}(x)_i$ containing all glossaries of query words $Gloss_{WND}(Q_{words}(x)_i\ )$ and all examples $Examples_{WND}(Q_{words}(x)_{ij})$ of each gloss from the x^{th} sentence using synsets (sets of synonyms) found in the WordNet dictionary.

4) Frequency of buzz words from the query string: Next, the occurrence (frequency) of each buzz word from all query strings will be determined and summed. In Figure 1, T5 and T9 are those buzz words which do not belong to any query words and F1 and F2, are the frequencies of T5 and T9 in string1 respectively. Sum1 is the sum of F1 and F2, F3 and F4 are the frequencies of T5 and T9 in string2 respectively, and Sum2 is the sum of F3 and F4. These sums can be found by applying Equation 7:

$$BQ_{WF}(x)_i = Sumfrequency(\bigcup_{i=1}^{n}(Frequency\ of\ B_{words}i\ )_{At}\ \bigcup_{j=1}^{n}\ \ (Q_{string}(x)_j)) \ (7$$

Where x = 1, 2, 3…n, will determine the total number of frequencies of each buzz word $BQ_{WF}(x)_i$ from all query strings $Q_{string}(x)_j$ of xth sentence. Type equation here.

5) Target word: The query word of the greater sum from the query string will be considered as the target word. Suppose in Figure 1, sum1 is greater than sum2. As sum1 is generated from string1, and string1 is generated from query word T2, T2 can, therefore, be considered as the target word using Equation 8:

$$T_{words}(x) = \bigcup_{i=1}^{n} \begin{cases} (Q_{word}(x)_i\ , & if\quad Large\ (BQ_{WF}(x)_i) \\ \end{cases} \quad (8)$$

Where Large is a function to determine the largest sum and Query word $Q_{word}(x)_i$ can be considered as the target word $T_{words}(x)$ from xth sentence.

B. Generation of the Sense/Concept

First, a string of target word will be generated from the concatenation of all definitions and its examples. In Figure 2, T2 is a target word where a string of T2 is generated using Equation 9.

$$T_{string}(x) \quad = Gloss_{WND}(T_{word}(x))\ \ Union\ Examples_{WND}(T_{word}(x)) \quad (9)$$

where x = 1, 2, 3…n, will determine a complete string of query words $Q_{string}(x)$ containing all glossaries of the query word $Gloss_{WND}(T_{word}(x))$ and all examples $Examples_{WND}(T_{word}(x))$ of each gloss from xth sentence using synset in the WordNet dictionary.

Now the sense/concept will be generated by creating a substring from an already generated string (from the target word). This sense/concept contains all sentences which contain all buzz words except target words. The entire process can be conducted using Equation 10.

$$Concept_{Tword}(x) \begin{cases} (T_{string}(x)\ \ ,\ \ \bigcup_{i=1}^{n}\ \ if\ \ (B_{words}(x)_i \\ \qquad (\notin\ \ (T_{word}(x))\ \in\ (T_{string}(x)\ \ ) \\ \end{cases} \quad (10)$$

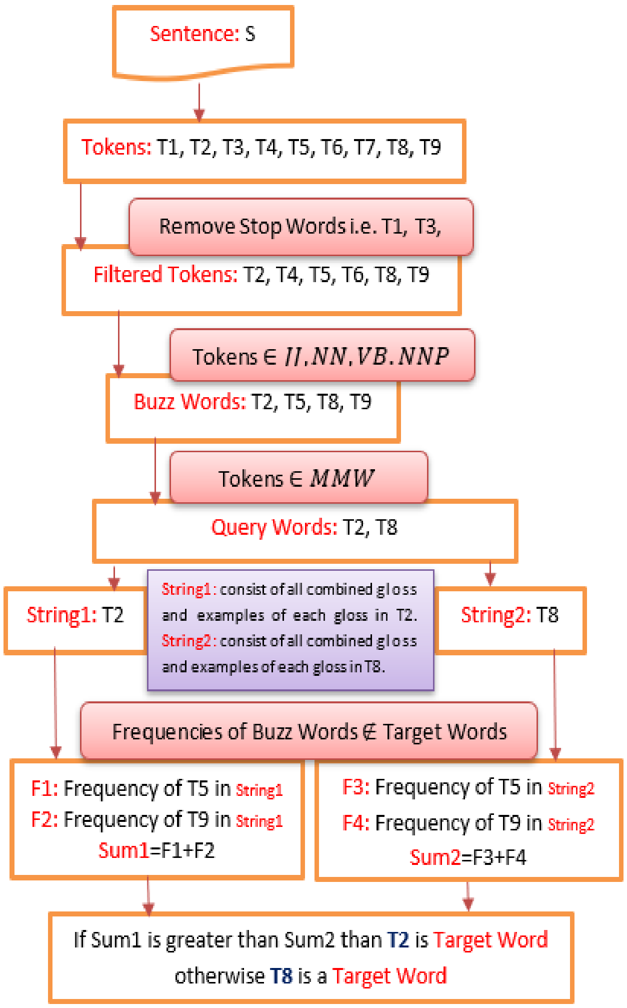


Fig. 1. Diagram of Proposed Framework for Detection of Target Word.

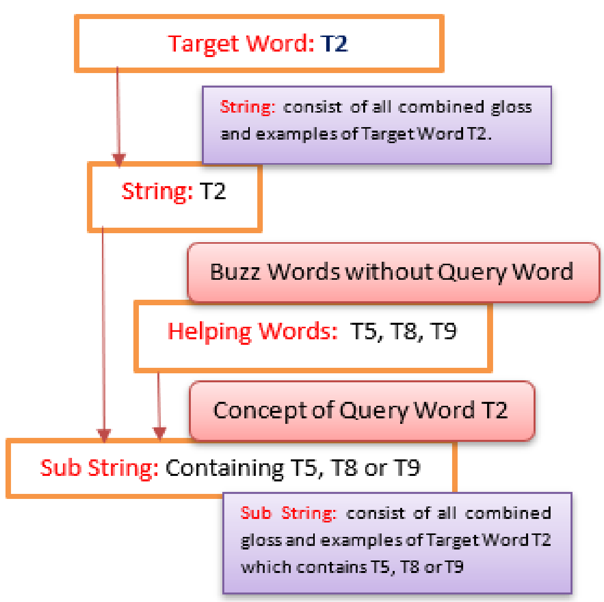


Fig. 2. Diagram of Proposed Framework for Detection of Sense of Target Word.

Where x = 1, 2, 3…n, will determine a $Concept_{Tword}(x)$ concept of the target word in the x[th] sentence. This concept consists of those glossary examples of the query word $T_{string}(x)$ containing buzz words $B_{words}(x)_i$.

IV. Samples Based on Proposed Work

In Table 2, the sentence "The researchers said the worms spend part of their life cycle in such fish as Pacific salmon and striped bass and Pacific rockfish or snapper".

By progressing through the following steps to identify Tags; Buzz Words; Query Words; strings of Query Words; the Frequency of buzz words from the strings; the sum of the Frequency of buzz words with respect to strings; and the largest Query Word as a Target Word, the target word "bass" is determined (based on the initial methodology).

Similarly, in Table 3, there is a further example "Sweet date can be used as the last course of a meal", where the target word "date" has been identified by the initial part of the proposed methodology.

TABLE I. Analysis of Example-1

Query Words ['part', 'bass']	**Buzz Words**									
	'researcher'	'worm'	'life'	'fish'	'pacific'	'salmon'	'bass'	'rockfish'	'snapper'	Sum
String1: from ['part']	0	0	0	0	0	0	0	0	0	0
String2: from ['bass']	0	0	0	3	0	0	0	0	0	3

Table 1 contains the analysis of example-1, where the frequency of each buzz word has been determined from the strings of query words ["part", "bass"]. The sum from string2 is 3, i.e. the largest, therefore, the target word will be "bass".

The sense/concept will be generated from those definitions and examples of the target word belonging to helping words (i.e. all buzz words without a target word). From Table 2, a target word was "bass", and in Table 4, the concept of "bass" was generated related to the sentence. From Table 3, a target word was "date", and in Table 5, the concept of "date" was generated related to the sentence.

TABLE II. SOLVED EXAMPLE-1 TO DETECT THE TARGET WORD

Sentence	The researchers said the worms spend part of their life cycle in such fish as Pacific salmon and striped bass and Pacific rockfish or snapper
Tag	[('The', 'DT'), ('researchers', 'NNS'), ('said', 'VBD'), ('the', 'DT'), ('worms', 'NNS'), ('spend', 'VBP'), ('part', 'NN'), ('of', 'IN'), ('their', 'PRP$'), ('life', 'NN'), ('cycle', 'NN'), ('in', 'IN'), ('such', 'JJ'), ('fish', 'JJ'), ('as', 'IN'), ('Pacific', 'NNP'), ('salmon', 'NN'), ('and', 'CC'), ('striped', 'VBD'), ('bass', 'NN'), ('and', 'CC'), ('Pacific', 'NNP'), ('rockfish', 'NN'), ('or', 'CC'), ('snapper', 'NN')]
Buzz Words	['researcher', 'worm', 'part', 'life', 'cycle', 'fish', 'pacific', 'salmon', 'bass', 'pacific', 'rockfish', 'snapper']
Query Words	['part', 'bass']
Strings of ['part', 'bass']	String1, String2
Frequencies of buzz words from strings	researcher:0 in String1, worm:0 in String1, life:0 in String1, cycle:0 in String1, fish:0 in String1, pacific:0 in String1, salmon:0 in String1, pacific:0 in String1, rockfish:0 in String1, snapper:0 in String1, researcher:0 in String2, worm:0 in String2, life:0 in String2, cycle:0 in String2, fish:3 in String2, pacific:0 in String2, salmon:0 in String2, pacific:0 in String2, rockfish:0 in String2, snapper:0 in String2,
Sum of frequencies of buzz words with respect to strings	Sum String1=0 , Sum String2=3 ,
Largest Target Word as Query Word	Bass

TABLE III. SOLVED EXAMPLE-2 TO DETECT THE TARGET WORD

Sentence	sweet date can be used as last course of meal.
Tag	[('sweet', 'NN'), ('date', 'NN'), ('can', 'MD'), ('be', 'VB'), ('used', 'VBN'), ('as', 'IN'), ('last', 'JJ'), ('course', 'NN'), ('of', 'IN'), ('meal', 'NN'), ('.', '.')]
Buzz Words	['sweet', 'date', 'last', 'course', 'meal']
Query Words	['date', 'last']
Strings of ['date', 'last']	String1, String2
Frequencies of buzz words from strings	sweet:2 in String1, course:0 in String1, meal:0 in String1, sweet:0 in String2, course:0 in String2, meal:0 in String2,
Sum of frequencies of buzz words with respect to strings	Sum String1=2 , Sum String2=0 ,
Largest Target Word as Target Word	Date

TABLE IV. CONCEPT OF DETECTED TARGET WORD IN EXAMPLE-1

Query Word	Bass
Buzz Words	['researcher', 'worm', 'part', 'life', 'cycle', 'fish', 'pacific', 'salmon', 'bass', 'pacific', 'rockfish', 'snapper']
Helping Words	['researcher', 'worm', 'part', 'life', 'cycle', 'fish', 'pacific', 'salmon', 'pacific', 'rockfish', 'snapper']
Concept of bass	[u'part : Gloss:the lowest part of the musical range', u'part : Gloss:the lowest part in polyphonic music', u'fish : Gloss:the lean flesh of a saltwater fish of the family Serranidae', u'fish : Gloss:any of various North American freshwater fish with lean flesh (especially of the genus Micropterus)', u'fish : Gloss:nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes']

TABLE V. CONCEPT OF DETECTED TARGET WORD IN EXAMPLE-2

Target Word	Date
Buzz Words	['sweet', 'date', 'last', 'course', 'meal']
Helping Words	['sweet', 'last', 'course', 'meal']
Sense/Concept of date	[u'sweet : Gloss: sweet edible fruit of the date palm with a single long woody seed']

V. RESULTS AND DISCUSSIONS

The evaluation strategy of WSD is based on the correctness of sense selection of an ambiguous word invoked in a context according to human judgment.

A. Dataset Preparation

In preparing the dataset, the consisting of 300 hotel reviews and MMWL (as previously defined). Approximately 66 % of the contexts were selected from hotel reviews containing ambiguous words from the MMWL for this purpose. The sample listing of the said datasets is presented in Table 6, where S1 has multiple meanings of the word "chair", S2 has "class", and S3 has "brand". These words are also listed in the MMWL.

TABLE VI. SAMPLE SENTENCES FROM THE PREPARED DATASET

S1	"Last year the chair of the food Department is retired"
S2	"The stay in the hotel was awsome. As a flight attendant, I see a lot of high class hotels and also know their service"
S3	"Diazepam is an example of the chemical (generic) name of a sedative. It is marketed by some companies under its generic name and by other companies under brand names such as Valium or Vazepam."

The proposed framework develops a sense of detected target words using the number of frequencies (occurrences) of buzz words from the query words. The work in this study identified 100 % of the target words (ambiguous words) from 66 % (containing vague words) of the context relating to hotel reviews, and 84 % concept/sense was generated from the dataset as shown in Table-7.

TABLE VII. DETECTED TARGET WORDS AND THEIR SENSES

Total Sentence	106
Sentences with No Query Words (Not Belongs to MMWL)	35
Filtered Sentence	71 (66%)
Detected Target Words out of 66 %	71 (100%)
Not Detected Concept from 66%	11 (15%)
Detected Sense	60 (84%)

This study was based on WSD-1, where only the WordNet dictionary was used. A sense detection of 84 %, was achieved by combining several other dictionaries, i.e., the Oxford dictionary, where the accuracy increased. Because the string of query words has been generated from its definition and examples, occasionally, the definition of a word could not be found in WordNet. The MMWL consisted of 266 obscure words. Updating the list would be useful for the remaining context given that buzz words can be generated from sentences, but if the buzz word is not in the MMWL then the number of query words is zero. Target words are reliant upon query words, and this is the reason why the sense cannot be generated.

VI. SUGGESTIONS FOR FUTURE WORK

In consideration of future work, if there is only a single query word, then there is no need for further processing as this query word can be considered as the target word. In proposed work, if there is a list of query words, then further processing will be carried out to detect a target word and the largest sum of frequencies, of buzz words from strings (a separate string for each query word) will identify the target word. If there is more than one sum of frequencies with the same score, then, this would be a viable case to perform future work to thereby calculate the distance of each buzz word from all query words to detect the target word. Also, additional work to detect the polarity sense of a word based on an opinion within in a sentence as defined in WSD-2 would be useful for future investigation.

REFERENCES

[1] S. Streat, "10 English Words with the most multiple meanings – Part 1," Succeed in English with Business, 2014. .

[2] S. Raviv, A., Markovitch, "Concept-based approach to word sense disambiguation," Proc. 26th AAAI Conf. Artif. Intell. Toronto, Canada, pp. 807–813, 2012.

[3] A. Rios, L. Mascarell, and R. Sennrich, "Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings," Proc. Conf. Mach. Transl. (WMT, vol. 1, pp. 11–19, 2017.

[4] K. H. Abualhaija, S., Zimmermann, "A novel method inspired by bee colony optimization for solving word sense disambiguation.," Swarm Evol. Comput. 27, pp. 188–195, 2016.

[5] J. C.-C. and R. N. D. Alessandro Raganato, "Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison," Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguist., vol. 1, pp. 99–110, 2017.

[6] T. Wang, W. Li, F. Liu, and J. Hua, "Engineering Applications of Artificial Intelligence Sprinkled semantic diffusion kernel for word sense disambiguation," Eng. Appl. Artif. Intell., vol. 64, no. May, pp. 43–51, 2017.

[7] A. Wawer and J. Kazimierza, "Supervised and Unsupervised Word Sense Disambiguation on Word Embedding Vectors of Unambiguous Synonyms," 120 Proc. 1st Work. Sense, Concept Entity Represent. their Appl., pp. 120–125, 2017.

[8] R. Paper, S. G. Rawat, M. B. Chandak, and N. A. Chavan, "An Approach for Improving Accuracy of Machine Translation using WSD and GIZA," Int. J. Comput. Sci. Eng., vol. 5, no. 10, pp. 256–259, 2017.

[9] D. De Cao and R. Basili, "Word Sense Disambiguation," http://www.d.umn.edu/˜ tpederse/Tutorials/ADVANCES-IN-WSD-AAAI-2005.ppt, 2009.

[10] C. Engineering, "A Survey on Word Sense Disambiguation Used In NLP," Int. J. Innov. Res. Comput. Commun. Eng., vol. 5, no. 3, pp. 5116–5117, 2017.

[11] S. M. Saqib, T. Naeem, and K. Mahmood, "Extraction of key / title / aspect words from document using wordnet," Accent. Trans. Inf. Secur., vol. 1, no. 3, pp. 36–43, 2016.

[12] A. Fulmari and M. Chandak, "An Approach for Word Sense Disambiguation using modified Naïve Bayes Classifier," Int. J. Innov. Res. Comput. Commun. Eng., vol. 2, no. 4, pp. 3867–3870, 2014.

[13] "100 Multiple Meaning Words Grouped By Grade," Home Speech Home. .

[14] "Free lesson plans and more for the school-based Speech Language Pathologist!" [Online]. Available: http://slplessonplans.com/files/mmwlist.pdf.

[15] R. Wanca, "Multiple Meaning Words List," M.S., CCC-SLP, 2012. .

[16] "15 English Vocabulary Words with Multiple Meanings." .

[17] G. 4 © S. T. R. Powerful Vocabulary for Reading Success: Teacher's Edition, "Multiple-Meaning Words to Know." .

[18] J. Sivakumar and A. Anthoniraj, "Automatic Word Sense Disambiguation Using Wikipedia," Sch. Comput. Sci. Eng. VIT Univ. Vellore-632014, 2007.

[19] K. Ċ. Pawe, M. Piasecki, J. K. Ě, and A. Indyka-piasecka, "Distributionally Extended Network-Based Word Sense Disambiguation in Semantic Clustering of Polish Texts," Int. Conf. Futur. Inf. Eng. ELSEVIER, vol. 10, pp. 38–44, 2014.

[20] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," Knowledge-Based Syst., vol. 89, pp. 14–46, 2015.

[21] S. N. Manke, "A Review on : Opinion Mining and Sentiment Analysis based on Natural Language Processing," vol. 109, no. 4, p. 8887, 2015.

[22] and J. R. Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, "Building a Sentiment Summarizer for Local Service Reviews," Proc. WWW-2008 Work. NLP Inf. Explos. Era., 2008.

[23] K. Lerman, K. Lerman, R. McDonald, and R. McDonald, "Contrastive summarization: an experiment with consumer reviews," in Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, 2009, no. June, pp. 113–116.

[24] I. Imam, A. Hamouda, and H. A. A. Khalek, "An Ontology-based Summarization System for Arabic Documents (OSSAD)," Int. J. Comput. Appl., vol. 74, no. 17, pp. 38–43, 2013.

[25] H. Xu, F. Zhang, and W. Wang, "Implicit feature identification in Chinese reviews using explicit topic mining model," Knowledge-Based Syst., vol. 76, 2015.

[26] W. Wang, T. Zhao, G. Xin, and Y. Xu, "Recognizing Comparative Sentences from Chinese Review Texts," Int. J. Database Theory Appl., vol. 7, no. 5, pp. 29–38, 2014.

[27] I. Gabsi, H. Kammoun, and I. Amous, "ScienceDirect ScienceDirect MeSH-based disambiguation method using an intrinsic information content measure of semantic similarity," Procedia Comput. Sci., vol. 112, pp. 564–573, 2017.

[28] and D. Palmer, M., Fellbaum, C., Cotton, S., Delfs, L. and H. T., "English tasks: All-words and verb lexical sample.," Proc. Senseval-2 2nd Int. Work. Eval. Word Sense Disambiguation Syst., pp. 21–24, 2001.

[29] H. T. Zhong, Z. and Ng, "It makes sense: A widecoverage word sense disambiguation system for free text.," ACL 2010, pp. 78–83, 2010.

[30] W. Senses, "Computing with Word Senses: Chapter 17," in Speech and Language Processing., 2016.

[31] and D. Palmer, M., Ng, H. T., "Evaluation of wsd systems," Agirre, E. Edmonds, P. (Eds.), Word Sense Disambiguation Algorithms Appl. Kluwer., 2006.

[32] S. S. K. P, P. C. R. Raj, and V. Jayan, "Unsupervised Approach to Word Sense Disambiguation in Malayalam," Procedia Technol. ELSEVIER, vol. 24, pp. 1507–1513, 2016.

[33] S. Banerjee and T. Pedersen, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet," A. Gelbukh CICLing 2002, LNCS 2276, pp. 136–137, 2002.

[34] B. T. Mcinnes and M. Stevenson, "Determining the difficulty of Word Sense Disambiguation," J. Biomed. Inform., vol. 47, pp. 83–90, 2014.

[35] G. K. Savova et al., "Word sense disambiguation across two domains : Biomedical literature and clinical notes," J. Biomed. Inform., vol. 41, no. 6, pp. 1088–1100, 2008.

[36] A. J. Yepes, "Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation," J. Biomed. Inform., vol. 73, pp. 137–147, 2017.

[37] S. M. Saqib, F. M. Kundi, and S. Ahmad, "Unsupervised Learning Method for Sorting Positive and Negative Reviews Using LSI (Latent Semantic Indexing) with Automatic Generated Queries," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 18, no. 1, pp. 56–62, 2018.

[38] A. O. Shakeel Ahmad, Sheikh Muhammad Saqib, "LSI Based Search Technique: Using Extracted Keywords and Key Sentences," VAWKUM Trans. Comput. Sci., vol. 14, no. 2, pp. 1–8, 2017.

[39] K. H. Ug, R. Rq, X. Dqg, U. Ri, Q. Dqg, and H. Lq, "$ QDO \] LQJ WKH LPSDFW RI 80 / 6 UHODWLRQV RQ ZRUG VHQVH GLVDPELJXDWLRQ DFFXUDF \," vol. 21, pp. 295–301, 2013.

[40] S. M. Saqib, K. Mahmood, and T. Naeem, "Comparison of LSI algorithms without and with pre-processing : using text document based search," Accent. Trans. Inf. Secur., vol. 1, no. 4, pp. 44–51, 2016.

[41] S. Singh and T. J. Siddiqui, "Role of Semantic Relations in Hindi Word Sense Disambiguation," Procedia - Procedia Comput. Sci., vol. 46, no. Icict 2014, pp. 240–248, 2015.

[42] Y. Hacohen-kerner and H. Badash, "Positive and Negative Sentiment Words in a Blog Corpus Written in Hebrew," Procedia - Procedia Comput. Sci., vol. 96, pp. 733–743, 2016.

[43] O. Nevzorova, A. Galieva, and V. Nevzorov, "Sentence context and resolving lexical ambiguity for special groups of words on the base of corpus data," vol. 198, no. Cilc, pp. 359–366, 2015.

[44] K. Barnard and M. Johnson, "Word sense disambiguation with pictures," Artif. Intell., vol. 167, no. 1–2, pp. 13–30, 2005.

[45] S. Muhammad and F. Masud, "MMO: Multiply-Minus-One Rule for Detecting & Ranking Positive and Negative Opinion," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 5, pp. 122–127, 2016.

[46] G. a. Miller, "WordNet: a lexical database for English," Commun. ACM, vol. 38, no. 11, pp. 39–41, 1995.

[47] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," Proc. 7th Conf. Lang. Resour. Eval. Lr., vol. 0, pp. 2200–2204, 2008.

[48] T. Naeem, F. M. Kundi, and S. M. Saqib, "Extracting Words' Polarity With Definition and Examples," Int. J. Comput. Sci. Inf. Secur., vol. 15, no. 2, pp. 180–190, 2017.

Evaluation of Gated Recurrent Unit in Arabic Diacritization

Rajae Moumen[1], Raddouane Chiheb[2], Rdouan Faizi[3], Abdellatif El Afia[4]
ENSIAS
Mohammed V University
Rabat, Morocco

***Abstract*—Recurrent neural networks are powerful tools giving excellent results in various tasks, including Natural Language Processing tasks. In this paper, we use Gated Recurrent Unit, a recurrent neural network implementing a simple gating mechanism in order to improve the diacritization process of Arabic. Evaluation of Gated Recurrent Unit for diacritization is performed in comparison with the state-of-the art results obtained with Long-Short term memory a powerful RNN architecture giving the best-known results in diacritization. Evaluation covers two performance aspects, Error rate and training runtime.**

Keywords—Gated recurrent unit; long-short term memory; arabic diacritization

I. Introduction

Natural languages require different processing steps in order to perform Natural Language Processing (NLP) tasks, such as Text-to-speech synthesis (TTS), speech recognition, sentiment analysis, information retrieval, etc. In the case of Arabic, an additional preprocessing step is mandatory: Diacritization, or diacritic restoration. Diacritics are signs placed below or above a letter indicating a different phonetic value.

Arabic is a semitic language with two varieties: Classical and Modern. Classical Arabic is the pure language spoken by Arabs; Modern Standard Arabic (MSA) is an evolving variety with constant new terms to meet the modern innovations and changes. Generally, Arabic (Classical or MSA) is transcripted without diacritics, leading to different ambiguities at various linguistic levels as explained in [1].

According to [2], in over 77% of cases, a non-vocalized word can have several possible diacritizations and consequently different possible meanings.

TABLE I. Possible diacritizations for the string “صدق”

Word	Diacritized word	Meaning
صَدَقَ	Sadaka	He is right
صَدَّقَ	Sad~aka	He believes
صدِّق	Sud~ika	We believe him
صِدْقٌ	Sidku	Truth
صُدُقٌ	Sudku	Dowries

Table 1 gives an example of this aspect and lists some of the possible diacritization forms of the string “صدق” and the inferred meaning.

Arabic diacritization received a lot of interest and went through different models: Rule based models, statistical models and hybrid models.

Rule based models rely on existing linguistic rules formulated, in most cases by human experts. They have proven an acceptable efficiency in diacritization, given the lack of linguistic resources. The major drawback of rule-based models is the laborious, costly and time-consuming task to formulate and maintain rules that covers all rich linguistic aspects of Arabic. Moreover, Rule based models require strong linguistic knowledge.

Statistical models attempt to learn a diacritization model from diacritized texts; by predicting the probability of distribution of a sequence of words or characters. Authors in [3] present a review of these methods, using Hidden Markov chains, n-gram or finite state transducers.

The weakness of statistical models is their reliability on large corpus of fully diacritized text. Their strength is that no linguistic knowledge or tools like Pos Taggers or morphological analyzers are needed.

Hybrid methods come from the statement that the strength of linguistic knowledge combined to statistic methods would yield to better results.

Recurrent neural networks (RNN) language models have been used to solve diacritization problem as a statistical or hybrid method. The results are impressive and the error is proven to be asymptotic with a DER of 5.08% over all characters using long short-term memory (LSTM) a powerful RNN architecture proving its performance in various tasks. Moreover, RNN have been used successfully without relying on any linguistic tools, and solely on diacritized corpus.

However, the superior performance of RNN comes at the cost of expensive model training, reaching days or weeks in some experimental settings requiring a significant computation capability.

As a solution to these issues, authors introduced new RNN architectures with simple internal architectures. GRU, introduced in [4] is one of these RNN; in some tasks, it seems to perform better on the training runtime, and maintain a comparable accuracy to LSTM. To our knowledge, diacritization has never been addressed by GRU.

In this paper, we present the results of our evaluation of GRU to enhance the diacritization results regarding its performance in training runtime and error scoring. We use the performances scored by LSTM in diacritization as a baseline. We show that we achieve to maintain the state-of-the art results with better scoring on the training and runtime.

II. Related Work

A. RNN Performance

The first motivation of this study is to enhance the performances of Arabic diacritization. The evaluation of GRU on Arabic diacritization is conducted in comparison with LSTM. In literature, many studies have made this evaluation in various tasks. For example, in [5], authors used a probabilistic approach to determine which RNN architecture is optimal. They evaluated thousand different RNN architectures and identified some that outperform LSTM and GRU in some tasks but not all. In [6], authors compared LSTM and variants over large-scale tasks such as speech recognition, handwriting recognition, etc. Authors conclude that variants of LSTM do not improve significantly the performance.

In [7], an empirical comparison between LSTM and GRU is performed for music and speech modeling. The study did not conclude the superiority of one on the other and then considers GRU to be a better choice since it uses less parameters.

We find the studies in literature inconclusive and insufficient to generalize over all tasks, taking into consideration the considerable differences that might remain in experimental settings and characteristics of the addressed problem.

B. Arabic Diacritization

The Arabic diacritization problem is mainly addressed as a classification problem over seven classes corresponding to the possible diacritics.

Diacritization is divided into two sub-types: The morphological diacritization giving satisfying results reaching an error of 3% to 4%, while syntactic diacritization is still to be improved with a rate of 9.9%.

Earlier approaches in diacritization are rule-based models, like in [8], where morphological analyzer is used for semi-automatic diacritization. Work in [9] presents “Alserag”, another rule based system working through three modules: Morphological analysis, syntactic analysis and morph-phonological module. The system scores 8.68% as diacritization error rate (DER) and 18.63% as word error rate (WER). The main drawback of rule-based models is their high development cost; and the fact that creating linguistic resources such as corpuses are laborious task that need to be reproduced over the studied language.

More recent studies benefit from the evolution of machine learning to learn a diacritization model from vocalized text, the study in [3] presents an overview of these models: Using Maximum Entropy, Hidden Markov Models (HMM) and weighted finite state machine.

In [10], a maximum entropy model is trained for sequence classification to restore the diacritic of each character. They used The Arabic treebank corpus, containing 600 documents form newspapers with over 340k words. The system achieves a DER of 5.5% and a WER of 18%.

More recently, Machine-learning models tend to rely on less and less external resources such as in the study in [11] where the model relies solely on diacritized text. To our knowledge, among systems depending on no external resources, this study gives the best results scored to now with a DER of 8.14% for the end-case character, and 5.08% for all characters.

III. Recurrent Neural Networks

RNN are computational approach based on large connected neural unit forming a directed graph.

In basic RNN we assume that $w = (w_1, \dots, w_T)$ is the input sequence, $h = (h_1, \dots, h_T)$ the hidden sequence and $y = (y_1, \dots, y_T)$ the output sequence, the basic RNN computes h and y by executing the equations (1) and (2) from t=1 to T iterations:

$$h_t = f(w_{hh}h_{t-1} + w_{ih}x_t + b_h) \quad (1)$$

$$y_t = w_{h0}h_t + b_0 \quad (2)$$

f is the activation function for hidden layers, w the matrix weights and b the bias vector, h the hidden bias vector.

RNN are suitable for capturing dependencies among sequential data types. The problem of RNN is that they remain weak on long-term dependencies as studied in [12] where authors proved the difficulty to capture long-term dependencies because of the “vanishing” or “exploding” of stochastic gradients.

Gated recurrent neural networks (GNN) have been proposed to resolve this problem; LSTM and GRU belong to the category of GNN.

A. Long Short-Term Memory (LSTM)

Introduced in [13], LSTM are special case of RNN capable to resolve long-term dependencies issue encountered in standard RNN by using a gating mechanism.

LSTM has the property to remember patterns selectively. Making them suitable for a number of sequence learning problems such as language modelling, translation, speech recognition, and Arabic diacritization.

Study in [11] uses LSTM to build a language-independent diacritizer trained solely from vocalized text without referring to any external tools. Authors run several RNN architectures and achieve with a 3-layer-Bidirectional LSTM to reach a DER of 4.85%.

In the work in [14], the problem is approached with bidirectional LSTM considering diacritization as a sequence transcription problem. The system does not require any previous treatment (lexical, morphological or syntactic) The WER scored in this study is 5.82%.

B. Gated Recurrent Unit (GRU)

First proposed in [4], GRU is generally incorrectly considered as a special-case of LSTM, because of the fact that global architecture is quite similar to LSTM. In fact, GRU is quite different form LSTM .It defines two gating signals instead of three in LSTM, an update signal and a reset gate.

GRU has no cell state. Unlike LSTM, it exposes the memory content at each time step. The transition between the previous memory content and the new memory contents is made using leaky integration controlled by the update gate.

GRU has shown its efficiency in many studies like [15] and [16], it achieves promising results in classification tasks and reduces the training runtime since few iterations are needed to update the hidden states and the internal structure of a cell is simplified.

However, GRU still comes second to LSTM in terms of performance. Therefore, GRU is mainly used in situations where fast training is needed with limited computation capability.

IV. Approach

A. Experimental Setup

Our goal is to set up an environment based on GRU to maintain state-of-the art results and enhance the computation efficiency. For this purpose, we use as a comparison pattern the approach used in [11], giving the best results to our knowledge in same conditions, relying solely on diacritized text.

We compare the error rate and runtime of both GRU and LSTM over same datasets and experimental setup. For this purpose, we use the setup described in Fig. 1. The GNN in the figure stands for the network used, namely GRU and LSTM.

We use single recurrent layer for our networks for limited computation capacity in one hand, and in the other hand to omit potential issues related to multilayer deep learning architectures.

The network needs exclusively two inputs for training, represented in two separate documents, the first one contains the diacritized text, and the second contains the equivalent undiacritized text.

In the whole process, we use the Buckwalter Arabic Transileration, an ASCII scheme, representing Arabic orthography strictly one-to-one.

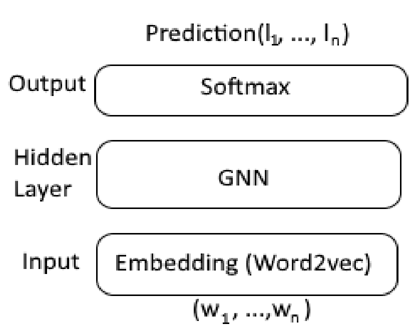


Fig. 1. Architecture of the GNN for Arabic Diacritization.

We tried different environments for experiments, and the results might be quite different to an environment to another; in this paper, we present the results of experiments with Python, Numpy and Theano [17]. We use Theano in order to parallelize computation of GPU and giving the best possible results with the limited computation capacity.

We first go through Word embedding, i.e. mapping the characters sequence into letter vectors. We use for this purpose word2vec [18] implemented with Theano and Lasagne Framework.

We use GRU as it has been introduced in [4] and we define the states as following in the equations (3), (4), (5) and (6):

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot s_t \quad (3)$$

$$s_t = g(W_h x_t + U_h(r_t * h_{t-1}) + b_h) \quad (4)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (5)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (6)$$

$\odot$ stands for element-wise multiplication.

x_t being the input vector

h_t being the output vector

z_t being the update gate vector

r_t being the reset gate vector

W, U and b the parameter matrices and vector

g being the chosen function, here we use *Tanh* as it has been proved to converge faster in practice.

We implement LSTM as following in (7), (8), (9), (10), (11) and (12):

$$f_t = \sigma\left(W_f.[h_{t-1}, x_t] + b_f\right) \quad (7)$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (8)$$

$$C'_t = tanh(W_c.[h_{t-1}, x_t] + b_c) \quad (9)$$

$$C_t = (f_t * C_{t-1} + i_t * C'_{t-1}) \quad (10)$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \quad (11)$$

$$h_t = o_t * \tanh(C_t) \quad (12)$$

x_t being input vector

h_{t-1}being the previous cell output

C_{t-1} being the previous cell memory

H_t being the current cell output

C_t being the current cell Memory

W, U = Weight vectors for gates (forget: f; candidate: c; i/p gate: i; o/p gate: o)

For training, test and dev, we use the data described in the section IV-*B.*

We use a softmax output layer because it ensures that the sum of possible output is 1, moreover, it defines a distribution overall output target classes, suitable to the addressed problem.

B. Dataset

The Dataset used in this work is a part of the corpus "Tashkeela" introduced in [19]. The corpus contains 75 millions of fully vocalized words extracted from 97 books mainly religious and media online from different sources in classical and modern Arabic language. We use the approach described in [10] by splitting the part of the corpus into three parts, Train/Dev/Test as described in Table 2.

TABLE II. DATA STATISTICS

	Training	Dev	Test
Words	560k	90k	80k

V. RESULTS AND DISCUSSION

A. Evaluation Metrics

To evaluate the accuracy of GRU we adopt the same metrics used in most of the diacritization studies, for instance [14], [11], [3]. The metrics are the Diacritics Error rate (DER) that compares the diacritization of the predicted word with the input word at a character level. In addition to the word error rate (WER), that compares the predicted word to the original at a word level, in other terms, if an error is detected on a character, the whole word is considered incorrectly diacritized.

To evaluate the performance of GRU, we consider the average epoch runtime.

B. Diacritic Error

Table 3 lists the results of the networks error rates, measured with diacritic error rate (DER) over all diacritics DER (ALL) and DER over the last character only DER (Last) on the dev dataset.

TABLE III. DER FOR GRU AND LSTM OVER DEV DATASET

	DER (ALL)	DER(Last)
LSTM	7.15	10.50
GRU	7.31	10.79

Experiments show that LSTM achieves better results than GRU in both the DER of all characters and the DER of the last character; however, the results obtained by GRU are still good.

a) Qualitative Analysis

To identify the errors we get from the system implemented with GRU, we proceeded in a direct way by classifying a sample set of the words incorrectly diacritized.

For classification, we use the diacritization scheme error proposed in [20].

Table 4 and Table 5 show the distribution of errors among the sub-categories. We notice that both networks behave in the same way, and distribution over categories is quite the same; we notice that most of the errors are made in Form/Spelling, this is to be expected as it has been reported in works like [21] that the diacritization tools are less performant in case-ending diacritics.

TABLE IV. GRU ANNOTATED ERROR CATEGORIES

Error Category	Sub-category	Nbr of Errors
Form/Spelling	Shadda	8
	Tanween	11
Grammar	Active-Passive Voice	4
Diacritization	Missing short Vowel	5
	Confused short Vowel	2
Morphology	Partial inflection	1
	Full-inflection	0
Overall		31

TABLE V. LSTM ANNOTATED ERROR CATEGORIES

Error Category	Sub-category	Nbr of Errors
Form/Spelling	Shadda	6
	Tanween	10
Grammar	Active-Passive Voice	3
Diacritization	Missing short Vowel	5
	Confused short Vowel	1
Morphology	Partial inflection	1
	Full-inflection	0
Overall		26

C. Training Performances

Fig. 2 compares average Epoch Runtime of GRU and LSTM.

The results do not include the embedding process training.

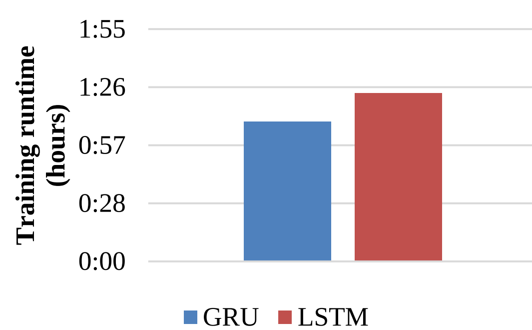


Fig. 2. Average Epoch Runtime of GRU & LSTM.

D. Discussion

We notice that GRU scores better results and outperforms LSTM in training. The Training time is reduced by 18.82%.

The evaluation showed that GRU gives comparable results to LSTM in diacritization accuracy and improves the training process. Consequently, we assume that GRU gives satisfactory results in Arabic diacritization. However, we consider our study to be completed by running different experiment settings over different datasets.

VI. Conclusion

In this paper, we presented a new approach for Arabic diacritization using gated recurrent unit. We showed that GRU outperforms LSTM in training runtime and gives an error rate comparable to LSTM. In future work, we intend to evaluate our approach over different datasets and integer the diacritization tool into a text-to-speech synthesizer for Arabic.

References

[1] A Azmi and R. S.Al Majed, «A survey of automatic Arabic diacritization techniques» *Natural language Engineering,* vol. 3, n° 121, pp. 477-495, 2013.

[2] M. Boudchiche and A. Mazroui, «Evaluation of the ambiguity caused by the absence of diacritical marks in Arabic texts: statistical study» *5th International Conference on Information & Communication Technology and Accessibility (ICTA),* Marrakech, Morocco, Dec. 2015.

[3] I. Zitouni and R. Sarikaya, «Arabic diacritic restoration approach based on Maximum Entropy Models» *Computer Speech and Language,* vol. 23, n° 13, pp. 257-276, July 2009.

[4] K. Cho, B. v. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, «Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation» *Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, October 2014.

[5] R. Jozefowicz, W. Zaremban and I. Sutskever, «An Empirical exploration of recurrent network architectures» *Proceedings of the 32nd International Conference on Machine Learning*, Lille, 2015.

[6] K. e. A. Greff, «A search space odyssey.» *IEEE Trans. Neural Networks. lean. Syst.,* 2016.

[7] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, «Empirical evaluation of gated reccurent neural networks on sequence modeling» *ArXiv preprint,* 2014.

[8] T. A. El-Sadany and M. A. Hashish, «Semi-Automatic Vowelization of Arabic Verbs» *10th National computer conference, Utilization of computers in development*, Jeddah; Saudi Arabia, 1988.

[9] S. Alansary, «Alserag: An Automatic Diacritization System for Arabic» *Intelligent Natural Language Processing: Trends and Applications,* 2017, pp. 523-543.

[10] I. Zitouni, J. S. Sorensen and R. Sarikaya, «Maximum entropy based restoration of Arabic diacritics» *21st International Conference on Computational Linguistics*, Sydney, Australia, July 2006.

[11] Y. Belinkov and J. Glass, «Arabic diacritization with Reccurent Neural Networks» *Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, September 2015.

[12] Y. Bengio, P. Simard and P. Frasconi, «Learning long-term dependencies with gradient descent is difficult» *IEEE Transactions on Neural Networks,* vol. 5, n° 12, pp. 157 - 166, Mar 1994.

[13] S. H. a. J. Schmidhuber, «Long Short-Term Memory» *Neural Comput. 9*, 1997, pp. 735-1780.

[14] G. A. Abandah, A. Graves, B. Al-Shagoor, A. Arabiyat, F. Jamour and M. Al-Taee, «Automatic diacritization of Arabic text using recurrent neural networks» *International Journal on Document Analysis and Recognition*, vol. Volume 18, pp. 183-197, June 2015.

[15] D. Bahdanau, K. Cho and Y. Bengio, «Neural machine translation by jointly learning to align and translate» *International Conference on Learning Representations*, Vancouver, 2015.

[16] A. Graves, «Generating sequences with recurrent neural networks» *Neural and Evolutionary Computing*.

[17] T. D. Team, « Theano: A Python framework for fast computation of mathematical expressions» 2016.

[18] M. Tomas, S. Ilya, C. Kai, C. Greg and D. Jeffrey, «Distributed Representations of Words and Phrases and their Compositionality» *Advances in neural information processing systems*, 2013.

[19] T. Zerrouki and A. Balla, «Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems» *Data in brief,* February 2017.

[20] G. Abuhakema, R. Faraj, A. Feldman and E. Fitzpatrick, «Annotating an arabic learner corpus for error» *Proceedings of the International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[21] N. Habash, R. Gabbard, O. Rambow, S. Kulick and M. Marcus, «Determining Case in Arabic: Learning Complex Linguistic Behavior Requires Complex Linguistic Features.» *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,,* Prague, Czech Republic, 2007.

Smile Detection Tool using OpenCV-Python to Measure Response in Human-Robot Interaction with Animal Robot PARO

Winal Zikril Zulkifli[1], Syamimi Shamsuddin[2], Fairul Azni Jafar[3]
Fakulti Kejuruteraan Pembuatan
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, Melaka, Malaysia

Rabiah Ahmad[4]
Fakulti Teknologi Maklumat dan Komunikasi
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, Melaka, Malaysia

Azizah Abdul Manaf[5], Alaa Abdulsalam Alarood[6]
Department of Computer Science,
Faculty of Computing and Information technology
University of Jeddah, Jeddah

Lim Thiam Hwee[7]
Psychosocial Department
SOCSO Tun Razak Rehabilitation Centre (TRRC)
Hang Tuah Jaya, Melaka, Malaysia

***Abstract*—Human-robot interaction (HRI) is a field of study that defines the relationship between humans and robot. In robot-assisted mental healthcare, there is still a lack in the methodology especially in evaluating the outcome. In this study, PARO; a robot in the shape of a cute baby seal is introduced as an adjunct therapy tool for six rehabilitation patients with post-stroke depression. Currently, the therapy outcome is measured using psychological tools. When a robot is introduced, a new measurement tool is needed to analyse the patients' response. Thus, this study constructs a tool using OpenCV-Python to detect the number of smiles when each patient interact with PARO. Smile is an indicator of positive emotion and that PARO helps to uplift patient's mood. The results were then compared with psychological evaluations. Both tools show congruent results. The number of smiles increased when patients were holding PARO and PARO helped all patients to manage their psychological distress. This indicates that smile detection is an effective supporting tool to indicate respond in human-robot interaction.**

***Keywords*—Human-robot interaction; OpenCV; PARO**

I. Introduction

Application of robots in healthcare setting are on the rise. These include robots used in surgery, to carry medicines, exoskeletons and robots for mental rehabilitation. Human-robot interaction (HRI) is a field that looks into the technicalities and relationship between humans and robots. HRI covers how humans interact with robots, and how the robotic systems are implemented in accomplishing tasks in human environments. HRI aims for effective and safe interaction for both entities. As HRI studies cover humanities, engineering, computer science, medical and social science; this makes HRI inherently a multi-disciplinary field of research.

This project is funded under the Fundamental Research Grants Scheme (FRGS) [FRGS/1/2016/SKK06/FKP-AMC/ F00321].

To date, robots have been used in therapy to treat patients with mental disabilities such as autism, Alzheimer and dementia [1-3]. With robotics involvement, it is expected that the condition of people who suffer from mental disabilities can be improved [3]. Psychotherapy alone may yield positive results, but combination treatments have the best outcome. This is why robotic therapy is suitable as is can be integrated as a tool in the current therapy program

To evaluate the outcome human-robot interaction, a few methods are usually used including surveys, interviews and bio sensing tools. For robotic intervention involving patients with depression, a specific tool is needed to assess the respond through the recorded videos when patients interact with the robot. This is to complement the current psychological tools used by therapists to assess patient conditions. Interaction can be recorded and the videos can be analysed during post-processing stage. This method does not involve asking questions to patients or putting devices on the patient's body.

Fig. 1. PARO the Seal Robot.

Machine learning is a method that allows systems to learn and improve from data and make decisions. In this study, a smile detection tool using OpenCV-Python was developed to detect the number of smiles when each patient interacts with robot. Through machine learning, the number of smiles are detected and serves as an indicator to recognize human's mood while interacting with the robot. This is based on the understanding that the generation of a real smile is produced when a person is feeling positive. Smiles are detected to indicate the mood of a person when they interact with the robot. More smiles means the person's mood is improving.

In this HRI study, an animal-robot PARO (Fig. 1) was used during therapy for patients with post-stroke depression at a multi-disciplinary rehabilitation centre. Most of previous researches have focused the use of PARO robot to treat patients with dementia [4]. There was also a study that investigated which function of PARO that can be associated to reduce psychophysiological stress [5]. PARO was chosen because of its cute and appealing look. It is commercially available, its weight is suitable to patients to hold and hug it and it is the only robot classified as a class 2 medical device by the U.S. Food and Drug Administration (FDA).

The aim of this study is to construct a tool to detect the number of smiles from recorded videos of interaction involving patients with depression and animal-robot PARO. The results of the number of smiles were compared to psychological assessments by a psychologist to evaluate its suitability in the current HRI study.

II. Methodology

The research methodology is shown in Fig. 2. The process starts with developing the research protocol that involves human and robot interaction. The model was based on literature review, subjective evaluation with public respondents in Malaysia and advice from experts in rehabilitation psychology and HRI. Approval from research ethic committee in Universiti Teknikal Malaysia Melaka (UTeM) was filed based on the experimental model. Permission was granted on 19th October 2016. The approval is important because this experiment involved humans as subjects and their rights, as well as the rights of the researchers need to be protected.

The research continued with OpenCV-Python program development for smile detection. This program was developed as a HRI evaluation tool to analyse the outcome of the experiment. Next step was the full-scale HRI study with six patients in the age between 35 to 52 years old. Two were males. All patients were diagnosed with post-stroke depression. The inclusion criteria were: diagnosed with depression after stroke, currently an in-patient at the SOCSO Tun Razak Rehabilitation Centre (TRRC) Melaka, able to converse in Malay language or English, able to stroke an animal and can hear clearly. They were currently undergoing rehabilitation therapy program at TRRC. Patients with depression at centres like TRRC undergo rehabilitation treatment before they are declared stable.

The set-up of the session involving patient, PARO and psychologist is shown in Fig. 3. The experiments were conducted in three separate sessions. First, the therapy room in TRRC were set-up with HD video camera, voice recorder and projector. Next, the patient was invited to enter the room. Then, the patient was given brief introduction about the experiment, his/her consent was taken, and a short video of PARO was shown.

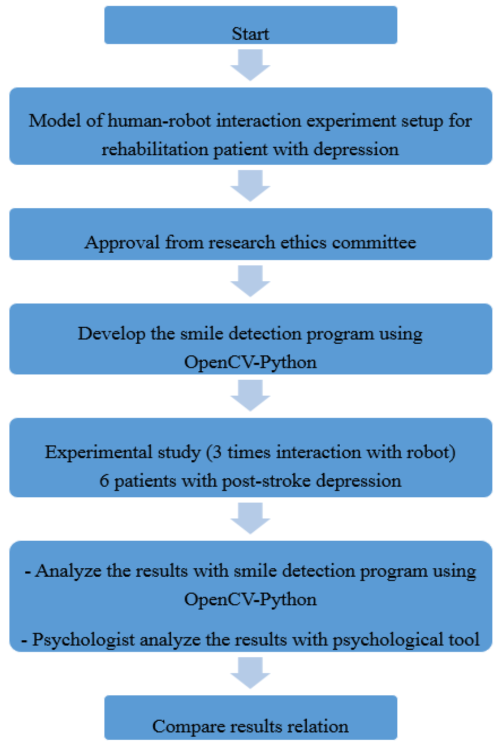


Fig. 2. Flowchart of the research methods involved in this study.

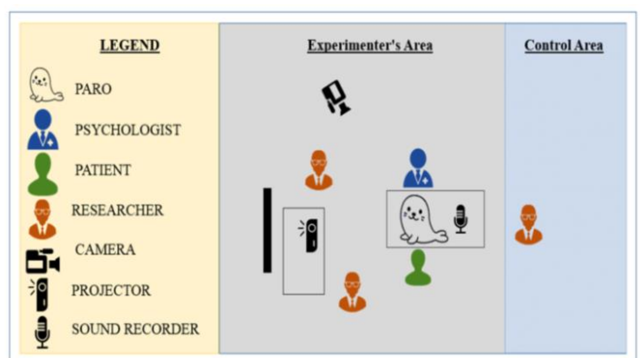


Fig. 3. Experimental layout.

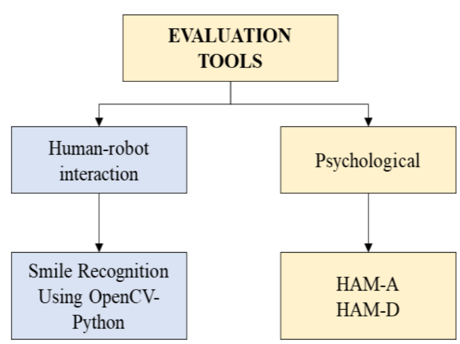


Fig. 4. Evaluation tools used in the study.

The following step was the psychological pre-evaluation on the patients by a rehabilitation psychologist from TRRC using the Hamilton Depression Rating Scale (HAM-D) and Hamilton Anxiety Rating Scale (HAM-A). HAM-D is a validated screening instrument for post-stroke depression [6] to provide indication of depression, guilt, suicide, insomnia (sleep problem) and as guide to evaluate recovery. HAM-A measures the severity of anxiety symptoms and is widely used in both clinical and research settings. The reason is to identify the patients' condition before robotic intervention and identify the changes in patients in respond to the robot.

Then, the 15 minutes session took place to obtain therapy recording without PARO. During this period, the psychologist began to introduce general topics to talk with the patient such as their favourite food, their activities during weekends and their family members. Indirectly, the psychologist was using cognitive behaviour techniques and goal directed behaviour during the therapy. After 15 minutes, interaction with PARO began. PARO was brought into the room by the experimenter and was put in front of the patient. With PARO in the room, the consultation therapy continued with PARO as a companion to the patient. The psychologist continued to have conversations with the patient. The 15 minutes therapy without PARO and 15 minutes with PARO cycle was repeated for all three sessions. At the end of each session, the patient was informed that PARO will be taken away. Total duration of recorded experiment was limited to 30 minutes because post-stroke patients cannot withstand long therapy sessions. They also get tired easily.

The interaction took place at a therapy room in TRRC for 3 sessions during one month. Each session lasted for 30 minutes and it included therapy time with PARO and without PARO. For the second and third session, the psychologist also carried out the HAM-D and HAM-A evaluations. The same protocol was applied for all the six patients recruited for the study.

A. Evaluation Tools: HRI and Psychological

To evaluate the outcome of experiment, the tools are used that are separated into two categories: human-robot interaction (HRI) and psychological evaluation as shown in Fig. 4. HRI evaluation consists of the smile recognition tool developed using OpenCV-Python. A smile is a positive emotion. To identify the changes in depression patients, smile has been chosen as an indication of the patient's positive mood. To identify the smile occurrences, the OpenCV-Python program analysed the recorded video from the patient-robot interaction session.

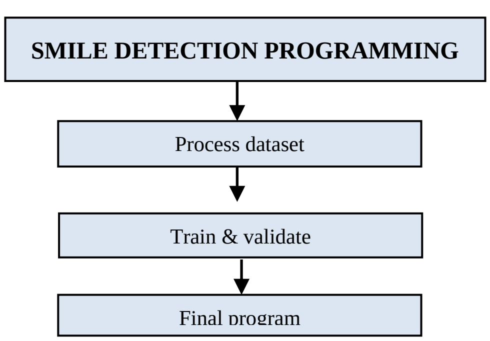


Fig. 5. Smile detection development program flow.

For the psychological tools, HAM-D and HAM-A were chosen because of their reliability and recommendations from the psychologist. HAM-D and HAM-A were used together because the observations on patients with depression always correlate with assessments of depression and anxiety.

B. Smile Detection Programming using OpenCV-Python

The development phase of the OpenCV-Python is shown in Fig. 5. The objective of the program is to identify and count the smile on patients during the experiment. First, a collective data on human smiles must be included for smile identification. This action followed by the second step which is identifying appropriate dataset processing library. OpenCV is an open source computer vision library for commercial and research use. OpenCV is one of the most widely used libraries in image processing. The OpenCV was chosen for its extensive library, it is simple to use and has extensive user network.

Once the dataset library had been identified, the development continued with detection and learning. For smile detection algorithm, Adaptive Boosting (AdaBoost) was used. Using AdaBoost improves the accuracy of the learning algorithm where the output of multiple "weak classifiers" are combined into a weighted sum that represents the final output of the boosted classifier.

Smile detection was targeted at the mouth region. In the smile detection mechanism, the videos were processed frame by frame and converted into a grey scale. Smile detection determines 'where' in a face image the 'smile' is located, and this is done by scanning the different face image scales and extracting the exact patterns to detect a smile. The classifier uses a single feature to define images as smiles or non-smiles which are stored in cascade data of the numbers of 'smiles'. The program is trained using the pre-determined dataset in OpenCV for program environment adaptation. The program analysed the recorded videos from the experiment.

To prepare a dataset, raw images of smiling faces were needed. Previously, Chang et al. used ten individual subjects and captured four images from each of them [7]. This resulted with 40 sample images for the database.

Fig. 6. Sample dataset of 300 smiling faces of Malaysians.

However, in this study, the developed program in OpenCV-Python is able to capture ten images within 100 milliseconds. These images were acquired by running a video through an OpenCV-Python program to capture still images. Using video recordings of thirty different individuals, 10 smiling images were extracted from each individual. This makes a total of 300 smiling faces (Fig.6) for the database. The steps to run this program are divided into three main program sections which are: detector, dataset creator and trainer. For performance validation, the dataset was divided into training and validation. The verification of this data used the 5-fold cross validation [8]. Then, the output images were stored using unique identification codes. The images are labelled with target vector '1' for smiling and '0' for not smiling. Images that are labelled with not smiling are used as negative image dataset.

The HRI smile detection tool was made with OpenCV-Python program to introduce a new dataset with acceptable accuracy for smile identification. With the new dataset, the accuracy of the HRI smile detection program was improved. The new dataset also introduced sample images from Malaysians.

III. Results and Discussion

In this study, the HRI smile detection program was used to analyse the video recordings. The performance of the program was tested before it was applied in the experiment study.

A. *Performance of the Smile Detection Program*

To evaluate the effectiveness of this program, an accuracy test was carried out. The test used three types of dataset for tracking. Each dataset consists of pictures of human faces. 'Smile' pictures were labelled with '1'. The 'not smile' pictures were label with '0'.

To test for accuracy, dataset of images was fed into the smile detection program. The first test used the dataset of 300 images from 10 Malaysian subjects. To test for accuracy, this dataset used 5-fold cross-validation. The technique performs learning and training at the same time.

The second dataset used to test for accuracy was GENKI. The GENKI database, developed by MPLab is an expanding database of image containing faces spanning from a wide range of illumination conditions, geographical locations, personal identity, and ethnicity [9]. The database contains 25,064 images. The third dataset was the Japanese female facial expression (JAFEE) [10]. The database contains 213 images from 60 Japanese subjects. The results on program accuracy are shown in Table 1.

Dataset of the 300 smiling faces of Malaysians scored the highest 90% accuracy. This high score is expected because the same dataset was used to teach/train the program. The results with JAFEE and GENKI dataset are at 76% and 70%. These scores were acceptable to run smile detection on the videos for the main experiment. In image processing, the acceptable image detection accuracy should be more than 70% [11]. Therefore, this smile detection program developed in this study has acceptable accuracy with more than 70% score in the testing result.

TABLE I. Accuracy Results For Dataset Testing

No.	Dataset	Accuracy
1	300 smiling faces of Malaysians (5-fold cross-validation)	90%
2	JAFEE	76%
3	GENKI	70%

B. *Results of Smile Detection for Experiment with PARO*

The smile detection program was used to analyse the recorded videos from three sessions for each of the six patients. Fig. 7 shows the results of the smile detection program. For each session, the time is separated into 15 minutes spent without the robot and 15 minutes with the robot.

Patient 1, Patient 3 and Patient 4 have shown a continuous increase of the number in smiles starting from the first session until the third session with PARO. Their smile increases with each session in sequel. For Patient 2, Patient 5 and Patient 6 there was also an increase in the number of smiles from the first to the second session. However, the number of smile drops after the second session. When comparing the number of smiles during sub-sessions spent with the robot and without the robot, it is seen that all patients, except for Patient 1, show an increase in the second sub-session with PARO.

Looking at the results of smile detection for the sessions without PARO, it can be seen that the number of smiles also increase except for Patient 5 and Patient 6. For the other 4 patients, they were feeling happier because they are responding well to their rehabilitation program at TRRC. Nevertheless, the presence of PARO managed to increase their happiness and positive mood as the session with PARO resulted with more smiles than the session without PARO. This is a positive outcome. The results indicate that the animal robot PARO gave a positive impact on them emotionally.

On the other hand, Patient 1 in the second session recorded a higher number of smiles in the period without PARO than with PARO. This interesting finding indicate that though PARO is cute and likeable, there is also a small number of people that do not like the robot or find it attractive. Specifically, for patients undergoing rehabilitation therapy, introducing PARO as a tool during therapy may also relate to the patient's acceptance of a robot considering that this is the first study of its kind in Malaysia

C. *Results of Psychological Tools*

Evaluation using the HAM-D and HAM-A were carried out by the psychologist and the results are tabulated in Table II. When using the HAM-D instrument, the score of 0 to 7 is within normal limit, score of 8 to 13 indicates 'mild severity' and score between 14 to 17 falls under 'moderate severity'.

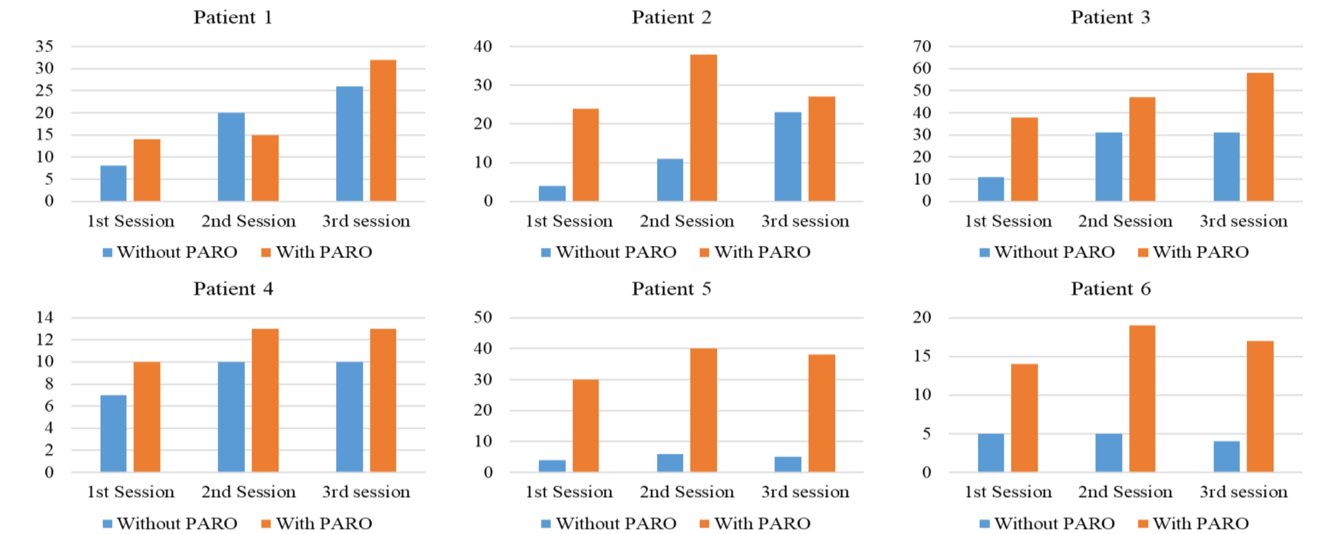


Fig. 7. Results for HRI smile detection program.

Before the experiment, Patient 1 was diagnosed with ‘mild severity’ depression level. After second session, Patient 1’s depression level rise but still in the same level. After the third session, Patient 1 level of depression went down to level ‘within normal limit’. This was better compared to the first evaluation.

Patient 2 and Patient 5 were diagnosed with ‘mild severity’ depression before the experiment started. Their condition showed improvement after second and third session with PARO. Their depression level had improved to ‘within normal limit’.

For Patient 3 and Patient 6, their depression level started with ‘mild severity’ and continued with the same level in the second session. After the third session, Patient 3’s depression level improved to ‘within normal limit’. On the other hand, depression level for Patient 6 also decreased into ‘mild severity’ level. Patient 4 depression level showed no changes, comparing before the experiment started and after the third experiment session, which was in the ‘mild severity’ level. Overall, depression level of all six patients had decreased and their mental health were getting better.

TABLE II. SCREENING RESULTS OF PSYCHOLOGICAL TOOLS

Screening	**HAM-D**			**HAM-A**		
Session	*Pre-experiment*	*2nd*	*3rd*	*Pre-experiment*	*2nd*	*3rd*
Patient 1	9	13	7	11	14	15
Patient 2	8	5	5	20	21	18
Patient 3	12	9	6	23	18	17
Patient 4	13	13	11	15	11	11
Patient 5	8	6	3	10	8.8	6
Patient 6	15	14	10	18	18	15

With HAM-A, the score of 0 to 15 is within normal limit, score 16 to 20 is ‘mild severity’ and the score of 20 to 25 falls under ‘moderate severity’. Before the experiment, all patients was screen with HAM-A for assessment on their anxiety level. Patient 1, Patient 4 and Patient 5 level of anxiety were ‘within normal limit’ since before the experiment and until the third session. Even though the Patient 1 showed an increased score anxiety, but the scores were still within the normal level.

On the other hand, Patient 4 and Patient 5 scores had decreased. For Patient 2, her level of anxiety was screened with ‘moderate severity’ before the experiment and after the second session. In the third session, her anxiety level reduced to ‘mild severity’ level. Patient 3 was also screened with ‘moderate severity’ before the experiment. But the anxiety level dropped to ‘mild severity’ level after second and third session. This is in contrast to Patient 6 who was screened with ‘mild severity’ before the experiment and after the second session. After the third session, the level of anxiety fell to level ‘within normal limit’. Overall, all the patients’ score of anxiety had decreased, except for Patient 1. This result shows that most patients are recovering from their anxiety condition.

D. Relation of HRI Tool with Psychological Tools

To find out whether the results from smile recognition program is congruent with psychological tools assessment, both data were analysed using SPSS software with analysis of variance (ANOVA). The purpose of using ANOVA analysis is to find out the statistical models from the tool’s results and associate their data for estimation through mean square, *F-test* (*F*), *degree of freedom* (*df*) and *significant* (*Sig.*).

Table III shows the results output from ANOVA. Mean square value was used as denominator for F ratio. The F statistic is defined as the ratio between the two independents. In this study, the independent is between HRI smile recognition tools with HAM-D and HAM-A. The test statistic for F value resulted with 9.264, 12.157 and 12.157. Using an α value of 0.05, we have *Sig.* = 0.001, 0.000 and 0.000. The *df* value of 5 and 12 were applied with these levels.

Since the test statistic is much smaller than the critical value, the null hypothesis is accepted (>= 0.5) of equal population means. This concluded that there is a (statistically) significant among the population means. So, this statistic test proves that the results from HRI smile recognition tool were significant at the predefined *df* levels with the HAM-D and HAM-A tools.

All patients who took part in this study were ordinary (typical) individuals, which means that they were not diagnosed with psychiatric illness. They were admitted to the rehabilitation facility due to their post-stroke condition. Therefore, it is relevant to use a smile detection program with the trained database from ordinary people too. The smiles produced by these patients show congruency with the results from psychological assessment tools. Results were different (in terms of scores) between each patient because, as suggested by the rehabilitation psychologist all patients react differently when they interact with animal robot PARO.

Each patient has their traits. However, in the results with six patients, there are significant between depression and smiles. More smiles mean their mood were improving. PARO give the patients more smiles during the interaction. It is through interacting (i.e., touch, hug, stroke, pet and kiss) and communicating with PARO. This behaviour improves the moods of the patients and as a results, they smiled more. When mood improves, depression is reduced.

Anxiety as a mental condition always have comorbidity results with an individual's depression level. Based on the statistical results level of anxiety have relationship with smiles. In this study, the patients' anxiety level reduced when their number of smile increased. With the presence of PARO during the interaction, it is suggested that the patients were recovering and their positive emotion increased. The patients' mood also improves after each interaction.

Overall, results from HRI smile recognition tool was similar with psychological tools results. Thus, smile recognition can be used along with psychological tools to evaluate the interaction of patients with animal-robot PARO.

TABLE III. ANOVA TEST ON HAM-D AND HAM-A

Tools	df	Mean Square	F	Sig.
HRI (Smile Recognition)	5	577.967	9.264	0.001
	12	62.389		
HAM-D	5	59.869	12.157	0.000
	12	4.924		
HAM-A	5	59.869	12.157	0.000
	12	4.924		

IV. CONCLUSION

The study aim was achieved by the development of a smile recognition tool using OpenCV-Python for use in recorded videos of HRI experiment. The tool was created using a new embedded dataset consisting of 300 smiling faces. By using 5-fold cross validation technique to train and validate, the image recognition accuracy obtained was 90%.

Conclusively, the smile recognition tool successfully recognised smile occurrences in the HRI experiment involving patients with post-stroke depression and animal-robot PARO. The results are compared with the psychological tools. Statistically, the smile detection tool has significant results with the psychological tools. This suggests that the results produced by HRI smile detection tools can be used to evaluate response during the human-robot interaction.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Ministry of Higher Education Malaysia, Universiti Teknikal Malaysia Melaka, University of Jeddah and SOCSO Tun Razak Rehabilitation Center (TRRC) Melaka for their support. This project is funded under the Fundamental Research Grants Scheme (FRGS) [FRGS/1/2016/SKK06/FKP-AMC/ F00321].

REFERENCES

[1] C. Breazeal, "Toward sociable robots," in Robotics and Autonomous Systems, 2003, vol. 42, no. 3–4, pp. 167–175.

[2] S. Shamsuddin, N. A. Malik, H. Hashim, H. Yussof, F. A. Hanapiah, and S. Mohamed, "CCIS 376 - Robots as Adjunct Therapy: Reflections and Suggestions in Rehabilitation for People with Cognitive Impairments," Ccis, vol. 376, pp. 390–404, 2013.

[3] C. Bühler, "Robotics for rehabilitation - A European (?) perspective," Robotica, vol. 16, no. 5, pp. 487–490, 1998.

[4] S. Šabanovic, C. C. Bennett, W. L. Chang, and L. Huber, "PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia," in IEEE International Conference on Rehabilitation Robotics, 2013.

[5] R. Aminuddin, A. Sharkey, and L. Levita, "Interaction with the paro robot may reduce psychophysiological stress responses," ACM/IEEE Int. Conf. Human-Robot Interact., vol. 2016–April, pp. 593–594, 2016.

[6] A. F. G. Leentjens, F. R. J. Verhey, R. Lousberg, H. Spitsbergen, and F. W. Wilmink, "The validity of the Hamilton and Montgomery-Åsberg depression rating scales as screening and diagnostic tools for depression in Parkinson's disease," Int. J. Geriatr. Psychiatry, vol. 15, no. 7, pp. 644–649, 2000.

[7] C. Chang et al., "Multi-level Smile Intensity Measuring Based on Mouth-Corner Features for Happiness Detection," pp. 181–184, 2014.

[8] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in Encyclopedia of database systems, Springer, 2009, pp. 532–538.

[9] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward practical smile detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 11, pp. 2106–2111, 2009.

[10] J. Michael, M. K. Lyons, and J. Gyoba, "Japanese female facial expressions (JAFFE)," Database Digit. images, 1997.

[11] J. T. Mundt, N. F. Glenn, K. T. Weber, T. S. Prather, L. W. Lass, and J. Pettingill, "Discrimination of hoary cress and determination of its detection limits via hyperspectral image processing and accuracy assessment techniques," Remote Sens. Environ., 2005.

Functionality Gaps in the Design of Learning Management Systems

Tallat Naz, Momeen Khan
Dept. of Computer Science
IIC University of Technology
Phnom Penh, Cambodia

***Abstract*—This research paper focuses on various gaps associated with the Learning Management System (LMS) and their remedies. LMS is a software application platform upon which multiple tasks related to online tutoring are created. For organizations, it's crucial that the risks associated with any automated process are kept as low as possible. This should also pertain to selecting the LMS platform for educating new professionals to the organization. To execute this, organizations, should carry out due research before incorporating any system as their primary LMS. Even though, they provide a lot of benefits for organizations integrating such platforms. Choosing faulty LMS for training recruits can lead to a variety of issues later on. Thus, it becomes essential to select the best LMS platform available in the market, and the one suits the organization's needs. The work proposed in this paper is listing together a number of problems that exists in any given LMS framework and trying to alter them according to the needs of the organization so that they provide a feasible solution and deliver a better guidance to the recruits.**

***Keywords*—Learning Management System (LMS); shortcomings of LMS; functional gaps in LMS; LMS design issues; remedies for gaps in design of LMS**

I. Introduction

Learning has always been an essential pursuit for humankind since ancient times. With the advent of the internet, the quantity and quality of information available to a person now have increased exponentially. For the traditional learning method, the Internet has allowed to be side-lined gradually and is replaced by electronic learning methods [1]. These all electronic-learning methods are collectively called e-learning. It is the broad term used to define any formal teaching done with the help of electronic devices. Internet and computers form essential components of the e-learning method, but is not limited to these. E-learning includes everything from early implementations such as pre-recorded CD & DVD, to the latest innovation such as live, interactive classes, where the teacher and students can interact even when they are on the other side of the planet. E-learning offers a variety of advantages over the conventional teaching methods such as cost-saving measures and ability to learn around the clock. With the advancement of mobile technology, the new concept of m-learning has also developed, that is e-learning done through mobile devices [2]. It is estimated that the percentage of using e-learning methods is steadily rising in the last few years. According to statistics, nearly 99% of higher education institutions currently adopt and run LMS platforms [3]. According to an Educause Centre for Analysis and Research (ECAR) survey, 86% of teachers use LMS and out of which 56% use it on a daily basis [4]. 83% of students use LMS in at least one course with 56% using it in most or all courses [5]. This demonstrates that teachers and students value the LMS as an enhancement to their educational experience.

In this paper, we focus on the difficulties that are encountered by the organizations that rely on LMS platforms for their day-to-day usage. These organizations can be educational or corporate. A number of strategies are also provided for correcting these difficulties. These are the problems that can have undesired results for the organizations and can often result in the delay of the training or educational process.

This paper helps in contribution towards the ever-present problems that many organizations face, i.e. the training of recruits. For this purpose, implementing an LMS framework is convenient and practical. Thus, it becomes crucial that the LMS selected can fulfill all the necessary requirements that an organization can have and manage them well. Some of the common functionalities of modern LMS are shown in the fig. 1.

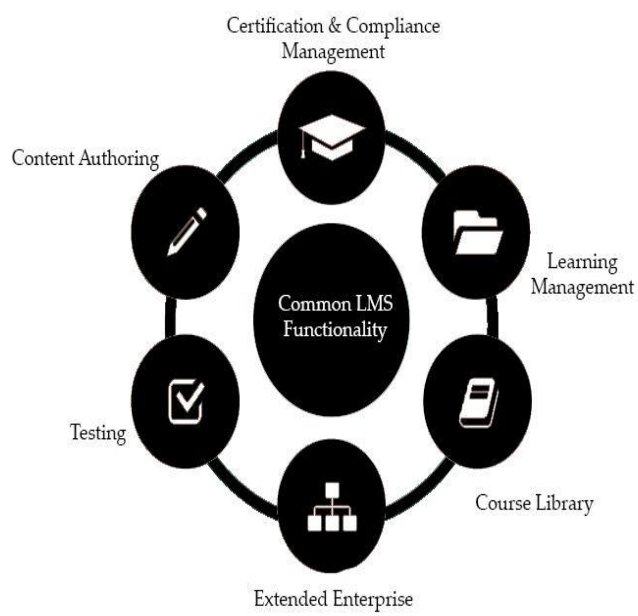


Fig. 1. Functionalities Common to All Modern Lms.

II. LMS AND ITS MODELS

Learning Management System (LMS) is an online platform that allows a user to carry out many tasks related to e-learning. LMS can be understood as the mechanism that powers e-learning. The functions of an LMS can range from essential services such as providing educational content to tracking a user's progress throughout the course's timeframe and conducting a regular assessment. LMS can also perform other processes such as chat boxes for users, periodic tests and other necessary tasks related to teaching. All organizations that have to educate recruits, either students or professional, can benefit from LMS. But deciding the correct LMS to incorporate is the major hurdle organizations have to go through. A variety of options are available that differ from each other in functionality and expenses incurred[6]. Two different models of LMS are explained below.

A. Massive Open Online Courses (MOOC)

MOOCs are sites that furnish an online course with a many choices. These choices are free and open, provide generally shared educational modules, and convey open-ended results. MOOCs coordinate social networking, data that are available online open on the web, and are advanced by top professionals in the field of study[7]. Most strikingly, MOOCs expand on the engagement of researchers who self-compose their support as per learning objectives, earlier information and abilities, and comparative interests.

B. Software based LMS

The same concept of MOOC when performed through the use of software is said LMS software. There are two types of LMS software that is widely available, each with its advantages and disadvantages.

1) Open Source LMS: These are software that is free-to-use for everyone and organizations can modify upon the existing platform provided by the developer community. The most popular example of this category is Moodle [8]. They offer free of cost LMS solution but come with many hidden costs such as hosting fees, extra storage space, and more tech support [9].

2) Proprietary LMS: Proprietary LMS platform is those that are developed by for-profit organizations. Blackboard is an example of the patented LMS with the maximum market share [10]. These are comparatively more secure and reliable than open-source LMS. They also enable the creation of training courses. But the higher cost associated with might keep smaller companies at bay.

III. ANALYSIS OF VARIOUS COMPETITORS OF LMS PLATFORM

The LMS platform market has many competitors that have cemented their place amongst the consumers. Some of the top systems consist of both open-source and proprietary platforms. A quick analysis of the market share provides us with some important information.

A. Market Share in American Higher Education Institutions

The market share of all the top LMS competitors which are utilized in institutes with more than 2000 students is analyzed below. The two systems that are at the top, that is, Blackboard & Canvas are both proprietary LMS. The market share of Moodle has remained stable over the last few years but has fallen in terms of total usage. When comparing with institutes with less than 2000 students, the trends remain the same and Blackboard is again the majority holder, as shown in the fig. 2.

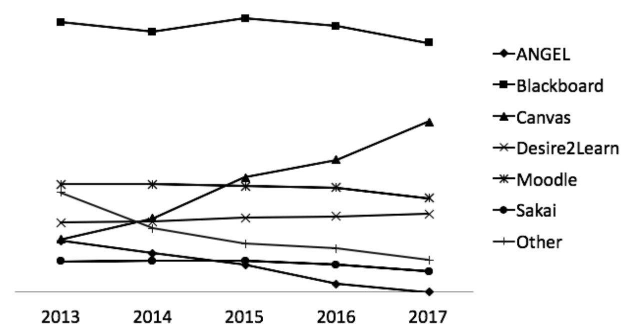


Fig. 2. Market Share in US [11].

B. Global Scenario of LMS

When observed in the global standings, we get approximately the results with Moodle jumping to second position. Canvas is nearly non-existent in other top markets for LMS, such as UK & Canada as shown in fig. 3.

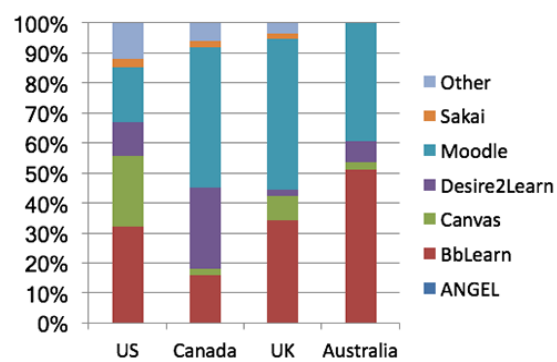


Fig. 3. Holdings in 4 Biggest Markets of LMS.

IV. SHORTCOMINGS OF LMS

Most organizations have to adhere to the standards approved by the regulatory bodies of EU and USA. As LMS is the integral part of the development and learning process, they have to keep the vigilance on the LMS they operate on. Due to the several risks that are associated with using an LMS with any flaw, it is imperative for the organizations to select the best product that fulfils all their requirements.

The various potential gaps of the Learning Management System are addressed below:

A. Lack of Version Control

The first gap of LMS is the lack of version control for the training content, which can result in two major issues. The first and foremost is the increased rate of human error. Because of the higher human error rate, the collected content can be deemed unreliable. This can increase the risk of wrong information being transverse. The other one is an absence of automation for version training and reviewing. This can lead to considerable in-efficiencies for the administrative department. There are many Corrective Action/Preventive Action (CAPA) directives which will be breached due to this gap.

B. The Absence of Computerized Electronic Markings or e-Signatures

The absence of computerized electronic markings or e-signatures builds a level of deficient in preparing results and record maintenance, which leads to inconsistencies in the documentation process. E-signatures require extra managerial action as they don't automatically add to every one of the courses associated with the training program and as a result, can confuse the students.

C. Deficiency of Audit Trail

The third shortcoming is the deficiency of audit trail data on curriculum profiles and learner's training items. Finite data on what was changed and who rolled out the improvement, failure to hold the information for the required time and day, and the absence of date and time-stamp capacity. This hinders the preparation program control and observing, and lessens the impact of the ability to build up an extensive, end-to-end errand review trail. The downside of an LMS to record both old and new execution of benchmarks makes the framework inconsistent for Code of Federal Regulations (CFR). The need to physically screen which is contributing the framework changes, the date and time the progressions were made, and who rolled out the improvements, can likewise cause the LMS not agreeable to CFR mandates.

D. Managerial and Administration Consent

Another gap of LMS is that all managerial and administration consent should be performed physically, outside the LMS if the preparation design endorsement processibility is inadequate. Administration unequipped for surveying singular preparing plans against worker sets of expectations with a specific end goal to guarantee precision when enlisted and refreshed when there is an adjustment in part. With expanding occurrence of difficult work fundamental to make up for the absence of electronic forms, the margin for the human mistake would likewise increment. Therefore, extra work-force must be sent to carry out a manual audit of the preparation program.

E. Insufficient "Audit-Ready" Reports

Insufficient "audit-ready" reports for managers and administrators indicate the real-time qualification's status of learners. Administrators and managers must get immediate notification and access to the qualification and the compliance level of their employees. These records also need to be presentable during an audit. They must have all necessary data so that it seems it is represented from a trustworthy and accurate source.

F. No Automated Role-Based Training Management Features

As the highlights of LMS and coordinating usefulness that represents the part based preparing process are inadequate, the staff engaged with the quality framework can't be assembled into automated parts relying on their activity capacities. Development, evaluation, approval, and assignment of individual training plans have to be done manually when there are no automated role-based training management features. This raises the issue of human mistakes and can give the supervisors the impression that the education plans are assigned inconsistently. It can indicate that the employees are not adequately trained before performing the activity.

G. Incapable of Tracking the Training Equivalencies

Administrators and managers can't unmistakably assess, set down, allot and report comparable preparing performed in a few preparing modalities/strategies.

The current LMS's are unequipped for following the preparation equivalencies between different learning exercises. The functionality accessible to address those equivalencies is constrained, and just a similar preparing technique/sort can be made identical. For example, English and German adaptations of related material can be made similar, yet slightly different types of learning sources are past this parameter. Along these lines, a read and sign preparing on a SOP or an educator drove preparing on a SOP isn't created as, or seen as a comparable in the LMS.

V. Addressing the Gaps

The gaps mentioned above of the LMS platform can be corrected by carrying out the following processes.

A. Acquiring Adaptation Control Capacities

Acquiring adaptation control capacities can dispose of manual compromises for a developing volume of different material forms. It ensures the organization against the issuance of warning letters, and other review remediation orders. It improves the working efficiency of both the system and quality organization.

B. Electronic Marking and Records Consistent Or E-Signature Production

Electronic marking and records consistent with stringent necessities of FDA, CFR and EU Annex will permit that the e-signature must be produced toward the fruition of preparation. It must keep up security and review trails to guarantee it is an immaculate record. It must produce the approval of frameworks to guarantee unwavering quality, exactness, the capacity to observe adjusted or invalid records, and predictable, planned execution. It must generate the time of completion and exact duplicates of documents in both electronic and comprehensible frame appropriate for assessment, examination, and replicating by the administrative office.

C. Date and Time-Stamped Review Trails

The time and date of operator activities and operations that create, change, or delete electronic records can be freely recorded by secure, computer produced, date and time-stamped review trails..

D. Intermittent, Computerized Surveys of Preparing Plan Usefulness

Limit and minimize the reliance on manual and administrative actions. It can be adjusted to coordinate the interior preparing record maintenance disciplines that will boost responsibility, oversight, and control of preparing plan surveys.

E. Automated Reports Preparations

Automated reports preparations and notices progressively that can detail the capability and consistency status of students. It informs both framework heads and directors using email when a worker's task is past due. It offers current access to preparing records continuously. It produces "review prepared" reports in pdf format and arranges them according to date/time stamps, report criteria, page numbers, and so forth.

F. Full Automation of the Critical Preparing Plans

Full automation of the critical preparing plans from representative on-boarding to progressing capability for those staff engaged with GxP activities. It guarantees making of steady preparing records for similar courses, and reliable reporting and tracking. It empowers the staff to satisfy the preparation objective necessary to get to access system or conduct operations and guarantees the opportune individuals are prepared at the ideal time on the correct material.

G. Following of Training Equivalency between Different Learning Sources

Limits the human mistake caused by physically taking care of representative preparing fruitions and oversight of records. A top-tier report for each kind of preparing can be united by system managers, making it less demanding to run reports and react to internal info demands.

VI. Discussion

Observing the above-mentioned risks, it is imperative that the LMS selected for the usage in organizations are the best product possible and a product that meets all the requirements. Many of the top products available in the market have gained market penetration which hinders the decision making of the organizations. To avoid that, organizations should survey all the available options. Many new products are streamlined for particular tasks which are better suited for the organization. Any error on the part of LMS can have significant effects on the working of these organizations. To avoid these mistakes, it is essential that the organizations take in mind all the different options available for the LMS selection and choose the best option possible.

VII. Conclusion

Learning management system is playing a vital role in the learning and development process of any organization. If applied accurately, the rectifications offered by this research paper will help the organizations to respond to evolving requirements that comes with training new professionals. It is of utmost importance that the recruits are trained as best as possible. The LMS platform created after correcting all the variety of issues mentioned above can help an organization to provide a well-organized experience.

The future application of the proposed drawbacks and their improvement can help the organizations to make an informed decision about the various drawbacks present in LMS platforms and have to be avoided. There are a number of opportunities available for future development that can be pursued by making this particular study as the basis for advance research.

Acknowledgment

We would like to express our gratitude to Prof. Dr. Chhuon Chanthan and IIC University of Technology, Phnom Penh, Cambodia for providing us good support to conduct this work.

References

[1] Andersson, A., (2008). Seven Major Challenges for e-learning in Developing Countries:Case Study eBIT, Sri Lanka, International Journal of Education and Development using ICT, Vol 4, Issue 3

[2] Hunsinger, J. (2005). How to determine your readiness for mobile e-learning.Information policy. http://ipolicy.typepad.com/informationpolicy/2005/04/how_to_determin.html

[3] Dahlstrom, E., Brooks, D. C., & Bichsel, J. (2014). The current ecosystem of learning management systems in higher education: Student, faculty, and IT perspectives. *Research report. Louisville, CO: ECAR.* https://net.educause.edu/ir/library/pdf/ers1414.pdf

[4] Berking, P., & Gallagher, S., (2016). Choosing a Learning Management System (Ver. 7.0). Advanced Distributed Learning (ADL) Co-Laboratories, 14, 40-62.

[5] Brown, M., Dehoney, J., & Millichap, N. (2015). The Next Generation Digital Learning Environment. *A Report on Research. Educause Learning Initiative paper*. Retrieved from https://net.educause.edu/ir/library/pdf/eli3035.pdf

[6] Yuan, L. & Powell, S. (2013). MOOC and Open Education: Implications for Higher Education. CETIS: Centre for Educational Technology, Interoperability and Standards, White Paper, WP03, http://publications.cetis.org.uk/2013/667

[7] Diver, P., & Martinez, I. (2015). MOOCs as a massive research laboratory: Opportunities and challenges. *Distance Education*, 36(1), 5–25. http://dx.doi.org/10.1080/01587919.2015.1019968

[8] Feller, J., & Fitzgerald, B. (2000). A framework analysis of the open source software development paradigm. *In Proceedings of 21st International Conference on Information Systems (pp. 58-69). Australia: Brisbane.*

[9] Aydin, C. C., & Tirkes, G., 2010: Open source learning management systems in e-learning and Moodle. *InEducation Engineering (EDUCON)*, 2010 IEEE (pp. 593-600).

[10] Pankaja, N., & Mukund Raj, P. K. (2013). Proprietary software versus open source software for education. *American Journal of Engineering Research*, 2(7), 124–130.

[11] *edutechnica.com*

Blockchain Traffic Offence Demerit Points Smart Contracts: Proof of Work

Aditya Pradana, Goh Ong Sing, Yogan Jaya Kumar, and Ali A. Mohammed
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—In Malaysia, a new regulation of traffic offences demerit points has been over a debate. Therefore, a blockchain model is formulated to solve this issue. It serves a purpose to be a Proof of Work (PoW) of a blockchain system. This model contains application layer and blockchain layer with smart contract inside. The smart contracts act as a conditional filter which follows the regulation rules. It contains three contracts starting from the declaration of each offence's demerit points and fines until the penalties when a certain amount of demerit points is collected, including revocation of driver license. The contracts will be automatically executed when such conditions are fulfilled. A transaction schema is also designed to match the schema of a traffic offence system. This model is deployed in online environment with two servers synced to each other to prove the decentralized characteristic of blockchain. It is developed using NodeJS while preserving JSON format for transaction between server and client. A user interface is also provided as a simulation media where a traffic officer can input offences and send it to blockchain server while public users or the driver itself can check the status of the driver license recorded on the blockchain. Government officer can monitor the records through a dashboard analytics provided which contains graphs and charts based on the records. This interface is used as media to do evaluation which produces satisfying results. The evaluation shows that the smart contracts are executed properly as compared to real regulations.**

***Keywords*—Blockchain; proof of work; smart contract; demerit points; decentralized system; distributed ledger**

I. Introduction

A revolution of peer-to-peer trust-free systems has been rising with blockchain technology on its core. This technology provides robust interactions among the peers which previously was maintained by third-party providers. It is able to replace the role of third-party providers by utilizing its smart contract which rules have been agreed by the involved parties. It opens up possibilities to be implemented in crucial party including government. Government policies which involve public as its object has been questioned with the existence of corruption by other parties. As new regulation is being implemented, it is better to see if blockchain technology can assist the regulation to be transparent and clear of any trust issues.

A case in Malaysia where a new regulation has been over debate by public is considered as an opportunity to test the capability of blockchain technology over the existing system. The regulation stated that road traffic offenders will be punished after a certain amount of demerit points reached. According to The Sun, driving licenses will be revoked when road traffic offenders reached 100 Kejara demerit points [1]. The details of each offences are shown in Table 1. A blockchain Proof-of-Work (PoW) is developed based on this regulation.

A blockchain or sometimes referred as trust-free technology [2] is a distributed ledger system with decentralized consensus mechanism. The ledger is cryptographically secured to avoid security issues on the records [3]. In technical perspective, it is also considered as a shared database to perform transactions among users without interference of an intermediary. The database contains transactional data stored in an encrypted infinite chain of blocks or ledgers. Each block contains the transaction along with timestamping algorithm to order it in time sequence [4]. The consensus mechanism requires a set of programmable rules dependent to the specific system to be developed. Nowadays, it is projected as the core of an alternative approach for trust-free systems. It has main characteristics such as distributed, immutable, trust-less, and decentralised. With its characteristics, it enables a verification of a record by public users without a need of intermediary institution [5]–[7]. These characteristics suit the main purpose of having an automated trust-less traffic offence system. Looking at its distributed characteristic, blockchain provides a wide variety of nodes or computers to be involved in a network in order to distribute the computing power. For example, Amazon AWS and Ethereum are two contrasting model. AWS is a private set of computing power bought and maintained by Amazon so only Amazon can contribute on the computers. Meanwhile, a blockchain organisation called Ethereum provides freedom by allowing anyone to install their software, so they can contribute to their network. Distribution helps to reduce risk of tampering, fraud and cybercrime [8]. Systems are more difficult to attack and taken down since more nodes are connected. Blockchain will provide transparency to the system since all parties involved have access to view the status of a driving license.

This research work is sponsored by UTeM Zamalah Scheme from Universiti Teknikal Malaysia Melaka.

TABLE II. KEJARA DEMERIT POINTS [1]

No.	Offence	Points	Fines
01	Driving under the influence of alcohol or drugs / Intoxication.	15 points	RM 300
02	Driving dangerous / reckless.	15 points	RM 300
03	Illegal racing on the road / Street Racing.	15 points	RM 300
04	Inconsiderate driving.	15 points	RM 300
05	Failure to provide breath, blood or urine sample when requested by a police officer, without a justifiable reason.	15 points	RM 300
06	Failed to follow traffic light.	10 points	RM 200
07a	Driving over the speed limit By 40 km/h	10 points	RM 200
07b	Driving over the speed limit By 26 km/h – 40 km/h	8 points	RM 150
07c	Driving over the speed limit By 1 km/h – 25 km/h	6 points	RM 100
08	Fail to give priority to ambulance, firefighter, police, custom, or Road Transaction Department car (with siren)	8 points	RM 150
09	Fail to stop at junction	8 points	RM 150
10	Offences related to overtaking and obstructing an overtaking vehicle.	8 points	RM 150
11	Offences committed at a pedestrian's crossing.	8 points	RM 150
12	Offences related to driving left lane.	8 points	RM 150
13	Careless driving.	8 points	RM 150
14	Ignore traffic sign or regulation.	5 points	RM 80
15	Using exhausted tire.	5 points	RM 80
16	Operating a motor vehicle on a cordoned off roadway.	5 points	RM 80
17	Overtaking at a double line.	5 points	RM 80
18a	Failure to keep the probationary license on one's person while operating a motor vehicle.	5 points	RM 80
18b	Failure to display identification at an easily-accessible place, according to the diagram in the sixth table of the rules.	10 points	RM 200
18c	Failure to keep the alcohol level in one's breath, blood and urine at 0.00.	5 points	RM 80

In trust-less characteristic, parties who involved in blockchain are able to perform digital transactions even if the parties do not trust each other. For example in the current system, central authorities like banks act as ledgers without the involvement of other parties with a purpose of avoiding the problem of duplication. They are not able to record or duplicate any transactions since central authorities have full control of it. Different with blockchain, it distributes the ledgers to may nodes and synchronise the ledgers using consensus method. It allows involving parties who don't trust each other can trust that the transaction is happening and valuable. It won't stop only to this kind of trust since the processes are shared among the parties and the records of transactions are immutable. It opens a massive range of potential digital transactions and possibilities that couldn't have happened with the current management style of central authorities. In a traffic offence system, government, police departments, and citizen might not have full trust in each other. It might even reduce a decision affected by political reason. Therefore, having blockchain is important.

Another important characteristic is its immutability. Removing or updating a transaction is impossible in blockchain because the transaction has been agreed upon and shared across the distributed network. Over time, when the transactions are increased, it becomes even harder to undo since the transactions are also shared across the network. Bitcoin as a public ledger provider allows everyone to explore the blockchain registered in Bitcoin and check the amount of Bitcoins in anyone's account. It also allows tracking of any transactions happened, thus it can be used to track supply chains. In other scenarios such as in networking perspective, blockchain can be used to check which users accessed certain files on a network [9], [10]. In a traffic offence system, any police officer cannot undo the demerit points and fines being given to a driver thus it reduces the possibilities of having bribe.

The last characteristic which makes blockchain technology different from other system is the decentralised network. As mentioned before, transactions are distributed across network which forms in a decentralised network. Decentralised network helps to reduce monopolies or issues with intermediary thus removes unnecessary costs for centralised investment. It allows increasing the competition in the market by increasing pressure on involved parties to become more efficient in doing any transactions. In addition, involving parties allowed to transact without requirement for trust disrupts the current management styles of central authorities who facilitate trusts such as banks and lawyers. Transactions directly between peers can reduce the step of involving middle-men which further increasing market efficiency. With this characteristic, a traffic offence system requires less interference of other parties which are not involved directly with the system or act as intermediary.

A blockchain needs another layer to make a use case works properly. With smart contract in this layer, a blockchain can have the rules based on regulation so that the transaction stored follows the regulation. Smart contracts are basically computer programs that can automatically execute the terms of a contract. When a pre-configured condition in a smart contract among involving parties is met then the parties involved in a contractual agreement can be automatically made transactions as per the contract in a transparent manner [11]. Smart contract makes the involvement of intermediary institution possible to be reduced. In 2015 Visa and DocuSign demonstrated smart contracts for leasing cars without the need to fill in forms. Smart contract in this PoW is applied based on the regulations where each offence has its own demerit points and fines, a driving license will get penalty after reaching 100 demerit points, and the penalties given are based on number of suspensions occurred. All of the contracts are automatically applied when police officer input an offence. It can detect the current state of a driving license and compare to the conditions of the contracts.

II. Blockchain Application Review

There is a number of blockchain system developed in the recent years. The first concept of blockchain system that was put in operation is the cryptocurrency [5]. Bitcoin is designed as an alternative way to perform independent transaction which traditionally involved governments, central authorities and banks. It is performed in form of electronic and peer-to-peer cash system.

Bitcoin was introduced in 2008. Since then, blockchain has started to have a new role from a mechanism to verify cryptocurrency to a broader field of commercial and economic applications. It has disruptive impact that is not limited to a specific industry [12] but rather enables the creation of a distributed, tamper-free, and transparent record of almost anything [13]. These factors are caused by its potential in reducing intermediary.

Another application related to Bitcoin is developed based on Bitcoin scripting language. It presented a reasonable protocol for transactional data where the commercial deal, the way data is delivered, and payment is performed, is essential. It is developed in such a way that if the buyer hasn't received the data then the seller is not able to redeem the payment and the buyer cannot receive the data if payment hasn't been made [14]. Another application which is inspired by Bitcoin is Cecoin. It follows the characteristics of Bitcoin such as immutable and public verifiable. Unlike Bitcoin, it utilised certificates to be treated as currencies and recorded on blockchain. The users can have access to a verification system. It verifies the validity of certificates based on smart contracts to ensure ownership consistency. It allows user to have multiple public-key certificates. Cecoin incorporates a modified algorithm based on Merkle Patricia tree which in later stage implemented as a distributed Certificate Library. It enables efficient retrieval and verification of certificates, and quick operations [15].

The application of blockchain is not always related to financial services but also Internet of Things (IoT). Hybrid-IoT is an application designed with a hybrid blockchain architecture for IoT. Then, the connection among the Proof of Work (PoW) sub-blockchains employs a BFT inter-connector framework, such as Polkadot or Cosmos. The PoW sub-blockchains formation, guided by a set of guidelines based on a set of dimensions, metrics and bounds [16]. Different application in IoT implemented and experimented a combination of bigdata and IoT technology to preserve renewable energy. Projects which are related to renewable energy are always get attention due to current environmental issues such as global warming. It introduces and implementation of blockchain technology for energy prosumer service model. This allows various energy sources to be connected to various users and producers. It also improves energy efficiency by analysing energy pattern of users. A transaction model that can collect, utilize, and process data more efficiently by combining the above technologies has also been presented [17].

In other domain such as industrial field, an application of blockchain technology related to the 4th Industrial Revolution (Industry 4.0) is presented with an example where blockchain is employed to facilitate machine-to-machine (M2M) interactions and launch a M2M electricity market in the context of the chemical industry. It involves one electricity consumer and two electricity producers performing transaction with each other in a blockchain environment. The research found out that this technology has significant under-researched potential to support and enhance the efficiency gains of the revolution and identifies areas for future research [18].

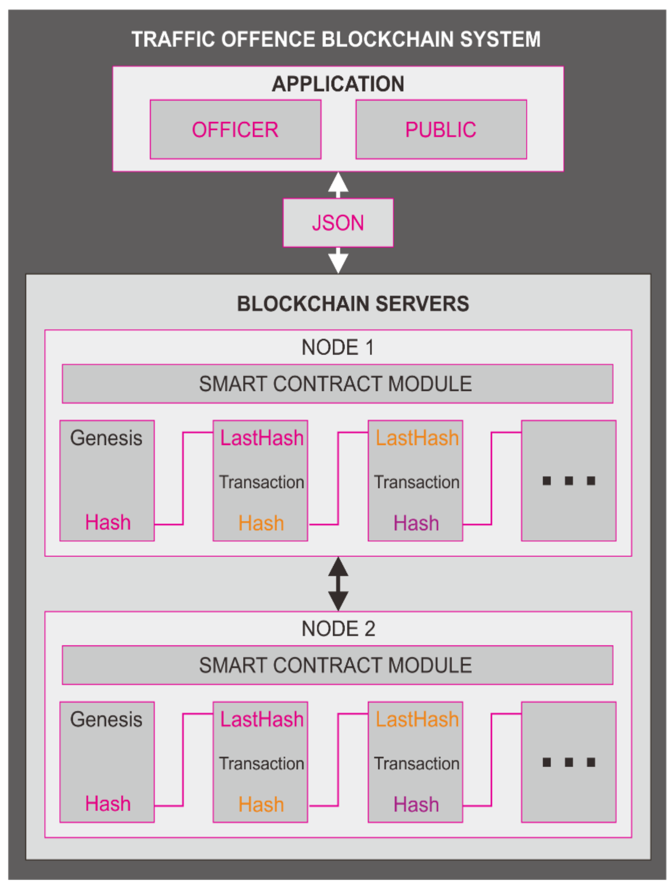


Fig. 1. Traffic Offence Blockchain Model.

Based on the reviews, it can be concluded that most of the blockchain implementations are related to financial and IoT. Although blockchain application is starting to be implemented in various fields [19], it hasn’t reach governmental regulation field such as traffic offence system.

III. Traffic Offence Blockchain Model

A model of traffic offence blockchain system is designed based on the requirements. As shown in Fig. 1, there are two main parts interact with each other by sending JSON objects namely Application layer and Blockchain Servers layer.

A. Application Layer

The first part is Application layer which acts as front-end interface to users such as police officer interface and public user interface. Both interfaces are connected to each other in which public users can retrieve the information or block that police officer has sent to blockchain servers.

B. Blockchain Servers Layer

For this PoW, two servers acting as nodes have been set up with Node 1 acts as the main server without removing the possibility to use Node 2 as the main server. Each node has the same layers inside with smart contract module as the first layer processing the transaction. It has been configured to match the regulations so that the transaction received is processed according to the configured conditions and allow the transaction proceeds to the blockchain layer. As its name suggests, it contains a chain of blocks which started with genesis block and followed by the transactions. A genesis block is a starting point of a blockchain which contains empty transaction and a pre-defined hash as agreed by involved parties. The hash will be used in the next block along with the transaction of the current block to create a hash of the current block. This hash will be used in the next block and the process repeated each time a transaction is received. In simple way, the hash of last block will link to the next block to make a chain. In each block contains transaction that has been processed by smart contract module. It contains its created time and a schema based on the input from police officer.


```
{
    "timestamp": 1529768832473,
    "lastHash": "004554370c634d640c8c64d4807c5b627a
                dbf92cd565a0fb6dfebe9b3c7e2a82",
        "hash": "007446fcd2417b8bf608d1f9041d4a520f
                6d47156b5a04616f0bd4a3ed45809a",
        "data": {
            "createdTime": 1529768832443,
            "cert": {
                "driverLicenseNumber": "SBH8451",
                "offenceDate": "01/18/2018",
                "offenceType": "13",
                "demeritPoints": 96,
                "isDriverLicenseSuspended": false,
                "noOfSuspended": 0,
                "penalty": "Active",
                "fines": 150,
                "totalFines": 1850
            },
            "error": []
        },
        "nonce": 616
}
```


Fig. 2. Transaction Schema in a Block.

C. Transaction Schema

A schema will be recorded in each block sent to a blockchain. The schema contains the details of a traffic offence. It will be based on inputs from traffic officer or any party who is involved in this system. The schema is designed based on requirement of a traffic offence system with added details to support smart contract implementation. An example of a block with the schema is shown in Fig. 2.

Transaction schema of the block is designed to match the real regulation rules which later will be used as input for smart contract. The details of the schema are as follows:

- Driver License Number: It shows the driver license number. It will act as identifier where the demerit points and other contract refer to.
- Date of Offence: It shows the occurrence date of the offence. It is stored in MM/DD/YYYY format.
- Offence Type: It contains selected offence type from 22 types as listed in Table 1.
- Demerit Points: It contains the sum of demerit points collected on each offence in a driving license.
- Fines: It contains total fines based on demerit points collected.
- Penalty: It shows the penalty given to the driver which refers to the third smart contract.
- Driver License Status: It may contain Active, Suspended or Revoked.

D. Smart Contract Model

Based on Malaysia regulation stated under Sections 35, 35A, 37 and 38, the 1987 Road Transportation Act and the Motor Vehicles (Demerit Points) Rules 1997, Demerit point is basically a way to reduce traffic offences and thus reduce road accidents. A smart contract model is developed based on the regulation. There are three contracts applied as Proof of Work such as:

- First contract: each traffic offence has its demerit points referring to the regulation and fines will be applied based on the demerit points as follows as shown in Table 1.
- Second contract: based on the accumulated demerit points, if a driver has exceeded 100 points, the driver license will be suspended and follows the penalty of the third contract.
- Third contract: Once a driver’s demerit points racked up to 100 points or above, the first driving license suspension will be given. The first suspension is no longer than 6 months. If after the suspension, a driver makes another offence regardless of the demerit points, it will get a penalty of driving license suspension for no longer than 12 months. If another offence is made within 5 years after the second suspension, then the driving license will be revoked.

E. Decentralised Network Model

Node 1 and Node 2 are connected in such a way that transactions between them are synchronised. It serves a purpose to enhance the security where attackers who attack one of the servers and somehow modified the block will be detected by consensus system that compares the blockchain with another server. If the blockchain hash is not identical to other servers then the modified block will be replaced by the original one. Furthermore, it will act as a backup module for the blockchain since the blockchain is not stored in a database. In case of one of the servers is down then the other server still keep the blockchain. Due to the blockchain is only stored in the memory, it will be erased when the server is down and when it is back online, the blockchain is empty and start from genesis block again. Therefore, decentralised module is needed since it will automatically synchronise the blockchain to the empty server.

IV. Development

A blockchain application is developed based on the model explained before. The development is started by setting up two servers to fulfil blockchain concept of decentralisation. Each server acts as a node of blockchain decentralised network. Node 1 is Ubuntu 16.4 server located in US and Node 2 is CentOS 5.4 server located in Netherland. Both servers have been installed Node JS framework to run Javascript scripts on server side. The function of sending (POST) and retrieving data (GET) is handled by Express JS as Web API.

The application has three main parts, such as block generator, schema, and blockchain generator. In schema part, each of schema detail is declared based on the input. The data type of each of the inputs is validated before sent to block generator to ensure no data type error. Block generator has a function to generate a block or ledger based on the schema. It generates a block which contains block number, timestamp, hash from last block, hash of current block, the schema itself, and a nonce. A timestamp is attached to each block in order to keep track when the block is being generated. Before generating any block from a schema, block generator will create an empty block called as genesis block. As stated in the previous section, genesis block is needed to start a blockchain thus the first block generated will hash it as part of Hash from Last Block. The hashing method is using SHA-256 since it is a one-way encryption thus difficult to decrypt. As for Current Hash, it uses Last Hash, current schema, and timestamp to generate a hash. Since Current Hash is based on Last Hash, it makes blockchain immutable to any changes. A change to a specific block will need hash from previous block and repeated until the genesis block which is almost impossible to achieve. Nonce is a number showing the occurrence number of mining process to reach a matching hash. Every time a block is retrieved, it will be added to blockchain with additional condition from smart contract module.

All of three main parts are executed in a main file which requires parameters such as Peers' IP Address, HTTP Port, and Web Socket. It is required to connect among nodes so that each record can be updated on each node. In this development, each server has been set up with one HTTP Port and one Web Socket. After the connection is established, any application can point to either nodes to send a block and both nodes will be updated accordingly.

A. Smart Contract Implementation

The first contract filters the condition of each offence to define its demerit points and fines. According to Table 1, a condition can be written as shown in pseudocode below.

Contract 1

IF offenceType = 01 || 02 || 03 || 04 || 05

THEN demeritPoints = 15, fines = 300

IF offenceType = 06 || 07a || 18b

THEN demeritPoints = 10, fines = 200

IF offenceType = 07b || 08 || 09 || 10 || 11 || 12 || 13

THEN demeritPoints = 8, fines = 150

IF offenceType = 07c

THEN demeritPoints = 6, fines = 100

ELSE

THEN demeritPoints = 5, fines = 80

The second contract is based on the first contract's demerit points. It will be executed if the total of demerit points reached 100. It will assign the variable of isDriverLicenseSuspended to be true and occurrence of suspension increased by 1 occurrence.

Contract 2

IF totalPoints >= 100

THEN isDriverLicenseSuspended = true

THEN noOfSuspended = +1

The third contract executed when second contract is fulfilled. Based on the number of suspensions, it will assign penalty to the driving license. When it reached 3 occurrences, the driving license will be revoked.

Contract 3

IF noOfSuspended = 1

THEN penalty = Suspended for 6 months

IF noOfSuspended = 2

THEN penalty = Suspended for 12 months

IF noOfSuspended = 3

THEN penalty = Suspended for 6 months and License to be revoked

IF noOfSuspended > 3

THEN penalty = Driver License revoked

B. Supporting User Interface

As shown in Fig. 1, the application layer contains an interface for police officer and public user. The interface is developed to show the input and output of the blockchain in readable form. The flow of registering a block is started by submitting offence details from Traffic Officer UI which contains Driver License Number, Offence Date, and Offence Type as shown in Fig. 3. As described in Fig. 1, the application layer sends data in JSON format to the server and retrieve the result in the same format.

The details will be processed in smart contract module to see if any conditions are met and then hashed and stored in blockchain. If the process is successful, the transaction will appear in the interface. Once an offence is registered as a block in blockchain, anyone can access it via Public interface as shown in Fig. 4. User can input the driving license number and check the status of it. An analytic dashboard is also provided to assist government to analyse the policies.

Fig. 3. Traffic Officer User Interface.

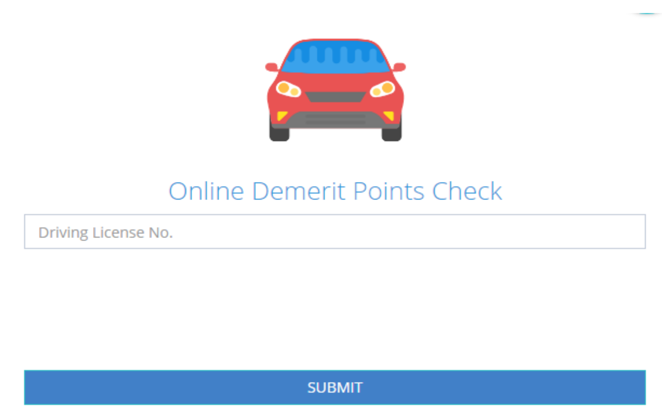


Fig. 4. Public User Interface.

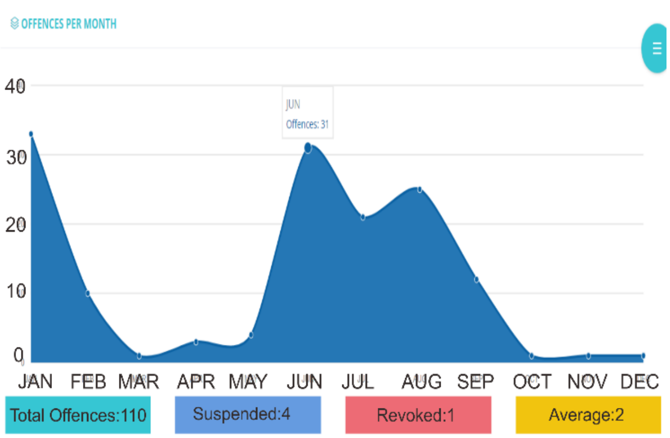


Fig. 5. Traffic Offences in a Year.

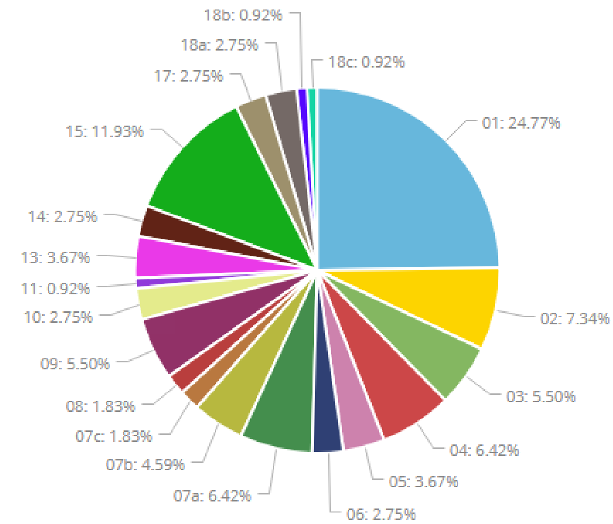


Fig. 6. Types of Offences Chart.

V. Evaluation

An evaluation is conducted using black-box approach to test the functionality of the system. This test has a purpose to check the rigidness of the smart contracts in filtering various conditions in transactions and to verify the creation of blocks in the blockchain. A set of instructions are given to a group of participants to ensure the variety of circumstances occurred on the blockchain.

A total of 110 blocks were received during the test. All of the blocks are properly recorded on the blockchain system proven by the charts shown in Fig. 5 and Fig. 6.

In Fig. 5, the chart provides offences on each month and few other statistics such as total offences, total license suspended, total license revoked, and average offences per month. The values are achieved from the test. It shows that the schema is stored properly as a block in blockchain and confirms the smart contracts are implemented properly with a total of 110 Offences, 4 License Suspended and 1 License Revoked. This test has shown that all smart contracts are executed. The chart shows total offences in January is 11, February is 14, March is 7, April is 4, May is 12, June is 11, July is 13, August is 6, September is 30, October is 1, November is 0, and December is 0. It can be concluded that there is a tendency of the users to pick the date close to the date when the test was conducted which is September.

The pie chart shown in Fig. 6 is provided to keep track on which offences have mostly occurred. It simply derives the Offence Type in the schema and sums it up. Similar to Fig. 5, the values are taken from the participants during test. It shows that most of the offences are Offence 01 which is “Driving under influence of alcohol / drugs” with 24.77%, Offence 15 which is “Using exhausted tire” with 11.93%, and Offence 02 which is “Driving dangerous/reckless” with 7.34%. Such statistics could assist the government to review the current policy and consider the next regulations to be composed.

```
{
    "no": 110,
    "difficulty": 2,
    "blockchain": [
        {
            "timestamp": 0,
            "lastHash": "000000",
            "hash": "f1r5t",
            "data": [],
            "nonce": 0
        },
```


Fig. 7. JSON Output of Total Block on Both Servers.

An evaluation is also conducted to check the decentralized model between Node 1 and Node 2. It checks the total blocks in the blockchain in each server as well as the first block after the genesis block and the last block in each server. It is to ensure the blockchain is synchronized between the servers.

Fig. 7 shows the total blocks in the blockchain i.e. 110 blocks. It is shown in JSON format as discussed in earlier section. It is retrieved from Postman application by sending GET request to Node 1 address which is https://myblockchain.asia:3001/blocks. The same output is retrieved from Node 2 address which is https://myblockchain.name:3001/blocks. It proves that all blocks are properly stored with the same number of blocks and thus the decentralized model is working well.


```
{
    "timestamp": 1536506064368,
    "lastHash": "f1r5t",
    "hash": "00e5758e2388adf51b3365627fe0dcc8a9af5d128abd0d222fa0954f98fce577",
    "data": {
        "createdTime": 1536506064295,
        "cert": {
            "driverLicenseNumber": "ASSA1234",
            "voilationDate": "08/27/2018",
            "voilationType": "01",
            "noOfVoilation": 15,
            "isDriverLicenseSuspended": false,
            "noOfSuspended": 0,
            "penalty": "Active",
            "fines": 300,
            "totalFines": 300
        },
        "error": []
    },
    "nonce": 447
},
```


Fig. 8. JSON Output of the First Block on Both Servers.


```
{
    "timestamp": 1536681004321,
    "lastHash": "00651625a41a9106b6166f310af0ea7973f134111cce7fe8af341b87c353e772"
    "hash": "00d9cf6a6ca7349b336f69405e1b383af0d909b92d3c26add1876452141ef1f4",
    "data": {
        "createdTime": 1536681004320,
        "cert": {
            "driverLicenseNumber": "30759188",
            "voilationDate": "01/24/2018",
            "voilationType": "15",
            "noOfVoilation": 70,
            "isDriverLicenseSuspended": false,
            "noOfSuspended": 0,
            "penalty": "Active",
            "fines": 80,
            "totalFines": 1240
        },
        "error": []
    },
    "nonce": 21
}
```


Fig. 9. JSON Output of the Last Block on Both Servers.

The next step is to look at the first block after genesis block whether Node1 and Node 2 produced identical first block. An output shown in Fig. 8 is retrieved by using the same request on each server. The first block contains the hash from genesis block which is “fir5t”. It also contains its own hash which is “00e5758e2388adf51b3365627fe0dcc8a9af5d128abd0d222fa0954f98fce577”. The nonce of this block is 447 which is a unique number appear once in a blockchain. Node 1 and Node 2 produced identical JSON output for the first block. In conclusion, the decentralized model has shown a success on this step.

The last step is to check on the last block of the blockchain on both servers. So far, the total blocks and first block are identical thus the last block should have similar output. Fig. 9 shows that hash of the last block is “00d9cf6a6ca7349b336f69405e1b383af0d909b92d3c26add1876452141ef1f4”. The hash is identical on Node 1 and Node 2. It implies that the decentralized model has been implemented properly and worked as expected.

VI. Discussion

From the evaluation conducted, it can be seen that the results reflect the blockchain model are properly implemented. All inputs are successfully created as blocks which are stored in a blockchain. Demerit points and fines received from the inputs are matched with the real regulations as shown in Table 1. It executed Contract 1 based on the Offence Type inserted by the users. There is four Driver License Suspended from the inputs. It shows that Contract 2 is executed by four different Driver License since its total demerit points exceed 100 points. Contract 3 is automatically executed once a Driver License has executed Contract 2 and executed another Contract 1. Based on the inputs, only one Driver License execute Contract 3. In conclusion, all smart contracts are properly executed. The decentralized model is also proven to synchronize blocks created to all servers.

In real-world situation the smart contracts will be very helpful since it is only defined once, and it can filter any inputs since then. However, the smart contracts must be agreed over the involved parties to ensure minimal modification in the future. It reduces human errors in deciding how many demerit points are assigned in a particular offence. It also eases the process of record keeping with thousands of driver license registered and their respective demerit points, fines, and status recorded. With blockchain characteristic of immutable, once a smart contract is triggered and record the transaction as a block, then no one can modify it. It reduces the chances of misusing the regulation for negative purposes. For example, in normal system a database record can be easily modified in order to reduce the demerit points, reduce the fines, or even change the status of the driver license. As long as a person has an access to the database, it can be done. However, in blockchain even if a person has access to the server, he cannot do anything with the record for several reasons such as each of the block stored in a blockchain is secured with unique hash. A hash is unique and based on timestamp of the generated block therefore when a block is regenerated on purpose, the timestamp will be different and thus the hash will also be different. Another reason is the consensus method applied in decentralized network model. Any changes in a server will be verified first and compared to other servers before executed. Having blockchain technology in a traffic offence system is also helpful to prevent attack from inside and outside.

VII. Conclusion and Future Work

A blockchain Proof of Work with smart contract has been developed. The system shows that it is possible to implement and apply blockchain in a government policy such as traffic offence regulation. The system has successfully developed to fulfil blockchain characteristics of being distributed, immutable, trust-less, and decentralised. It also shows that the details of traffic offence can be converted into a block schema and later cryptographically secured using hash to secure the transaction. Although the overall system has been proven to work properly, there might be issues with the performance if the setup is changed with additional nodes. Being a blockchain system means that forking issues still exists when clients mining the data and found the identic nonce. Therefore, it is required to have a real-world performance test to this model before implemented in real policies.

Acknowledgment

This research work is sponsored by UTeM Zamalah Scheme from Universiti Teknikal Malaysia Melaka.

References

[1] C. Ramendran, “Offenders who chalk up 100 Kejara demerit points will have driving licence revoked,” The Sun, 2017. [Online]. Available: http://www.thesundaily.my/news/2219522. [Accessed: 06-Jul-2018].

[2] R. Beck, J. S. Czepluch, N. Lollike, and S. Malone, “Blockchain – the Gateway To Trust- Free Cryptographic Transactions,” Twenty-Fourth Eur. Conf. Inf. Syst., pp. 5–16, 2016.

[3] M. Risius and K. Spohrer, “A Blockchain Research Framework,” Bus. Inf. Syst. Eng., vol. 59, no. 6, pp. 385–409, 2017.

[4] B. Gipp, N. Meuschke, and A. Gernandt, “Decentralized Trusted Timestamping using the Crypto Currency Bitcoin,” pp. 1–6, 2015.

[5] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System,” Www.Bitcoin.Org, p. 9, 2008.

[6] B. Notheisen, J. B. Cholewa, and A. P. Shanmugam, “Trading Real-World Assets on Blockchain,” Bus. Inf. Syst. Eng., vol. 59, no. 6, pp. 425–440, 2017.

[7] T. Puschmann and R. Alt, “Sharing Economy,” Bus. Inf. Syst. Eng., vol. 58, no. 1, pp. 93–99, 2016.

[8] M. Nasser et al., “Cyber-Security Incidents: A Review Cases in Cyber-Physical Systems,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, 2018.

[9] M. F. Ali, N. Azman, and N. Harum, “A Novel Multiple Session Payment System,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, pp. 237–245, 2018.

[10] 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.

[11] D. Efanov and P. Roschin, “The all-pervasiveness of the blockchain technology,” Procedia Comput. Sci., vol. 123, pp. 116–121, 2018.

[12] D. Wörner, T. von Bomhard, Y.-P. Schreier, and D. Bilgeri, “The Bitcoin Ecosystem: Disruption Beyond Financial Services?,” in ECIS 2016 Proceedings, 2016.

[13] R. Böhme, N. Christin, B. Edelman, and T. Moore, “Bitcoin: Economics, Technology, and Governance,” J. Econ. Perspect., vol. 29, no. 2, pp. 213–238, 2015.

[14] S. Delgado-Segura, C. Pérez-Solà, G. Navarro-Arribas, and J. Herrera-Joancomartí, “A fair protocol for data trading based on Bitcoin transactions,” Futur. Gener. Comput. Syst., 2017.

[15] B. Qin, J. Huang, Q. Wang, X. Luo, B. Liang, and W. Shi, “Cecoin: A decentralized PKI mitigating MitM attacks,” Futur. Gener. Comput. Syst., 2017.

[16] G. Sagirlar, B. Carminati, E. Ferrari, J. D. Sheehan, and E. Ragnoli, “Hybrid-IoT: Hybrid Blockchain Architecture for Internet of Things - PoW Sub-blockchains,” pp. 1–10, 2018.

[17] J. Hwang et al., “Energy Prosumer Business Model Using Blockchain System to Ensure Transparency and Safety,” Energy Procedia, vol. 141, pp. 194–198, 2017.

[18] J. J. Sikorski, J. Haughton, and M. Kraft, “Blockchain technology in the chemical industry: Machine-to-machine electricity market,” Appl. Energy, vol. 195, pp. 234–246, 2017.

[19] F. Hawlitschek, B. Notheisen, and T. Teubner, “The limits of trust-free systems: A literature review on blockchain technology and trust in the sharing economy,” Electron. Commer. Res. Appl., vol. 29, pp. 50–63, 2018.

Features Optimization for ECG Signals Classification

Alan S. Said Ahmad[1, *]
Department of Physics, Faculty of Science
University of Zakho
Kurdistan Region, Iraq

Majd Salah Matti[2], Adel Sabry Essa[3]
Department of Computer Science, Faculty of Science
University of Zakho
Kurdistan Region, Iraq

Omar A.M. ALhabib[4]
College of Science
International University of Erbil
Kurdistan Region, Iraq

Sabri Shaikhow[5]
Department of Cardiology
Azadi Teaching Hospital Duhok
Kurdistan Region, Iraq

***Abstract*—A new method is used in this work to classify ECG beats. The new method is about using an optimization algorithm for selecting the features of each beat then classify them. For each beat, twenty-four higher order statistical features and three timing interval features are obtained. Five types of beat classes are used for classification in this work, atrial premature contractions (APC), normal (NOR), premature ventricular contractions (PVC), left bundle branch (LBBB) and right bundle branch (RBBB). Cuttlefish algorithm is used for feature selection which is a new bio-inspired optimization algorithm. Four classifiers are used within CFA, Scaled Conjugate Gradient Artificial Neural Network (SCG-ANN), K-Nearest Neighborhood (KNN), Interactive Dichotomizer 3 (ID3) and Support Vector Machine (SVM). The final results show an accuracy of 97.96% for ANN, 95.71% for KNN, 94.69% for ID3 and 93.06% for SVM, these results were tested on fourteen signal records from MIT-HIH database, where 1400 beats were extracted from these records.**

Keywords—Features optimization; cuttlefish; ECG; ANN-SCG; ID3; KNN; SVM

I. Introduction

Automatic diagnosis of electrocardiogram (ECG) it is very important in the field of heart disease diagnosis, that is why feature extraction and classification it is an important step to achieve a good diagnosis [1, 2].

Many techniques have been proposed to classify ECG beat using data preprocessing, feature extraction, and classification algorithms. Some of these techniques are, Ali Kraiem and Faiza Charfi have used C4.5 technique to classify ECG beats using morphological features from signals denoised by band pass filter [3]. Ataollah Ebrahimzadeh and Ali Khazaee they have used wavelet transform and time interval features with radial base function for classification of five types of beats [4]. Ataollah Ebrahim and Ali Khazaee they have proposed a method for using morphological and time features with support vector machine for classification of 5 beat types [5]. Yakup Kutlua and Damla Kuntalp used nearest neighborhood (KNN) for classification of 5 beat types and higher order statistic features [6]. Ali Khazaee have used genetic algorithm with radial base function for classification of 3 beat types and morphological and timing interval features [7]. Ebrahimzadeh, Shakiba and Khazaee used higher order statistics with time interval features with radial base function and bees algorithm for classification of 5 beats [1]. Raju, Rao and Jagadesh used discrete wavelet transform as features and PCA with ANN to classify 5 beat types [8]. Inbalatha and Kalaivani have used wavelet transform with principle component analysis for features and K-nearest neighborhood for classification of 2 beat types [9]. Alan and Majd have proposed a method of using higher order statistics with time intervals as features and Artificial Neural Networks to classify 5 arrhythmia beat types [10].

Another strategy has been proposed in this work. This technique is comprising of four stages. To start with, ECG flag preprocessing utilizing denoising dependent on Discrete Wavelet Transform (DWT). The database of the flag records that are utilized is MIT-BIH database [11] in which two atrial untimely constrictions (APC) records are chosen, three ordinary (NOR) records, three untimely ventricular compressions (PVC) records, three remaining group branch (LBBB) records and three right package branch (RBBB) records are utilized. Second, highlights extraction from each flag's beat and standardized for advancement and grouping, these twenty-four higher request factual and three planning interim highlights will be utilized. Third, include choice by utilizing Cuttlefish improvement calculation. Fourth, grouping utilizing Artificial Neural Network Scaled Conjugate Gradient (ANN-SCG) classifier calculation. Figure 1 delineates this work.

The remainder of this paper is as follows: section 2 contains an overview of the preprocessing technique used, Section 3 talks about the feature extraction process, and section 4 explains the usage of optimization with classification, while section 5 describes the datasets used, and sections 6 illustrated the results and discussion. Finally, section 7 describes the final conclusions.

*Corresponding Author.

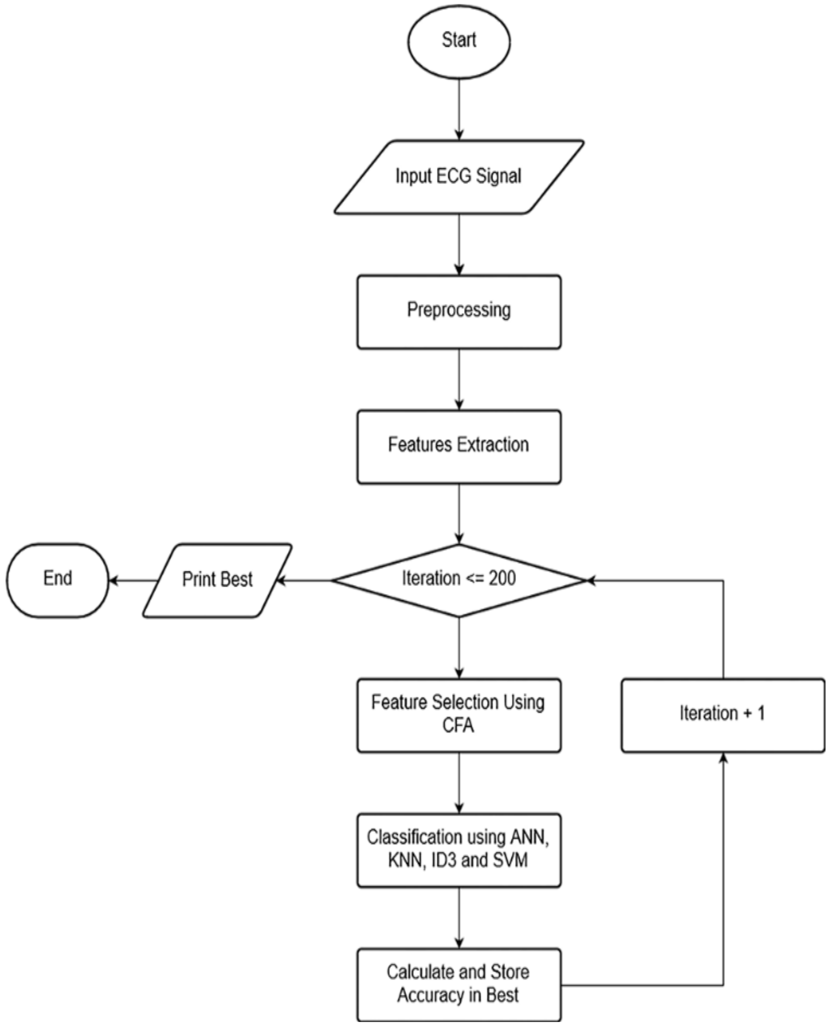


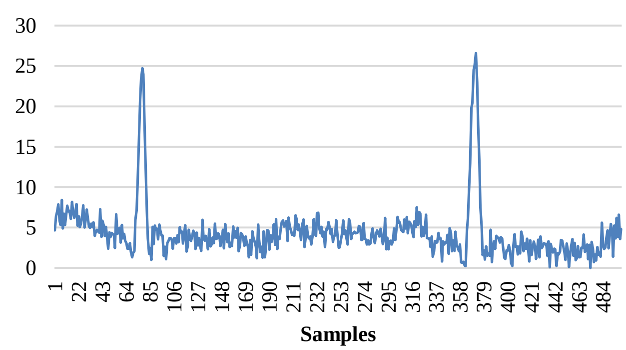
Fig. 1. The proposed system's diagram.

Fig. 2. Block Diagram of DWT.

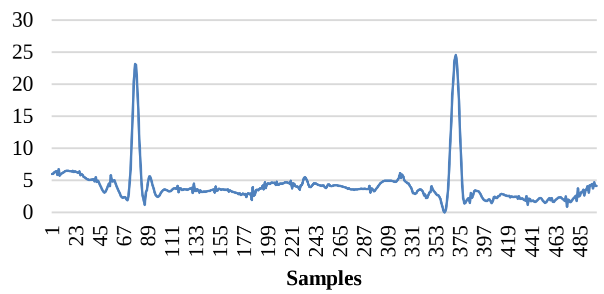
II. Preparing Signals for Preprocessing

Noise elimination from signals is very important and challenging step in signal processing [12]. Many techniques are available for noise elimination as filtering, thresholding and others. The used denoising technique in this work is wavelet shrinkage DWT method for its effective denoising results and minimum computation complexity [13]. Where denoising refers to removing noise from signal [14, 15] as in figure 3. DWT is consisting of three main steps, first signal analysis to its approximation and detailed coefficients. Second, thresholding the details coefficients. Third, synthesis of the analyzed signal to its original signal [16] as illustrated in the block diagram in figure 2.

*Corresponding Author.



Noisy Signal



Denoised Signal

Fig. 3. The denoising process on any signal.

III. Extraction of Features

A. Higher Order Statistics

Higher order statistics it's a very popular and good tool used to extract some features from nonlinear signals as ECG signal. The first and second order statistics are not sufficient for representing nonlinear signals. So the third and fourth order statistics will be used to represent each beat selected from ECG record. While variance and mean are contained in the first and second order statistics, higher order statistics contain higher order moments and cumulants [1, 6].

To extract these features, R peak must be detected at first. In any ECG signal a windows of size -300 ms to 400 ms is used which represents a 252 samples, this 252 samples will be normalized to mean of zero and standard deviation of unity. Then these 252 samples will be grouped into eight small groups starting from 30-45, 45-83, 84-112, 112-122, 122-145, 150-205, 207-225 and 230-252. Then for each small group second, third and fourth order of the cumulant is calculated. As a result we will have 24 statistical features as detailed in figure 3 [1, 6].

-300:400 window = 252 samples

30-45	45-83	84-112	112-122	122-145	150-205	207-225	230-252
2nd 3rd 4th order statistics	2nd 3rd 4th order statistics	2nd 3rd 4th order statistics	2nd 3rd 4th order statistics	2nd 3rd 4th order statistics	2nd 3rd 4th order statistics	2nd 3rd 4th order statistics	2nd 3rd 4th order statistics

Fig. 4. Explanation of the statistical features.

B. Timming Features

Three timing features are extracted using equations (1), (2) and(3) for every R peak, these features are: previous time interval RR_{prev} and next time interval RR_{next} and time interval ratio IR. Where IR_i is the current time interval ratio, T_i is the current R peak, T_{i-1} is the previous R peak and T_{i+1} is the next R peak [4, 5]. Figure 4 explains the three used timing intervals. The final number of feature will be 27.

*Corresponding Author.

$$RR_{prev} = T_i - T_{i-1} \quad (1)$$

$$RR_{next} = T_{i+1} - T_i \quad (2)$$

$$IR_i = \frac{RR_{prev}}{RR_{next}} \quad (3)$$

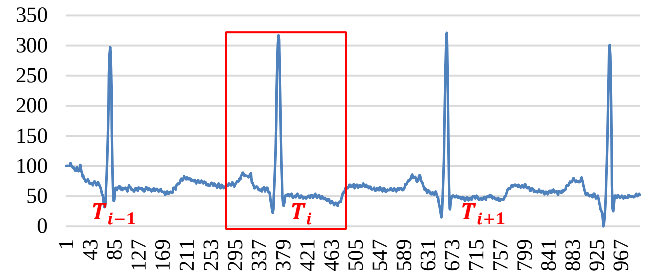


Fig. 5. Explanation of the timing features.

IV. Cuttlefish Optimization Algorithm (CFA)

It is an optimization algorithm that is inspired by the behavior of color changing mechanism of the cuttlefish, to find the optimal solution for any problem [17, 18].

The skin of cuttlefish is comprising of three distinct layers of cells including chromatophores, leucophores and iridophores. Each layer gives distinctive shading when the light occurrence on the skin. CFA utilizes two fundamental procedures which are reflection and perceivability. The reflection procedure is the reenactment of the light reflection system, while perceivability is the recreation of perceivability of coordinating examples. These two procedures are utilized to find the worldwide ideal arrangement in the calculation. The graph in figure 6, outlining the skin structures (chromatophores, iridophores and leucophores) with two model states (upper, lower) and three unmistakable beam follows (1, 2, 3), demonstrates the modern means by which cuttlefish can change reflective shading [19].

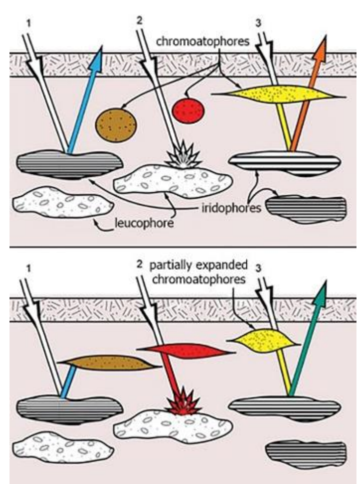


Fig. 6. The main three skin structures (chromatophores, iridophores and leucophores) of CFA.

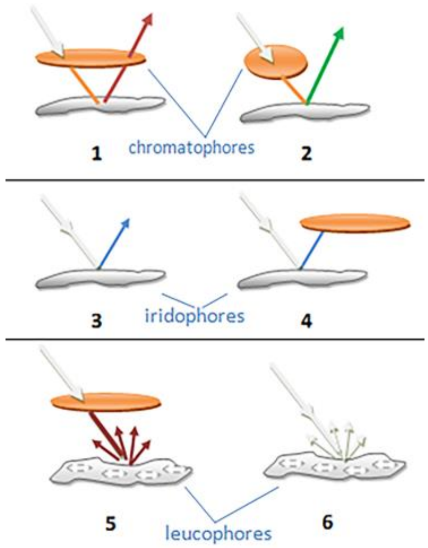


Fig. 7. Reorder of the 6 cases in figure 6.

A. Initialization

The algorithm starts by creating and initializing the population P of size N with random subsets, where each subset is consisting of two parts selectedFeatures and unselectedFeatures. Followed by the calculation of fitness value for each subset in the population. Finally, keep the best solution or (best subset) in bestSubset and avBestSubset, then remove one feature from selectedFeatures in bestSubset. After initialization the CFA will perform its main function which is consists of six cases, these cases are illustrated in figure 7, and in the next steps:

B. Group 1, Case 1 and 2

In the first group which include cases 1 and 2, the algorithm will start by descending sort of the population *P* based on fitness value for each subset in the population. Part of the population will be selected for applying CFA equations starting at $i = 1,2,\ldots,K$. Where K is a random number selected between 0 and N/2.

New subset for each i will be generated by combining Reflection (R) and Visibility (V) as in equation (4), where R is the degree of reflection and V is the visibility degree of the final view of the matched pattern. The main reflection and visibility equations of the algorithm for the current cases are as follows:

$$newSubset_i = Reflection_i \oplus Visibility_i \quad (4)$$

$$Reflection_i = random\ (subset[R]) \subset P_i(selectedFeatures) \quad (5)$$

$$Visibility_i = random\ (subset[V]) \subset P_i(unselectedFeatures) \quad (6)$$

where $Reflection_i$ is a subset with size of R and its elements are selected randomly from selectedFeatures and $Visibility_i$ is another subset with size of V whose elements are created randomly from unselectedFeatures. Where R and V values can be calculated as follows:

$$R = random\ (0, size\ of\ selectedFeatures)$$

$$V = size\ of\ selectedFeatures - R$$

C. Group 2, Case 3 and 4

In the second group which include cases 3 and 4, to calculate the Reflection we have only to remove one element from selectedFeatures of bestSubset as in equation (6) where the visibility can be calculated by selecting only one feature from unSelectedFeatures of bestSubset as in equation (7) and finally by combining the results of these two equations a new subset will be produced.

These two equations will be repeated T times, where T is a small fixed number selected from the size of selectedFeatures of bestSubset. If newSubset is better than bestSubset then replace newSubset with bestSubset.

$$Reflection = bestSubset(selectedFeatures) - bestSubset(selectedFeatures)\ [R] \quad (6)$$

$$Visibility = bestSubset(unselectedFeatures)[V] \quad (7)$$

where R is the index of the removed feature from selectedFeatures of bestSubset, and V is the index of selected feature from unselectedFeatures of bestSubset. The calculation of R and V are as follows:

$$R = random\ \big(0, size(bestSubset(selectedFeatures))\big)$$

$$V = random\ \big(0, size(bestSubset(unSelectedFeatures))\big)$$

D. Group 3, Case 5

In the third group which is about case 5, the new subset will be generated by the combination of the equations (9) and (10) into equation (8). These equations will be repeated to size of selectedFeatures of avBestSubset times. The values of reflection are equals to selectedFeatures in avBestSubset only and the value of visibility will be a single value selected from selectedFeatures in avBestSubset according to index i.

In this way, we can produce R new subsets, the value of R is equal to the size of selectedFeatures each subset representing the matched pattern by removing one feature from selectedFeatures at each time. If newSubset is better than bestSubset then replace newSubset with bestSubset.

$$newSubset_i = Reflection - Visibility_i \quad (8)$$

$$Reflection = AVbestSubset - selectedFeatures \quad (9)$$

$$Visibility_i = AVbestSubset - selectedFeatures[i] \quad (10)$$

where, i is the index of the feature in selectedFeaturest hat will be removed. $i = 1,2,\ldots,R$. Where R is the size of selectedFeatures.

E. Group 4, Case 6

In the fourth gathering which is case 6, any approaching shading from the earth will be reflected as it very well may be spoken to by any irregular arrangement. In the underlying calculation, this case is utilized to produce irregular arrangements. Likewise, we utilize this case as an irregular generator procedure to create arbitrary arrangements. The quantity of ages is equivalent to N-K.

The new age will be begun at the area K in the wake of arranging the populace P in diving request. On the off chance

*Corresponding Author.

that the newSubset is superior to the current avBestSubset, the current avBestSubset is supplanted with newSubset. The procedure of irregular arrangements is the equivalent as that utilized with the introduction procedure which is depicted previously.

F. *Stopping Condition*

The algorithm will stop when the iterations number reaches its max limit which is 50 iterations.

G. *Fitness Function*

CFA needs to evaluate every subset of the population and assign a fitness value to it. In this work four different leaner algorithms or classifiers are used where each is used separately with the CFA.

- ANN

Artificial Neural Network it is a very popular technique in the field of classification and machine learning, because its performance is very good [20]. The general structure of ANN is composed three main layers one input layer, one or more hidden layers and one output layer. In this work the used structure of the network is one input layer with N input neurons because the CFA will control the number of input layer, for the hidden layer we have two layers each one is of 40 neurons and one output layer with 5 neurons due to the five classes we have, as in figure 8. Scaled Conjugate Gradient classifier algorithm is used which is a supervised classification algorithm. The advantages of this classifier are its fast speed and it does not require a lot of memory [21, 22].

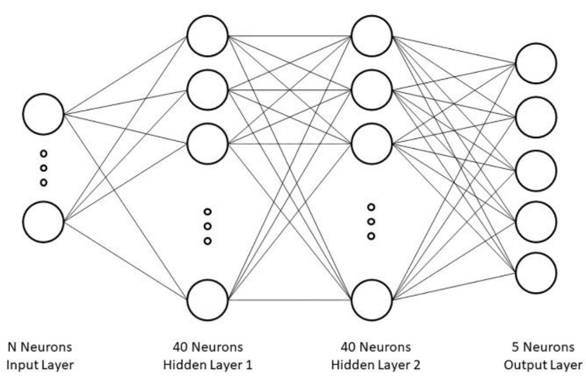


Fig. 8. The structure of the used neural network.

- KNN

The K-Nearest Neighbor is statistical classification algorithms based on closest training examples in the feature space [23]. The KNN algorithm has few parameters to set for classification [23]. These parameters are K and Distance metric. Where of K is a value which is selected in a way that it gives maximum classification accuracy [23]. Whereas the Distance metric parameter used in this work is Minkowski.

The Minkowski provides a concise, parametric distance function that generalizes many of the distance functions available as Euclidean, Manhattan, Chebyshev, … etc. The advantage is that mathematical results can be shown for a whole class of distance functions, and the user can adapt the distance function to suit the needs of the application by modifying the Minkowski parameter [24].

- ID3

The Interactive Dichotomizer 3 (ID3) it is a decision tree learner algorithm. Decision Trees (DTs) are accurate and small, in which results in a reliable and fast classification results. Because of its speed and reliability, decision trees are popular classification tool [25]. DTs are used to build classification rules in the form of top down Decision Tree [26]. Where the leafs contains class names and non-leafs are decision nodes [26].

- SVM

Support Vector Machines in general are binary classifiers, but you may want to classify your data into more than two classes [27]. To solve this problem multiclass SVM is needed which is done by using the One Against One(OAO) approach.

The OAO strategy, otherwise called (pairwise coupling), (all sets) or (round robin), comprises in building one SVM for each combine of classes. Hence, for an issue with n classes, n(n-1)/2 SVMs are prepared to recognize the examples of one class from the examples of another class. Grouping of an obscure example is finished by the greatest casting a ballot, where each SVM votes in favor of one class [27].

V. Datasets And Evaluation Metering

A. *Datasets Used*

MIT-BIH database [11] is used to use some ECG data records. In table 1 you have 5 classes in the first column, each class represents a specific arrhythmia disease. Eight records have were selected from the database and distributed on each class [1] as in table 1. From each record signal used 65 beats were selected for training and 35 for testing, more details are in table 1.

TABLE I. Summery of the Data used and Its Divisions

Class name	**Selected records**	**Used beats in training**	**Used beats in testing**
NOR	100m, 105m, 215m	65, 65, 65	35, 35, 35
PVC	207m, 209m, 232m	65, 65, 65	35, 35, 35
APC	106m, 223m	65, 65	35, 35
LBBB	109m, 111m, 214m	65, 65, 65	35, 35, 35
RBBB	118m, 124m, 212m	65, 65, 65	35, 35, 35
		910 beat	490 beat

B. *Evaluation Criteria*

For this work four equations have been used to evaluate the performance:

- The sensitivity:

$$Se = \frac{T_P}{T_P+F_N} \times 100\% \quad (11)$$

*Corresponding Author.

- The specificity:

$$Sp = \frac{T_N}{T_N + F_P} \times 100\% \quad (12)$$

- The positive predictivity:

$$Pp = \frac{T_P}{T_P + F_P} \times 100\% \quad (13)$$

- The accuracy: It is very important for testing the performance [2].

$$Acc = \frac{T_P}{Total\ number\ of\ beats} \times 100\% \quad (14)$$

Where T_P (true positive) represents the correct classified beats number for any class, F_N (false negative) represents the incorrect classified beats number in the other used classes, T_N (true negative) represents the correct classified beats number for all other classes and F_P (false positive) represents the incorrect classified beats number for any class.

VI. Discussion and Results

The final results of this work will be illustrated in this section. As a total number of beats 1400 beat is selected for training and testing the proposed system. Fourteen ECG records were used and distributed on five classes as in table 1, 100 beat is extracted from 106 and 223 records for APC, a total of 300 beats were selected from NOR records (100, 105 and 215) 100 from each, for PVC records (207, 209 and 232) 100 beat is selected from each record, for LBBB records (109, 111 and 214) 100 beat is selected from each record, and for RBBB records (118, 124 and 212) 100 beat is selected from each record.

In this work CFA is used optimization for features and ANN classifier is used to classify features into five classes. Figure 9 gives a good vision of the performance of the proposed method in this work for 50 iterations of CFA and Figure 10 gives the comparison with the works of others.

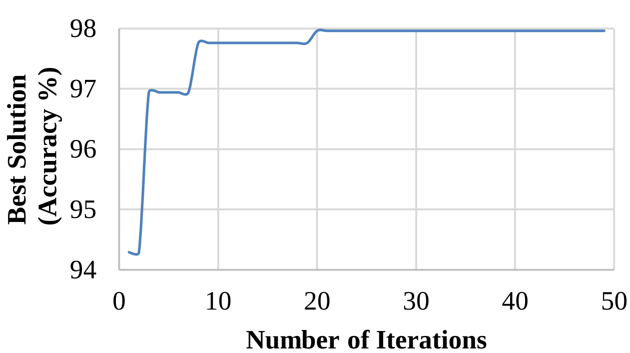


Fig. 9. The progress of best solution in every iteration.

As illustrated in figure 9 the final accuracy was 97.96% after training, which is a very accurate and high result when this result is compared with others in the same field. Also the third column in table 3 explains the accuracy of all class types used, 97.14% for APC, 98.10% for LBBB, 97.14% for NOR, 98.10% for PVC and 99.05% for RBBB.

Table 2 gives the complete statistics of the total number of beats used for these variables : true positive (T_P), false negative (F_N), true negative (T_N) and false positive (F_P).

TABLE II. Full Details About the Number of Beats used and the Correctly and Incorrectly Classified Beats for Each Class

	Total Beats	TP	TN	FP	FN
APC	70	68	412	2	8
LBBB	105	103	377	2	8
NOR	105	102	378	3	7
PVC	105	103	377	2	8
RBBB	105	104	376	1	9

The sensitivity (Se), specificity (Sp) and positive predictivity (Pp) for each class type is detailed in table 3.

TABLE III. The Classes used and the Sensitivity, Specificity and Accuracy for Each Class

Class Type	Se %	Sp %	Pp (Acc) %
APC	89.47	99.52	97.14
LBBB	92.79	99.47	98.10
NOR	93.58	99.21	97.14
PVC	92.79	99.47	98.10
RBBB	92.04	99.73	99.05
Overall	92.13	99.48	97.96

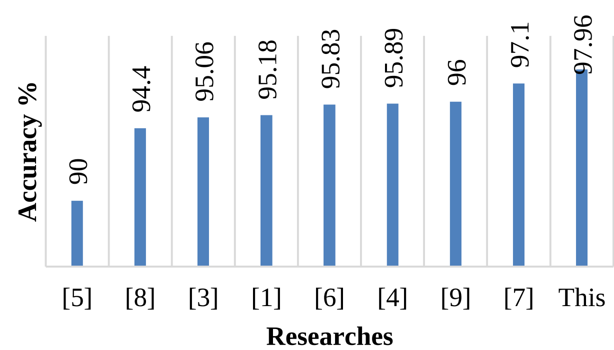


Fig. 10. Comparison with the work of other researchers.

As stated before, four classifiers have been used in this work ANN, KNN, ID3 and SVM each gave different accuracy after training phase as listed in table 4.

TABLE IV. The Comparison of Accuracy Obtained from the Four used Classifiers

ANN	KNN	ID3	SVM
97.96 %	95.71 %	94.69 %	93.06 %

VII. Final Conclusions

This work is about using a hybrid technique for classifying ECG signals into five classes, using 27 features (24 statistical with 3-time interval features). Fourteen signals were selected from MIT-BIH database [11], the records were selected according to five ECG arrhythmia classes. Four classifiers are

*Corresponding Author.

used to show a comparison of each one, ANN-SCG, KNN, ID3 and SVM are used as classifiers.

1) From the experiments of this work it is shown that the usage of CFA optimization algorithm with ANN-SCG learner algorithm gives best classification results than ID3, KNN and SVM.

2) The comparison is done with different works, each work used different technique but ANN-SCG with CFA is the best choice.

References

[1] Ebrahimzadeh, A., B. Shakiba, and A. Khazaee, Detection of electrocardiogram signals using an efficient method. Applied Soft Computing, 2014. 22: p. 108-117.

[2] Li, H., et al., Genetic algorithm for the optimization of features and neural networks in ECG signals classification. Scientific Reports, 2017. 7: p. 41011.

[3] Kraiem, A. and F. Charfi, Arrhythmia Classification from ECG Signals Using Data Mining Approaches. 2009: p. 7.

[4] Zadeh, A.E. and A. Khazaee, An Efficient Technique for Classification of Electrocardiogram Signals. Advances in Electrical and Computer Engineering, 2009. 9(3): p. 5.

[5] Zadeh, A.E. and A. Khazaee, High Efficient System for Automatic Classification of the Electrocardiogram Beats. Annals of Biomedical Engineering, 2011. 39(3): p. 16.

[6] Kutlu, Y. and D. Kuntalp, Feature extraction for ECG heartbeats using higher order statistics of WPD coefficients. Computer Methods and Programs in Biomedicine, 2012. 105(3): p. 257-267.

[7] Khazaee, A., Automated Cardiac Beat Classification Using RBF Neural Networks. I.J.Modern Education and Computer Science, 2013. 5: p. 7.

[8] Raju, N.N.S.V.R., V.M. Rao, and B.N. .Jagadesh, Identification and Classification of Cardiac Arrhythmia using Neural Network. Helix, 2017. 7(5): p. 6.

[9] Mandala, S., et al. Performance analysis of wavelet-based denoising techniques for ECG signal. in 2017 5th International Conference on Information and Communication Technology (ICoIC7). 2017.

[10] Ahmad, A.S.S., et al., ECG Signal Classification Using Scaled Conjugate Gradient Learner Algorithm. International Journal of Medical Research & Health Sciences, 2018. 7(5): p. 7.

[11] Goldberger, A.L., et al., PhysioBank, PhysioToolkit, and PhysioNet. Components of a New Research Resource for Complex Physiologic Signals, 2000. 101(23): p. 215-220.

[12] Alyasseri, Z.A.A., et al., Hybridizing B-hill climbing with wavelet transform for denoising ECG signals. Inf. Sci., 2018. 429(C): p. 229-246.

[13] Srivastava, M., C.L. Anderson, and J.H. Freed, A New Wavelet Denoising Method for Selecting Decomposition Levels and Noise Thresholds. IEEE Access, 2016. 4: p. 3862-3877.

[14] Ahmad, A.S.S., et al., Denoising of Arrhythmia ECG Signals. International Journal of Medical Research & Health Sciences, 2018. 7(3): p. 11.

[15] Sadooghi, M.S. and S.E. Khadem, A New Performance Evaluation Scheme for Jet Engine Vibration Signal Denoising. Mechanical Systems and Signal Processing, 2016. 76(77): p. 201-212.

[16] Donoho, D.L. and I.M. Johnstone, Adapting to Unknown Smoothness via Wavelet Shrinkage. Journal of the American Statistical Association, 1995. 90(432): p. 1200-1224.

[17] Eesa, A.S., A.M.A. Brifcani, and Z. Orman, A New Tool for Global Optimization Problems - Cuttlefish Algorithm. International Journal of Computer and Information Engineering, 2014. 8(9): p. 5.

[18] Eesa, A.S., Z. Orman, and A.M.A. Brifcani, A Novel Feature Selection Approach Based on The Cuttlefish Optimization. Expert Systems with Applications, 2014. 42: p. 10.

[19] Kreit, E., et al., Biological versus electronic adaptive coloration: how can one inform the other? Journal of The Royal Society Interface, 2012.

[20] AlHamouz, S. 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.

[21] Khadse, C.B., M.A. Chaudhari, and V.B. Borghate. Comparison between three back-propagation algorithms for power quality monitoring. in 2015 Annual IEEE India Conference (INDICON). 2015.

[22] Mohamad, N., et al. Comparison between Levenberg-Marquardt and Scaled Conjugate Gradient training algorithms for Breast Cancer Diagnosis using MLP. in 2010 6th International Colloquium on Signal Processing & its Applications. 2010.

[23] Saini, I., D. Singh, and A. Khosla. Delineation of ECG Wave Components Using K-Nearest Neighbor (KNN) Algorithm: ECG Wave Delineation Using KNN. in 2013 10th International Conference on Information Technology: New Generations. 2013.

[24] Chouikhi, H., M.F. Saad, and A.M. Alimi. Improved fuzzy possibilistic C-means (IFPCM) algorithms using Minkowski distance. in 2017 International Conference on Control, Automation and Diagnosis (ICCAD). 2017.

[25] Salzberg, S.L., C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning, 1994. 16(3): p. 235-240.

[26] Eesa, A.S., Z. Orman, and A.M.A. Brifcani, A new feature selection model based on ID3 and bees algorithm for intrusion detection system. Turkish Journal of Electrical Engineering and Computer Science, 2015. 23(2): p. 8.

[27] Faziludeen, S. and P.V. Sabiq. ECG beat classification using wavelets and SVM. in 2013 IEEE Conference on Information & Communication Technologies. 2013.

*Corresponding Author.

Handling Class Imbalance in Credit Card Fraud using Resampling Methods

Nur Farhana Hordri[1], Siti Sophiayati Yuhaniz[2], Nurulhuda Firdaus Mohd Azmi[3]
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia

Siti Mariyam Shamsuddin[4]
UTM Big Data Centre
Universiti Teknologi Malaysia
Johor, Malaysia

***Abstract*—Credit card based online payments has grown intensely, compelling the financial organisations to implement and continuously improve their fraud detection system. However, credit card fraud dataset is heavily imbalanced and different types of misclassification errors may have different costs and it is essential to control them, to a certain degree, to compromise those errors. Classification techniques are the promising solutions to detect the fraud and non-fraud transactions. Unfortunately, in a certain condition, classification techniques do not perform well when it comes to huge numbers of differences in minority and majority cases. Hence in this study, resampling methods, Random Under Sampling, Random Over Sampling and Synthetic Minority Oversampling Technique, were applied in the credit card dataset to overcome the rare events in the dataset. Then, the three resampled datasets were classified using classification techniques. The performances were measured by their sensitivity, specificity, accuracy, precision, area under curve (AUC) and error rate. The findings disclosed that by resampling the dataset, the models were more practicable, gave better performance and were statistically better.**

Keywords—Credit card; imbalanced dataset; misclassification error; resampling methods; random undersampling; random oversampling; synthetic minority oversampling technique

I. Introduction

In the past decades when businesses were migrated and evolved to the online business and money was managed electronically in an ever-growing cashless banking economy, credit cards were gradually replacing the use of cash over its suitability [1]. Credit cards became the most popular mode of payment ever since. According to [2], credit card-based purchases can be categorised into two types: i) physical card purchase and ii) virtual card purchase. Most payments for online purchases were under virtual card purchases which few information were needed such as card numbers, expiration data, and secure codes. Along with the increasing numbers of the credit card users, the numbers of fraudulent transactions have been constantly increased. In the article [3] stated that, it is hard to find the identity and the location of the fraudsters since the evidences were hidden behind the internet. The merchants that were facing with the credit card fraudsters will bear all the costs including card issuer fees, charges, and administrative charges [4]. Consequently, the merchants must increase the price of the goods or give more discounts or reduce the incentives to conceal all the losses. Hence, an effective fraud detection system is vital to reduce the losses rate.

Before proceeding with the fraud detection system, it must be bear in mind that there had been an enormous increase in the amount of credit card dataset collected and processed by the organisations. Normally in the real dataset, the number of fraudulent is very rare as compared with the non-fraudulent transactions [5, 6, 7]. Conceivably with a skewed dataset, the performance of the system surely drops in terms of its accuracy. When a legitimate transaction is misclassified as a fraudulent transaction, it will affect the customer services and causes to lose trust from the customers and the financial institution [8, 9, 10]. Maes (2002) have provided some capacity that a fraud detection system should have in order to perform a good result [11]. The system should be able to: i) handle skewed distributions, ii) handle noise, iii) avoid the overlapping data iv) adapt themselves to new kinds of frauds, v) evaluate the classifier using good metrics, and vi) detect the behaviour of the frauds. Recent research in [12] stated there are three challenges to construct the fraud detection system. The challenges are: i) the data distribution evolves over time because of seasonality and new attack strategies ii) fraudulent transactions represent only a very small fraction of all the daily transactions and iii) the fraud detection problem is intrinsically a sequential classification task.

In 2017, Haixian et al. stated that it is difficult to detect the rare events due to their infrequency and casualness. Plus, it can result in heavy cost if misclassified the rare events. In their review paper, they have identified three main solutions to the challenges: resampling, cost-sensitive learning, and ensemble methods [13]. The most popular method is resampling methods which are used to rebalance the imbalanced dataset in order to alleviate the effect of the skewed class distribution in the learning process. Secondly, cost-sensitive learning which can be incorporated to both data level and algorithmic level. Lastly, ensemble method is used to improve the performance of a single classifiers that outperform. In a review paper [14], they specified two approaches should be performed to solve the imbalanced data problems: i) solution at data level by balancing the distribution of the majority and minority class trough methods of under sampling, over sampling or combination of both methods, ii) solution at algorithm level by modification in classifier methods or optimise the performance of learning algorithm.

Thus in [13 and 14], both review papers emphasized that there are no absolute methods that are more efficient in dealing with the class imbalance. They found out some insights about commonly-used methods in some domains.

Burez (2009) handled the imbalanced class in churn prediction by applying several methods. The methods are: i) evaluation metrics, ii) cost-sensitive learning, ii) resampling methods and iv) boosting. He used ROC analysis as for evaluation metrics and stochastic gradient boosting learner as for boosting. For cost-sensitive learning, he used random forest and random under sampling as for resampling methods [15]. The study in [16] proposed an efficient resampling method and obtained comparable classification results between random under sampling and random over sampling. The experiments were carried out using four large imbalanced Bioinformatics datasets. They have recommended 100%-under(0.75)-over method for obtaining comparable classification results to the over sampling results. In 2002, [17] has proposed Wrapper-based Random Oversampling (WRO) to handle class imbalanced problem. Wrapper is a pre-processing method that incorporates the classifier output to guide pre-processing. The method oversampled the minority class data randomly and the classifier is optimised. They evaluated the WRO with real dataset that they obtained from UCI repository. WRO has better results in most experiment compared to Synthetic Minority Over Sampling Techniques (SMOTE) and random over sampling. Research in [18] investigated the resampling methods specifically on data from Spotify users. They used the most common oversampling methods: random oversampling and SMOTE, and the most common under sampling method: random under sampling. Yan and Han (2018) proposed RE-sample and Cost-Sensitive Stacked Generalisation (RECSG) based on 2-layer learning models to solve the imbalanced problem in 18 benchmark datasets [19]. The experimental results and statistical tests showed that the RECSG approach improved the classification performance.

In reviewing the literature, resampling methods is the main focus of this study due to its simplicity and compatibility with existing classification models to handle the rarity event in massive credit card dataset. There is no research yet were found on the association between credit card fraud and resampling methods. Therefore, the aim of this study is to investigate the classification models' ability to classify the fraud and non-fraud transactions, and to examine if the different resampling methods could improve the performances of the models. The research methodology of the study is conducted in Section 2. Thereafter in Section 3, the experimental setup is described. Next, the results and discussions is presented in Section 4. This study ends with conclusion remarks and future works in Section 5.

II. Research Methodology

This section gives brief description of the methodology of this study. In addition, this section also discusses each step of the methodology. Fig. 1 displays the framework of research methodology of this study.

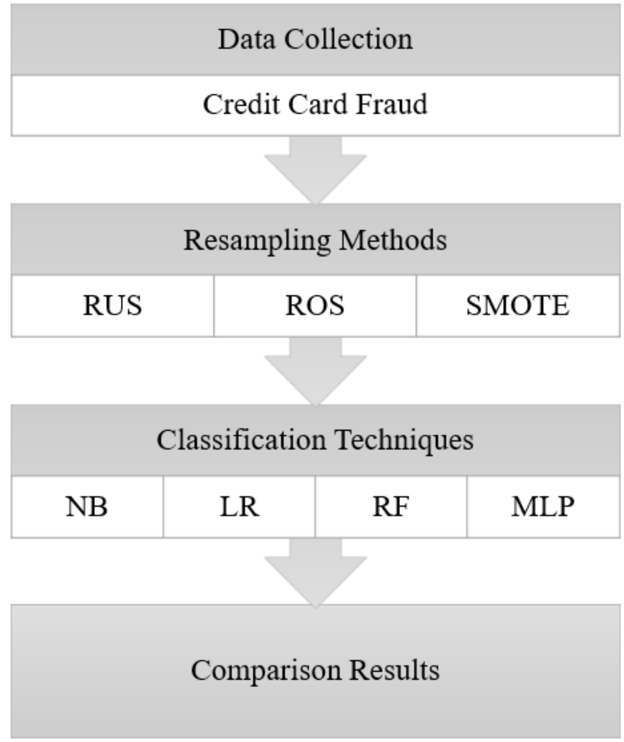


Fig. 1. Framework of Research Methodology.

A. Data Collection

One of the biggest problems associated with researchers in financial fraud detection is lack of real-life data because of sensitivity of data and privacy issue [5]. Hence, a publicly available dataset is downloaded from [20] to be used in this research. It has a total of 284,807 transactions made in September 2013 by European cardholders. The dataset contains 492 fraud transactions, which is highly imbalance.

B. Resampling Methods

Three widely-used methods for resampling in this study are Random Under Sampling (RUS), Random Over Sampling (ROS) and SMOTE. For undersampling, RUS is chosen, since it is considered both simple yet effective. ROS and SMOTE were chosen as oversampling methods because of its widely usage. Furthermore, ROS is an intuitive way of balancing data, whereas SMOTE is more complex creating synthetic samples using K-Nearest Neighbour (KNN). Table 1 below summarises the differences between the three resampling methods.

C. Classification Techniques

Credit Card dataset is a binary classification task. Either the transaction is classified as non-fraud (0) or fraud (1). After the data have been resampled accordingly, the models are needed to be trained using classifiers to evaluate the methods. Thus, in this study, four different classification techniques were explored: Naïve Bayes (NB), Linear Regression (LR), Random Forest (RF) and Multilayer Perceptron (MLP). A summary of the strength and limitations of the classifiers used in this study is given in Table II.

TABLE I. COMPARISON OF RESAMPLING METHODS

	RUS	ROS	SMOTE
Process	Shrink the majority data by randomly discarded the data from the dataset.	Expand the minority data by duplicated the data randomly.	Expand the minority data by extrapolating between preexisting minority instances which obtained by KNN.
Strength	Shorter convergence time.	Not possess any information loss, could produce better results.	Effective in improving the classification accuracy of the minority data.
Limitation	Loss the important information.	Trapped with overfitting due to multiple tied instances.	Data synthetic still possible to spread on both minority and majority data, hence reduced the performance of classification.

TABLE II. STRENGTH AND LIMITATION OF THE CLASSIFICATION TECHNIQUES

Classifier	Strength	Limitation
Naïve Bayes	Simplest classifiers and perform well in classification.	Assumes that all attributes are independent of each other given the context of the class.
Linear Regression	Provide optimal results when the relationship between independent and dependent variable are almost linear.	Sensitive to outliers and limited to numeric values only.
Random Forest	Require low computational power and suitable for real-time operations. The procedure is easy to understand and implement.	Easily to get overfitting when training set does not give underlying domain information. Whenever have new types of cases need retrain.
Multilayer Perceptron	Suitable for binary classification problems.	Retraining is required and need high computational power. Unsuitable for real-time operations.

III. EXPERIMENTAL SETUP

This section describes the division of the data in training dataset and the performance measures conducted throughout of this study. All the resampling techniques are implemented in Java framework of WEKA 3.8 for comparative evaluations. The parameters for the classification techniques were set accordingly by default. No further fine tuning of parameters to specific datasets can be beneficial, consideration of generally accepted settings is more typical in practice.

A. Data Division

The methodological approach taken by this study is motivated from research [15]. The research handled the imbalance problem in churn prediction by resampling the minority and majority classes based on ratio 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20 and 90:10 where churners proportionate with non-churners. Due to the results that the research have obtained and limitation of time, this study chose ratio 30:70 and ratio 50:50 (fraud:non-fraud) to divide the training dataset for this study. An overview of the dataset division, splitting and resampling, can be seen in Fig. 2. Following Fig. 2 is Table III and IV which have more details on dataset division for this research.

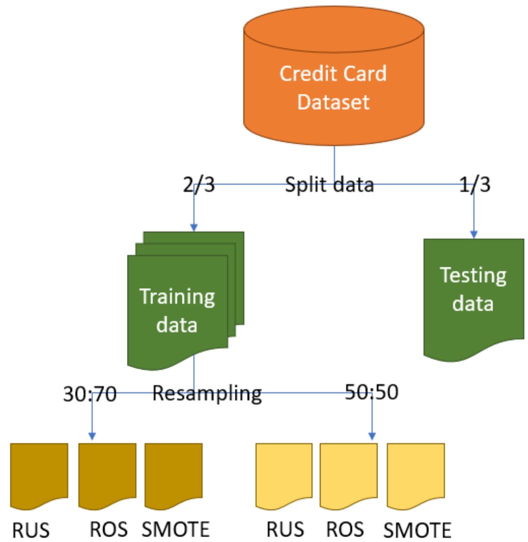


Fig. 2. Division of Dataset.

TABLE III. DIVISION DATASET BY RATIO 30:70

Data Division	Training Dataset	Resampling Method		
		RUS	*ROS*	*SMOTE*
Fraud	328	328	81233	81299
Non-Fraud	189544	768	189544	189543
Total	189872	1097	270777	270842

TABLE IV. DIVISION DATASET BY RATIO 50:50

Data Division	Training Dataset	Resampling Method		
		RUS	*ROS*	*SMOTE*
Fraud	328	328	189544	189543
Non-Fraud	189544	328	189544	189543
Total	189872	656	379088	379086

B. Performances Evaluation

In this study, performance evaluations were conducted to assess the performance of the classification methods for each resampling technique. The models have two fundamental errors may occur: classifying a fraud falsely as a non-fraud and classifying a non-fraud falsely as a fraud. These errors are more commonly known as false positive and false negative results. Other possible classifications will be correctly classified such as true positive and true negative results. The correlation between these are presented in a confusion matrix in Table V. Performance of four classifiers were compared in terms of Sensitivity, Specificity, Accuracy, F-Measure and Area Under Curve (AUC). These metrics are calculated using the confusion matrix as shown below.

TABLE V. CONFUSION MATRIX OF CREDIT CARD DATASET

	Classified as Fraud	Classified as Non-Fraud
Fraud	True Positive (TP)	False Negative (FN)
Non-Fraud	False Positive (FP)	True Negative (TN)

Table V was generated from the four measures: True Positive (TP) — the number of correctly classified as a fraud and it is really a fraud, True Negative (TN) — the number of correctly classified as non-fraud and it is really a non-fraud, False Positive (FP) — instances which were incorrectly classified as a fraud but it is a non-fraud and False Negative (FN) — instances which were incorrectly classified as non-fraud but it is a fraud.

$$Sensitivity = \frac{TP}{TP+FN} \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Accuracy = \frac{TP+TN}{\sum(TP+FP+TN+FN)} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$F\text{-}Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (5)$$

$$AUC = \frac{1}{2} \cdot (Sensitivity + Specificity) \quad (6)$$

$$Error = \frac{FP+FN}{\sum(TP+FP+TN+FN)} \quad (7)$$

IV. RESULTS AND DISCUSSIONS

This section discusses the results that were obtained from the experiments. Table VI and Table VII displays the summary of the comparison results for each classification techniques in three resampling methods by ratio 30:70 and 50:50, correspondingly. The results were compared in terms of sensitivity, specificity, accuracy, precision, F-measure, AUC and time taken to build the model in seconds. All the classifiers were performed well with an accuracy of 0.90 or more. Though, RF dominates with higher accuracy compared to other classification techniques for both ratios of each resampling method.

Table VI provides the information of ratio 30:70 for three resampling techniques. As can be seen in RUS, MLP has higher sensitivity if compared to other classification techniques but have slightly lower specificity than RF. The error rate for MLP and RF are 0.0319 and 0.0273, correspondingly. Thus, RF have approximately 2% of misclassification rate compare to MLP which is have 3% of misclassification rate. For ROS, LR and RF have accuracy 99% which they can correctly identified the fraud and non-fraud of the credit card dataset. However, RF is more precise compared to LR. On the other hand, LR only took 53.5 seconds to build the model while RF took about 343 seconds. The longest time taken to build the model is MLP which is 896 seconds. Meanwhile in SMOTE, RF has higher precision rate compare to other classification methods. It shows that RF often correctly classified non-fraud dataset with 0.9999 rate. Followed by LR (0.9862), MLP (0.9837) and NB (0.9328).

TABLE VI. COMPARISON RESULTS OF CLASSIFICATION TECHNIQUES BY RATIO 30:70

Resampling Methods: Random Under Sampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.85321101	0.97009103	0.93521898	0.92384106	0.88712242	0
Linear Regression	0.89602446	0.9869961	0.95985401	0.9669967	0.93015873	0.08
Random Forest	0.90825688	1	**0.97262774**	1	0.95192308	0.5
Multilayer Perceptron	0.9204893	0.98829649	0.96806569	0.97096774	0.94505495	4.52
Resampling Methods: Random Over Sampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.84758845	0.97382652	0.93595468	0.93279053	0.88815077	2.11
Linear Regression	1	0.99763642	0.9983455	0.99451532	0.99725012	53.5
Random Forest	1	1	**0.99995199**	1	1	343.43
Multilayer Perceptron	0.96569171	0.9989079	0.98894293	0.99736822	0.98127439	896.31
Resampling Methods: Synthetic Minority Over Sampling Technique						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.833294	0.97426441	0.93194925	0.93283212	0.88025831	2.51
Linear Regression	0.98990147	0.9940594	0.99281131	0.98620166	0.9880481	88.16
Random Forest	0.99924968	0.99957266	**0.99947571**	0.99900392	0.99912679	716.2
Multilayer Perceptron	0.98723232	0.99297785	0.9912532	0.98368713	0.98545653	1034.57

TABLE VII. COMPARISON RESULTS OF CLASSIFICATION TECHNIQUES BY RATIO 50:50

Resampling Methods: Random Undersampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.85321101	0.95731707	0.90534351	0.95221843	0.9	0
Linear Regression	0.91131498	0.97865854	0.94503817	0.97704918	0.94303797	0.05
Random Forest	0.92966361	0.97865854	**0.95419847**	0.97749196	0.95297806	0.45
Multilayer Perceptron	0.9204893	0.96036585	0.94045802	0.95859873	0.93915757	2.63
Resampling Methods: Random Oversampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.8497982	0.97318287	0.91149021	0.96940864	0.9056713	2.91
Linear Regression	1	0.99393278	0.9969664	0.99396943	0.99697559	72.17
Random Forest	1	0.99992614	**0.99996307**	0.99992614	0.99996307	451.36
Multilayer Perceptron	0.98128149	0.99772611	0.98950376	0.99768812	0.98941679	1536.34
Resampling Methods: Synthetic Minority Oversampling Technique						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.8335048	0.97364186	0.90357333	0.96934612	0.89630777	4.5
Linear Regression	0.99559467	0.99172747	0.99366107	0.99175934	0.9936733	131.6
Random Forest	0.99996307	0.99952517	**0.99974412**	0.99952538	0.99974418	991.69
Multilayer Perceptron	0.99217592	0.99018692	0.99118142	0.9902064	0.99119018	1499.7

In contrast, Table VII presents the information of ratio 50:50 for each resampling method. In RUS, LR and RF have similar specificity rate which is 0.97866 but different in sensitivity rate nearly at 0.0183. Although both techniques can classify the same number of fraud dataset, RF is still better in classifying the fraud dataset compare to LR. RF also expresses the effectiveness of classification in terms of high F-Measure compared to other classifiers. In the meantime, for ROS, RF has higher accuracy rate. Follows by LR, MLP and NB with 0.9969, 0.9895 and 0.9115, subsequently. Both LR and RF have equal numbers of sensitivity rate which is 1. This means that LR and RF have 100% correctly classified the fraud dataset. Meanwhile, RF gives comprehensive results even though in SMOTE. RF has a small differences of accuracy rate compared to LR and MLP with 0.00608 and 0.00856. Hence, RF can correctly classified fraud and non-fraud in the dataset since it has higher sensitivity rate and precision rate compared to other classification methods. It is important to view that, although NB only took a split of second to build the model, it has the lowest precision rate compared to the rest. NB also has the lowest accuracy rate and sensitivity rate. Albeit MLP has the longest time taken to build the model, MLP is doing well in classifying the fraud and non-fraud which add up to 99% correctness.

Table VIII is quite revealing in several ways. First, unlike the other tables, Table VIII is more focusing on performance of AUC and error rate. Secondly, RUS, ROS and SMOTE were compared with the original training dataset. The highest AUC and lowest error were printed in bold. For each of the case, the test statistics with the highest AUC and lowest errors were calculated and compared with models that were significantly worse. It can be seen in the table that although the original training set should have the lowest rate compared to other resampling methods in four classifiers, it has the worst performance in AUC. It is an example that the model does not have a good statistic.

For NB, the best AUC is obtained by the RUS when the training set is set to 30% of fraud and 70% of non-fraud (AUC=0.9117). It is only has 0.0002 differences compared to ROS when the training sets have the same ratio of fraud and non-fraud. SMOTE and the original training set do not differ significantly what concerns to AUC. While for LR, the closest AUC to 1 is ROS by ratio 30:70 with 0.9988. Followed by the ROS (50:50), SMOTE (50:50) and SMOTE (30:70). RUS in both ratios and the original training set have large differences in AUC compared to ROS with 30% fraud. Although the original training sets have the lowest error set, ROS (30:70) has the second lowest error rate. ROS with 30% of fraud and 70% non-fraud is significantly better in statistic, therefore it can be a better resampling method for LR technique.

TABLE VIII. MEAN AUC PERFORMANCE AND ERROR RATE FOR EACH RESAMPLING METHODS

Techniques: Naïve Bayes							
	Training Data	**RUS**		**ROS**		**SMOTE**	
Ratio		30:70	50:50	30:70	50:50	30:70	50:50
AUC	0.89271531	**0.91165102**	0.90526404	0.91070748	0.91149053	0.9037794	0.90357333
Error	**0.02278916**	0.06478102	0.09465649	**0.06404532**	0.0885098	0.06805075	0.09642667
Techniques: Linear Regression							
	Training Data	**RUS**		**ROS**		**SMOTE**	
Ratio		30:70	50:50	30:70	50:50	30:70	50:50
AUC	0.87036088	0.94151028	0.94498676	**0.99881821**	0.99696639	0.99198044	0.99366107
Error	**0.00057934**	0.04014599	0.05496183	**0.00165450**	0.0030336	0.007188693	0.006338931
Techniques: Random Forest							
	Training Data	**RUS**		**ROS**		**SMOTE**	
Ratio		30:70	50:50	30:70	50:50	30:70	50:50
AUC	0.88259557	0.95412844	0.95416107	**1**	0.99996307	0.99941117	0.99974412
Error	**0.00045821**	0.02737222	0.04580153	0.00004801	**0.00003693**	0.000524291	0.00025588
Techniques: Multilayer Perceptron							
	Training Data	**RUS**		**ROS**		**SMOTE**	
Ratio		30:70	50:50	30:70	50:50	30:70	50:50
AUC	0.88258238	0.95439289	0.94042758	0.9822998	0.9895038	0.99010508	**0.99118142**
Error	**0.00048454**	0.03193431	0.05954198	0.01105707	0.01049624	**0.00874680**	0.00881858

Similar in LR, the better performance of AUC using RF classifier is ROS by ratio 30:70 with 1. ROS (50:50) is following closely with AUC = 0.99996. From the table, ROS (50:50) have the smallest error rate as well after the original training set. Both ratios in ROS show significantly better in statistic for RF. When looking at MLP, what concerns of the performance of AUC, SMOTE gave better result in both ratios. Followed by ROS and RUS. Similar to error rate where SMOTE also gave the smallest rate compared to ROS and RUS. Yet, none of the resampling methods were significantly better in terms of error rate than the original training set.

V. Conclusions and Future Work

This study was set out with the aim to investigate the classification models' ability to classify the fraud and non-fraud transactions, and to examine if the different resampling methods could improve the performances of the models. It is interesting to note that in all four classifiers that have been applied, RF showed a robust performance in three resampling methods. RF succeeded to get higher accuracy compared to NB, LR and MLP for the resampling methods. It would be interesting to compare the classification techniques used in this study with other techniques such as Support Vector Machines, Neural Network and Genetic Algorithm.

Surprisingly, it has been found out that ROS was found to give convincing results if compared to SMOTE. Although SMOTE is quite effective in the literature, this is most probably due to some of the synthetic data resulting from the oversampling process were spreading on both minority and majority data, as discussed in Section 2, which is the main limitation of SMOTE. There were few researchers that have modified SMOTE to create more effective methods in improving the classification performance. Perhaps, this improve-SMOTE can be compared with the current resampling methods for credit card dataset in the future. Hence, these results may provide further support to the organisation to build better fraud detection system which can handle the skewed distribution and noise as well to evaluate the classifier using better metrics.

Acknowledgment

The authors honorably appreciate the Ministry of Education Malaysia, Universiti Teknologi Malaysia (UTM), UTM Big Data Centre and MyBrain15 for their support. This work is funded by Fundamental Research Grant Scheme (FRGS) with the title "Enhancing Data Analytics Algorithms using Deep Learning Approaches in Predicting Big Data Cyber-Enabled Crimes" (4F877).

References

[1] Ki, Y., & Yoon, J. W. (2018, January). PD-FDS: Purchase Density based Online Credit Card Fraud Detection System. In KDD 2017 Workshop on Anomaly Detection in Finance (pp. 76-84).

[2] Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. International Journal of System Assurance Engineering and Management, 8(2), 937-953.

[3] Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. IEEE ACCESS, 6, 14277-14284.

[4] Quah, J. T., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert systems with applications, 35(4), 1721-1732.

[5] Reddy, M. S., Indraja, S., & Nikhil, L. (2017). Implementation Of Neural Network For Cashless Transactions In Credit Card Transactions. International Journal of Recent Trends in Engineering & Research.

[6] Dal Pozzolo, A., Caelen, O., & Bontempi, G. (2015, September). When is undersampling effective in unbalanced classification tasks?. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 200-215). Springer, Cham.

[7] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In Computational Intelligence, 2015 IEEE Symposium Series on (pp. 159-166). IEEE.

[8] Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. IEEE Intelligent Systems and Their Applications, 14(6), 67-74.

[9] Chan, P. K., & Stolfo, S. J. (1998, August). Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In KDD (Vol. 98, pp. 164-168).

[10] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

[11] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002, January). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).

[12] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), 3784-3797.

[13] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220-239.

[14] Santoso, B., Wijayanto, H., Notodiputro, K. A., & Sartono, B. (2017, March). Synthetic Over Sampling Methods for Handling Class Imbalanced Problems: A Review. In IOP Conference Series: Earth and Environmental Science (Vol. 58, No. 1, p. 012031). IOP Publishing.

[15] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

[16] Batuwita, R., & Palade, V. (2010, July). Efficient resampling methods for training support vector machines with imbalanced datasets. In Neural Networks (IJCNN), The 2010 International Joint Conference on (pp. 1-8). IEEE.

[17] Ghazikhani, A., Yazdi, H. S., & Monsefi, R. (2012, May). Class imbalance handling using wrapper-based random oversampling. In Electrical Engineering (ICEE), 2012 20th Iranian Conference on (pp. 611-616). IEEE.

[18] Jagelid, M., & Movin, M. (2017). A Comparison of Resampling Techniques to Handle the Class Imbalance Problem in Machine Learning: Conversion prediction of Spotify Users-A Case Study.

[19] Yan, J., & Han, S. (2018). Classifying Imbalanced Data Sets by a Novel RE-Sample and Cost-Sensitive Stacked Generalization Method. Mathematical Problems in Engineering, 2018.

[20] Credit Card Fraud Detection. Accessed: June. 7, 2018. [Online]. Available: https://www.kaggle.com/dalpozz/creditcardfraud

Exploring Saudi Citizens' Acceptance of Mobile Government Service

Dr. Adnan Mustafa AlBar, Mashael A. Hddas
Department of Information Systems
Faculty of Computing & Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Mobile government is considered as an emerging technology that has been used in Saudi Arabia in order to enhance communication between the government and its citizens, it can also be considered as a mechanism through which the government can effectively respond to their needs and expectations. The current study seeks to propose and validate M-government adoption model to fully understand the varied variables affecting the adoption behavior. This model is based on Technology Acceptance Model (TAM) and DeLone and McLean Information Systems Success Model. The researchers will depend on a descriptive survey approach using structured questionnaires to investigate the extent of M-government acceptance by Saudi users. Structural equation modeling will be used as a method for statistical data analysis.**

Keywords*—*M-government; adoption; acceptance; citizen

I. Introduction

There is no doubt that the developed technology has permeated all walks of life, which has helped in doing things more easily than ever before, this in fact has had a great impact on many aspects of our lives such as health, education, transportation and communication.

The development in the field of technology has led to the introduction of many advanced tools that have a lot of useful applications for better communication. Nowadays, smart devices, iPhone and tablets have become important mechanisms for communication [1]. The wide spread of mobile devices around the world and the technological revolution surrounding their different applications have opened the floodgates for the introduction of M-government [2-3].

II. Research Motivation

Despite the fact that Kingdom of Saudi Arabia has made significant strides in the field of technological infrastructure development, M-government technology is still in its infancy, this confirms the idea that the Saudi government should pay more attention to integrate M-government services in its public sector in order to achieve its national objectives and reduce constraints imposed on citizens as a result of using traditional government services [4-5]. My research is motivated by the urgent desire to investigate the different factors that may affect the acceptance of M-government in Saudi Arabia through combining the dimensions of DeLone and McLean success model represented in (system quality, information quality and service quality) with the dimensions of TAM model represented in (Perceived Usefulness and Perceived Ease of Use) to determine their effects on attitudes, behavior intentions and actual use of the system. . Worthy here to mention that using the two models will enable the researchers to make full use of the strength areas of these models and as a result draw a full picture of the factors that affect M-government acceptance behavior among Saudi citizens. Identifying these factors can be considered as a must in order to move from the preliminary phase of experimentation to the actual use of the system for maximizing the benefits that can be gained from applying M-government in Saudi Arabia. Finally, the current study will provide some suggestions to help policy makers in improving the system of M-government in Saudi Arabia.

III. Similar Work

In a study [6] that has been carried out to investigate factors relevant to diffusion and adoption of M- Government in Greece, findings revealed that ease of use and compatibility could affect the adoption behavior. Another study [7] has been conducted to determine issues surrounding M-government services adoption in Jordon, results showed that public awareness, infrastructural limitations and lack of comprehensive legal framework were among the critical issues surrounding the use of M-government. At the local level a study has been conducted [8] to determine variables affecting M-government adoption in Saudi Arabia, findings revealed that satisfaction, mobility and trust were predictors of adoption behavior among citizens. Another study [9] analyzed the challenges and opportunities associated with the implementation of m-government services in Saudi Arabia. The results suggested that although a large percentage of population did not have access to mobile technologies, there was still a strong desire among users for the provision of M-government services. The previous studies and researches mentioned above have revealed that M-government technology is affected by varied factors. In order to identify these factors; the researchers will make full use of the three major dimensions of the DeLone and McLean Success Model (system quality, information quality and service quality) and two dimensions of the TAM Model (represented in perceived usefulness and perceived ease of use) to determine their effects on attitudes, behaviour intentions and use of M-government system in Saudi Arabia. This mixing between the two models in fact will be of great significance as it enables the researchers to combine the most common factors that can affect acceptance of M-government in Saudi Arabia.

IV. E-Government and M-Government in Saudi Arabia

All e. government initiatives issued by the Saudi government can be considered as a part of its National Information Technology plan in an attempt to restructure different processes taken place in the public sector [10]. Mobile government technology is still in the initial phase, this confirms the idea that the Saudi government should pay more attention to integrate M-government services in its public sector to generalize this experience all through the kingdom [11].

The study of M-government adoption behavior among Saudi citizens is of great importance as it helps us in identifying the different opportunities and challenges face the industry of M-government and as a result it will help in improving this service through getting a closer look on the different behaviors of Saudi citizens especially while taking into consideration the different challenges associated with M-government adoption such as security, privacy, usability and so on. In addition, the most important challenge that hinders the effective implementation of M-government lies in the adoption behavior of citizens, some researchers agreed that the Saudi government needs to overcome different inequalities that impede the effective adoption of M-government services [12].

V. Theory Background

The advancement in wireless communication has paved the way to the emergence of M-government. M-government can be seen as an effective and efficient solution for delivering services using mobile technology [2-13]. The goals and objectives of M-government are diverse including –but not limited to- saving time, cost and effort through providing efficient services to citizens using electronic channels [14]. It also aims at enhancing civic engagement among citizens [15] Many countries seek to implement M-government initiatives. The point of departure here is that the implementation of M-government heavily depends on many variables among them is the adoption behavior of citizens [16-17].

VI. Questionnaire

The researchers designed a structured questionnaire that would be used to reflect users' views regarding the critical factors that might affect the behavior intention and attitudes towards M-government, the dimensions of the questionnaire will be extracted from a model designed by the researchers through making full use of dimensions of DeLone and McLean Information Systems Success model and TAM model. The questionnaire consisted of three main parts: The first part was dedicated to gather demographic information related to gender, age, educational level, employment status, ICT usage experience and the actual use of M-government services. The second part included eight main dimensions represented in system quality, information quality, service quality, perceived usefulness, perceived ease of use, attitudes, actual use and behaviour intention. The final part included open questions to identify the different strategies that could be used or added to improve the quality of services being provided. The current questionnaire has Arabic and English versions. It depended on a 5-point Likert scale ranging from strongly disagree to strongly agree.

VII. Data Collection

The researchers will depend on the descriptive survey approach as the main study approach. Descriptive survey approach is concerned with the characteristics of whole sample of the study in an attempt to provide effective solutions to the problem. The survey approach makes full use of scientific methods by examining the different materials, analyzing data and reaching some results that can be generalized [18]. Online Surveys will be used to gather quantitative data for the proposed theoretical model in an attempt to reach as many people as possible. Surveys will be emailed to public and private organizations requesting them to forward these questionnaires to their employees. It will also be sent to universities to ask students and academic staff to participate in the survey, this process succeeded in gathering the responses from 1218 participants in the study.

VIII. Data Analysis

Microsoft Excel was used to code the data. In this step, we identified the variables of the study, the columns they occupied in the data file, their possible values. The five-point Likert scale was coded with values ranging from (Strongly Disagree =1, Disagree= 2, Neutral=3, Agree= 4, and Strongly Agree= 5). In order to test hypotheses and analyze data required for the current study, the researchers used Structural Equation Modeling (SEM) technique as the main instrument used in the analysis. It is a powerful multivariate analysis used to test the validity of hypothesized theoretical models. It can also be seen as a simple method to show numerous relationships simultaneously and evaluate relationships [19]. The analysis will be done using SmartPLS software. PLS is a SEM technique that simultaneously represents the theoretical relationship among latent variables and the relationships between latent variables and their indicators. The software package SmartPLS version 2 for SEM analysis and SPSS version 21 for descriptive analysis of demographic information would be used. The data set would be evaluated for missing values, invalid observations or outliers. When missing data for an observation exceeded 15%, it will be removed from the data set. When less than 5% of values per indicator would be missed, mean value replacement would be used. Otherwise, case-wise deletion would be used. After identifying outliers, they would be removed from the data if they represented less than 5% [20]. The path coefficients and the coefficients significance were used for hypotheses testing. Hypotheses testing in SmartPLS could be involved in the evaluation of t-values. All p significance levels ($p<0.01$, $p<0.05$ and $p<0.10$) were addressed. The larger p value of less than 0.10 could be regarded as appropriate for exploratory research [20].

IX. Discussion of Results

The descriptive statistics of the study showed that (52.6 %) of the study sample were males while (47.4%) were females. The majority of the participants' ages ranged from 20 to 29 years old, they had good knowledge of ICT, most of them used ICT applications for less than one hour a day and they had been using mobile phones for more than five years. The majority of them used mobile phones in order to make calls and send messages and most of them use M-government services for many different reasons, but the most prominent one was to obtain applications from the Ministry of Education. The results of the study showed that System Quality and Information Quality had not statistical significant effects on Perceived Usefulness. This result went in harmony with the results mentioned by [21] who showed that neither system quality nor information quality had positive effects on perceived usefulness. On the contrary, the results confirmed that perceived security had a positive effect on perceived usefulness. Worthy here mentioning that these results were disappointing, especially while keeping in mind that the researchers had expected that system quality and information quality could play critical roles in affecting perceived usefulness as our beliefs regarding the feasibility of a system could be formed in the light of the quality of the system as a whole and the quality of the information being provided by it. Information quality, service quality and perceived usefulness had positive effects on the perceived ease of use. Also, Service quality had a positive effect on perceived usefulness. This result went in harmony with the result of [22] who indicated that services quality (SQ) was found to have the strongest influence on adoption of M-government. Honestly, the researchers expected these results, especially while taking into account that information quality, service quality and perceived usefulness have been shown in the literature to be powerful factors that have the ability to influence M-government adoption behaviour. The results of the study showed that perceived ease of use and perceived usefulness had positive effects on attitudes towards M-government; the results were consistent with [23] who showed that perceived ease of use, near-term usefulness and long-term usefulness had significant and positive influences on the intention to use mobile government. These results were not consistent with [24] who showed that perceived usefulness had a positive effect on perceived value. Attitudes had positively affected behavior intention. This result went in harmony with [24] who showed that individual's intention to behave was based on the person's attitude. Behavior intention had positively affected actual use of M- government. This result went in harmony with the theory of reasoned action [25] indicating that one's intentions affected his actual behavior. Here the behaviour component can be seen as a driving force towards using M-government services, and as a result, once behaviour intention was motivated, actual use could be guaranteed. Significant testing of results using structural model path coefficients can be shown as in the following table:

Based on their sizes, the results showed that SeQ had the strongest effect on PU (0.5479), followed by PEOU (0.1776), while SQ, and IQ almost had no effect. Also, SeQ had a strong effect on PEOU (0.4048), followed by IQ (0.2635) and SQ (0.1406). Findings showed that both PU (0.5497) and PEOU (0.3041) had strong effects on Att. Also, these findings showed that Att had strong effect on BI (0.6086), and BI had a strong effect on AU (0.5799). The Path Coefficients can be shown as in the following table:

TABLE I. SUMMARY OF SIGNIFICANT TESTING OF RESULTS

	Path Coefficient (Bootstrap)	tt-values	PP- value	Significance Levels
SQ →→ PU	0.0485	1.22	0.22	*NS*
SQ →→ PEOU	0.1408	3.66	<0.001	***
IQ →→ PU	-0.0703	1.61	0.11	*NS*
IQ →→ PEOU	0.2638	6.06	<0.001	***
SeQ→→ PU	0.5493	11.89	<0.001	***
SeQ→→ PEOU	0.4061	9.72	<0.001	***
PEOU →→ PU	0.1773	4.20	<0.001	***
PEOU →→ Att	0.3053	9.66	<0.001	***
PU →→ Att	0.5497	21.06	<0.001	***
Att→→ BI	0.6105	25.76	<0.001	***
BI →→ AU	0.5819	22.73	<0.001	***

*p <0.10. **p <0.05. ***p <0.01.

TABLE II. PATH COEFFICIENTS

	PU	**PEOU**	**Att**	**BI**	**AU**
SQ	0.0479	0.1406			
IQ	-0.0695	0.2635			
SeQ	0.5479	0.4048			
PU			0.5497		
PEOU	0.1776		0.3041		
Att				0.6086	
BI					0.5799

X. MOST PROMINENT RESULTS AND RECOMMENDATIONS

The most important results can be shown as follows: system quality (SQ) has not a positive influence on the Perceived Usefulness (PU) of using M-government, System Quality (SQ) has a positive influence on the Perceived Ease of Use (PEOU) of using M-government, Information Quality (IQ) has not a positive influence on the perceived usefulness of using M-government, Information Quality (IQ) has a positive influence on the Perceived Ease of Use (PEU) of using M-government, Service Quality (SeQ) has a positive influence on the Perceived Usefulness (PU) of using M-government, Service Quality (SeQ) has a positive influence on the Perceived Ease of Use (PEU) of using M-government, Perceived Usefulness (PU) has a positive influence on Attitude (Att), Perceived Ease of Use (PEU) has a positive influence on Attitude (Att), Attitude (ATT) has a positive influence on Behaviour Intention (BI) to use M-government services, Behavior Intention (BI) had a positive influence on Actual Use (AU) of M-government. The researchers recommend improving the provision of services being provided by M-government as a mean of improving citizens' attitudes, intention behaviour and actual use of M-government, this can be done by making the governmental website more attractive, easing the browsing process, providing guidelines related to how to deal with the website, building trust in M-government transactions, providing internet connections especially to those in remote areas. The researchers also recommend developing the technological infrastructure within the Kingdom in order to provide citizens with easy access electronic services. Improving the rates of use of M-government will help in reducing the burdens imposed on many governmental agencies.

For future studies, the researchers recommend conducting similar studies but in other regions in Saudi Arabia using different methodological approaches, for example conducting studies using comparative approaches in order to compare levels of acceptance of M-Government Service between Asir and Jeddah to determine levels of acceptance and adoption behaviors and different factors affecting them.

ACKNOWLEDGMENT

I'd like to thank all researchers in the Department of Information Systems at Faculty of Computing & Information Technology for their help, I'd like also to thank those who are working at King Abdulaziz University, Saudi Arabia for easing my mission as a researcher especially when conducting the field study.

REFERENCES

[1] H. Abdelghaffar and Y. Magdy, “The Adoption of Mobile Government Services in Developing Countries: The Case of Egypt,” International Journal of Information and Communication Technology Research, vol. 2, pp. 333-341, 2012.

[2] E. Goyal and S. Purhoit, “Emergence of M. Government the way forward,” SIES Journal of Management, vol. 8, pp. 56-65, 2012.

[3] V. R. Rao, “Improving Government to Employee (G2E) Services through Mobile Technology – A Survey,” International Journal of Computer Applications, vol. 168, pp. 33-45, 2017.

[4] T. Ahmad, A. A. Ansari, A. Akhtar, and S. Parveen, “Current Review Of ICT And M-Government Services In Saudi Arabia,” International Journal of Computer Engineering and Applications, vol. VII, pp. 71-77, 2014.

[5] M. Alsenaidy and T. Ahmad, “A review of current state m government in Saudi Arabia,” Global engineers & Technologists Review, vol. 2, pp. 5-8., 2012.

[6] A. Vrechopoulos and M. Batikas, “Predicting the Adoption of Mobile Government Services”. In"Mobile and Ubiquitous Commerce: Advanced E-Business Methods", Vol. 4, M. Head,. Eds. IGI Global, 2009, pp. 63-78.

[7] O. Al-Hujran, “Toward The Utilization Of M-Government Services In Developing Countries: A Qualitative Investigation,” International Journal of Business and Social Science, vol. 3, pp. 155-160, 2012.

[8] S. Alotaibi and D. R., “A Conceptual Model for Examining Mobile Government Adoption in Saudi Arabia,” In: Proceedings of the European Conference on e-Government, ECEG. Academic Conferences Limited, pp. 369-375, 2015.

[9] A. Alssbaiheen, “M. Government Adoption in Saudi Arabia: Challenges and Opportunities,” International Journal of Technology and Human Interaction (IJTHI), vol. 11, 2015, DOI: 10.4018/ijthi.2015070104, Available online at http://www.irma-international.org/article/m-government-adoption-in-saudi-arabia/128403/

[10] N. Abanumy and P. J. Mayhew, “M- government Implications For E-Government In Developing Countries: The Case of Saudi Arabia,” Presented at The First European Mobile Government Conference, University of Sussex, Brighton, UK, 2005.

[11] C- M. Park and J- H. Bae, “Sources of Public Trust in Government: East Asian Evidence,” Prepared for delivery at the IIAS Study Group on Trust and Public Attitudes Workshop on “Trust in Public Administration” in Seoul, Korea on December 11-12, 2012.

[12] A. Alssbaiheen and S. Love, “The Opportunities and Challenges Associated With M-Government as an E-Government Platform in KSA: A Literature Review,” IJMBS, vol. 5, pp. 31-38, 2015.

[13] H. S. Al-Hubaishi, S. Z. Ahmad, and M. Hussain, “Exploring mobile government from the service quality perspective,” Journal of Enterprise Information Management, vol. 30, pp. 4-16, 2017.

[14] M. Abaza and F. Saif, “The adoption of mobile government services in developing countries,” *IJCSI International Journal of Computer Science Issues*, vol. 2, pp. 137-145, 2015.

[15] A. S. Albesher, “Civic engagement with traditional government and mobile government: Comparing rural to urban citizens in Saudi Arabia,” Doctor Thesis, Iowa State University, Ames, Iowa, 2016.

[16] Y. K. Dwivedi, M. A. Shareef, and N. Archer. “Examining Adoption Behavior Of Mobile Government,” The Journal of Computer Information Systems, vol. 53, pp. 39-49, 2012.

[17] S. Rees, “Designing User - Centric Mobile Government: A Checklist for Efficiency and Effectiveness,” European Conference on e-Government, vol. 2016, pp. 170-176, 2016.

[18] N. SALARIA, “Meaning Of The Term- Descriptive Survey Research Method,” *International Journal of Transformations in Business Management*, vol. 1, 2012, pp.1-7.

[19] J. C. Anderson and D. W. Gerbing, "Structural equation modeling in practice: A review and recommended two-step approach,” Psychological Bulletin, vol. 103, p. 411, 1988.

[20] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, A primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). Thousand Oaks, CA: SAGE, 2014.

[21] C. Wang, R. Fang, K. Park, Y. Feng, Z. Lu, and Y. Cui, “Perceived Usefulness, Perceived Security and Adoption of Mobile Government: An Empirical Research,” Advances in information Sciences and Service Sciences (AISS), vol. 4, pp. 234-244, 2012.

[22] N. Sahari, N. A. M. Zain, and A. Althunibat, “Modeling the factors that influence mobile government services acceptance,” African Journal of Business Management, Vol. 5, pp. 13030-13043, 2011.

[23] Y. Liu, H. Li, V. Kostakos, J. Goncalves, S. Hosio, and F. Hu, “An empirical investigation of mobile government adoption in rural China: A case study in Zhejiang province,” Government Information Quarterly, Vol. 31, pp. 432–442, 2014.

[24] T. R. Huggins, C. M. Yates, R. J. Cook., R. B. Tranter, J. R. Park, and P. T. Dorward, “Theory Of Reasoned Action And Its Integration With Economic Modeling In Linking Farmers’ Attitudes And Adoption Behavior – An Illustration From The Analysis Of The Uptake Of Livestock Technologies In The South West Of England,” International Farm Management Association, 14th Congress, Perth, Western Australia, August 10-15, 2003.

[25] W. Jen, T. Lu and P. T. Liu, “An Integrated Analysis of Technology Acceptance Behavior Models: Comparison of Three Major Models,” MIS Review, vol. 15, pp. 89-121, 2009.

The Role of user Involvement in the Success of Project Scope Management:

Jordanian Government IT Departments

Maha Alkhaffaf[1]
Management Information System Department, Jordan
The World Islamic Sciences and Education University
Amman-Jordan

Abstract—**Greater emphasis is now being placed on User Involvement as a factor imperative to Success in Project Scope Management. Although Project Scope Management Processes have a tendency to centre on various factors pertaining to the collecting criteria, defining scope and verifying scope, controlling scope is viewed as being fundamental to the management process as a whole. Furthermore, Success in Project Scope Management in the modern-day competitive business setting is recognised as resting on efficient and effective processes applied across Project Scope Management. One essential factor in achieving success in this arena is that of User Involvement. In this regard, the point is presented that Project Scope Management and User Involvement may be implemented in such a way so as to enhance Successful Project Scope Management. A questionnaire-centred survey approach utilizing Project Scope Management Processes and User Involvement to Successful Project Scope Management, encompassing management- and strategy-level employees, totalling 295, was applied in order to establish the link, both indirect and direct, between particular elements influencing four different IT departments at the governmental level. The data gathered underwent analysis through the use of SPLS (Smart Partial Least Square). This work provides a valuable contribution for professionals in the field, both in terms of researchers and practitioners, and further highlights the different ways in which project managers can arrange and modify Project Scope Management Processes in pursuit of their efforts to enhance the mediation of Successful Project Scope Management through User Involvement.**

***Keywords*—Gathering requirements; defining scope; verifying scope; controlling scope; and user involvement**

I. Introduction

A large number of businesses recognise the fact that a significant degree of their effectiveness depends on how Project Scope Management is applied and managed. Accordingly, there is a need for organisations to establish the Project Scope Management Processes and subsequently outline and identify the role adopted by User Involvement when it comes to attaining success in Project Scope Management. Furthermore, it is recognised that there are various difficulties in Project Scope Management [1] [2] [3], with the business viewed as fluid and changeable. The field involves a great variation of users in different projects in mind of achieving varying goals across differing settings. With this taken into account, as highlighted by [4], project management may be positioned in such a way so as to facilitate businesses in strategically structuring themselves to attain their business objectives and subsequently enhance their competitiveness across their industry.

Furthermore, a number of different researchers have presented the recommendation that Project Management (PM) seeks to redirect away from more conventional approaches to more generalised management principles [19], specifically when there are complex environments as the setting for projects [5]. Moreover, in the study of Ajelabi & Tang [2010], it was recognised that, with the passing of time and the greater wealth of experience and literature, [7] PM theory has provided a valuable instrument when it comes to change implementation across businesses. In addition, the work of Kwak & Anbari [2009] highlights the need for PM theory practitioners to encourage the adoption of PM theory as an academic discipline. Owing to the widely acknowledged value of the field, Project Scope Management has become recognised as an imperative consideration across different sectors in Jordan; therefore, in-depth and wide-ranging expertise in this area has become recognised as necessary.

Project Scope management processes across businesses and new project managers [9] undergoing training lead the overall process of the project, and are viewed as fundamental organisational change in line with project implementation success. In addition, a number of different project organisations, such as Oracle, SAP and Microsoft, amongst others, place much emphasis on the best, most innovative practices, such as those carried out in significant businesses, [10] i.e. IT departments in governmental institutions, which have in place environmental professionals employed in order to garner insight into the required experiences [11].

One problem with this particular solution is that Project Scope Management does not always encompass User Involvement; specifically, they provide management support, as well as support to the project user, with User Involvement not always incorporated within the team. In such instances, there is a need for the project management to be clear on the issues pertaining to User Involvement so as to ensure the necessary support is provided. Project Scope Management processes need to be assigned in such a way so as to include the most important and valid data, with this updated and related to the greatest possible degree. It is not always feasible for this to be achieved owing to the fact that varying degrees of precision are required in different areas. In order to ensure the data is kept informed and relevant, it is important that there are updates and feedback whenever necessary [6]. The issue is to determine the way in which Successful Project Scope Management can be achieved, whilst also ensuring a significant usage level and understanding. The most optimal situation is that all of the necessary data exists across the Scope Management processes, and that user information is kept updated and valid, with continuous development through User Involvement across all stages. When it comes to dealing with this particular issue, the aim was suggesting a conceptual framework relating to the Project Scope Management Processes, bringing together Successful Project Scope Management and User Involvement. Accordingly, this particular work completes an analysis on the link between Project Scope Management Processes, User Involvement, and the effects of such in line with Project Scope Management success.

This study is broken down into six key sections in an effort to describe Project Scope Management. First and foremost, there is an introduction into the most pertinent of considerations and the value acknowledged in Project Scope Management Processes and User Involvement. Secondly, a review of the relevant literature relating to Project Scope Management Processes is presented, along with the value of User Involvement in line with Project Scope Management success. Third, there is the presentation of the hypotheses and conceptual framework. Following is an explanation of the research methodology, with the fifth section providing the data analysis findings, whilst the sixth section draws its conclusions.

II. Litratur Review

A. *Project Scope Management Processes*

The country of Jordan is recognised as being in its developmental stages, with decentralisation presenting a number of challenges in administration and local governments, in combination with work processes undergoing globalisation and there being much significant development in ICT trends—all of which are recognised as having a key and significant effect on the organisational capacity of Jordan, Furthermore, businesses are called upon to implement plans and present innovative ideas[12]. There is much emphasis being placed on change, which is encouraging firms to establish their systems and projects [13], [14]. Accordingly, a significant wealth of knowledge in the literature in the area of Project Management, IT projects and Project Scope Management is now available [15]; [16]; [17]; [18]. Furthermore, various researchers have completed analyses on the effect of Project Management on project success. As an example, the study of Nikumbh & Pimplikar [2014] describes PM as being a skill identifiable as a human and material resource centred on leading and organising throughout a project's lifecycle, notably through modern management method development in such a way so as to attain the outlined aims of scope, cost, participation satisfaction, quality and time.

Furthermore, the Project Management Institute (PMI) recognises the key skills needing to be offered by an efficient and valuable project manager [20]. First and foremost, the key competencies are recognised as scope management, with scholars Sánchez & Schneider [2014] describing international project management organisations as having created their own project management guidelines upon knowledge areas [21], with the inclusion of scope management. Moreover, it has been stated by Marinho et al. [2014] that the majority of projects have come to experience restrictions in regards costs, scope and time, in addition to particular principles relating to quality [22].

Nonetheless, in an effort to teach management and businesses the key role of User Involvement in line with Successful Project Scope Management, there is a need to define the success of Project Scope through completing an evaluation on the approval of the user. As such, one fundamental aspect of Successful Project Scope Management is that of User Involvement. Moreover, as highlighted by PMI [2013] Project Scope Management Processes may be broken down into four different process groups, namely Collecting Requirements, Defining Scope, Verifying Scope, and Controlling Scope.

B. *The Value Recognised in user Involvement in Line with Project Scope Management Success*

During more recent times, User Involvement has become acknowledged as a resource encouraging and facilitating success in Project Scope Management across a number of business organisations. Various authors, including [10] Travaglini et al. [2014], recognise that stakeholder executive is one of the most important project success aspects owing to the fact that success in projects is significantly dependent on stakeholder satisfaction. Furthermore, project management experts are highlighted by [23] Seresht et al. [2014] as continuing to show a lack of consensus in relation to how project success may be defined and measured. The work of [2] Morris [2010] further emphasises that PM is becoming more and more widely used and in such a way so as to include the user across all arenas. It is important to recognise that future projects need to place greater emphasis on user-specific deliverables as quickly as possible.

Nonetheless, in the view of [7] Mian et al. [2010], a project is recognised as involving various individuals all working in unison on a common task, sharing the tasks, resources and responsibilities so as to achieve success. It has been stated by [3] Too & Weaver [2014] that actual PM encompasses a number of different objectives, in addition to an agreement between the project manager and user on how such objectives will be fulfilled. Furthermore, the point has been made by [24] Nenni et al. [2014] that a number of different professionals and academics in the field have examined the way in which processes and approaches can be improved in an effort to achieve efficiency improvements in attaining the project goals of a firm. Moreover, it is noted in the work of [15] Al Freidi [2014] that professional project management tools may be utilised so as to document and monitor the progress of a project, [8] which subsequently can lead to success. There is strict adherence to project planning and monitoring, as well as communication between the user and project manager, all as part of the management infrastructure applied throughout the lifecycle of the project.

A research by [25] Mir & Pinnington [2014] details that the framework presents a number of different factors underpinning project success, with the inclusion of business success, customers, efficiency, future preparation, and the influence of achieving a competitive edge.

In addition, the study carried out by [26] Purna [2012] highlighted communication management between the various parties in a project as being well-detailed in the literature, predominantly owing to the emphasis placed on this part of PM and its recognised influence in project success. As such, User Involvement across all of the Project Scope Management processes results in a greater degree of success in Project Scope Management. Furthermore, preliminary scope statements are identified by [27] Silvius & Schipper [2014] as highlighting the needs and expectations of stakeholders through user involvement across all a project's aspects.

III. Research Model and Hypotheses Development

In the recent past, Project Scope Management and User Involvement process have been identified as fundamental when striving to achieve Project Scope Management success. Accordingly, the research available in the field of Project Scope Management Processes and User Involvement in line with Project Scope Management success is examined in this work. In line with this, the researcher presents a framework with the aim of emphasising the casual links between the various Project Scope Management Processes (Collecting requirements, Defining scope, Verifying scope, controlling scope [20] and User Involvement with the aim of achieving improvement across Project Scope Management success.

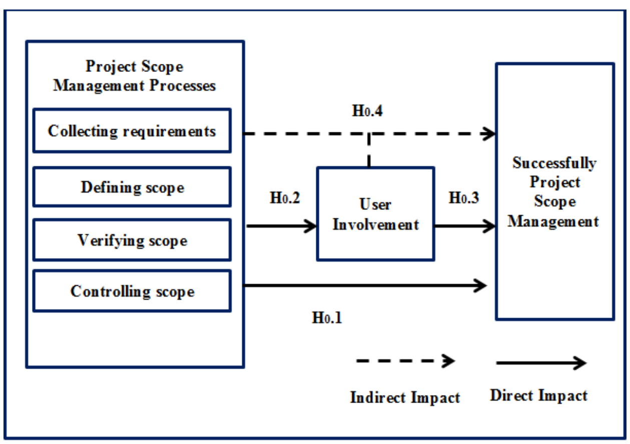


Fig. 1. Proposed Model.

In line with that stated above, Project Scope Management processes [20] implementation is recognised as appropriate in this work for various reasons: firstly, the technique of Project Scope Management facilitates project managers and supervisors in assigning the most appropriate volume of work to achieve successful project completion; secondly, the majority of project methodologies warrant that the scope of a project is outlined in the first instance; and lastly, the PMBOK Guide is concerned with providing a further breakdown of project management knowledge that, as a whole, is acknowledged as being good practice for Project Scope Management processes [28]. Throughout this particular work, User Involvement adopts a mediatory role in the link between Project Scope Management Processes and achieving success in Project Scope Management. The diagram below "Fig. 1" provides an overview of this work's model.

The above "fig 1" encompasses a total of 6 different factors, 4 of which are linked with Project Scope Management Processes whilst the remainder are linked with User Involvement and Successful Project Scope Management. The suggested model is recognised as comprising a number of different processes, as detailed as following:

A. Collecting Requirements:

This comprises explaining and detailing the functionality and overall structure of the products generated across the preliminary organisation of the project.

B. Defining Scope:

This phase relates to the review of all project charter elements, with the inclusion of the preparation of the necessary documents and the assets of the organisational processes applied in order to create a scope statement.

C. Verifying Scope:

This approach encompasses the formalisation of approval regarding the project deliverables.

D. Controlling Scope:

Relating to project scope changes, and the control of such, throughout the project's lifecycle.

E. User Involvement:

Centred on the process of assigning all users recognised as being of influence to the project team, with assignment to the project team, whilst also outlining their responsibility in Collecting requirements, Defining scope and Verifying scope.

F. Successful Project Scope Management:

Considers Project Scope Management, with the inclusion of those processes validating the project whilst addressing all of the components fundamental when seeking to achieve success in Project Scope Management.

At the first stage, the direct effects acknowledged as relevant in Project Scope Management Processes and Successful Project Scope Management will be measured in H.1, which includes a total of four different sub-hypotheses, detailed as follows:

H.1.1: Collecting requirements is recognised as having a significant effect on Successful Project Scope Management at ($\alpha \leq 0.05$).

H.1.2: Defining scope is recognised as having a significant effect on Successful Project Scope Management at ($\alpha \leq 0.05$).

H.1.3: Verifying scope is recognised as having a significant effect on Successful Project Scope Management at ($\alpha \leq 0.05$).

H.1.4: Controlling scope is recognised as having a significant effect on Successful Project Scope Management at ($\alpha \leq 0.05$)

Throughout this work, the direct effects acknowledged as relevant in regards Project Scope Management Processes and User Involvement will be measured in H.2, which includes a total of four different sub-hypotheses, detailed as follows:

H.2.1: Collecting requirements is recognised as having a significant effect on User Involvement at (α≤ 0.05).

H.2.2: Defining scope is recognised as having a significant effect on User Involvement at (α≤ 0.05) H.2.3: Verifying scope is recognised as having a significant effect on User Involvement at (α≤ 0.05). H.2.4: Controlling scope is recognised as having a significant effect on User Involvement at (α≤ 0.05). Subsequently, throughout this work, the direct link between User Involvement and Successful Project Scope Management will be measured in H.3, which includes one sub-hypotheses, detailed as follows:

H.3.1: User Involvement is recognised as having a significant direct on Successful Project Scope Management at (α≤ 0.05)

Lastly, throughout this work, the links between Project Scope Management Processes and their effect on User Involvement and Successful Project Scope Management will be considered through the application of H.4, which includes a total of four sub-hypotheses, detailed as follows:

H.4.1: User Involvement is recognised as mediating the link of Collecting requirements and Successful Project Scope Management at a significant level (α≤ 0.05)

H.4.2: User Involvement is recognised as mediating the link of defining the scope and Successful Project Scope Management at a significant level (α≤ 0.05). H.4.3: User Involvement is recognised as mediating the link of verifying scope and Successful Project Scope Management at a significant level (α≤ 0.05)H.4.4: User Involvement is recognised as mediating the link of controlling the scope and Successful Project Scope Management at a significant level (α≤ 0.05).

IV. Research Methodology

The questionnaire detailed the individual constructs to undergo measurement throughout the quantitative analysis [29]. Data were gathered through the adoption of a survey questionnaire [30] to determine the opinions of employees at the strategic and management level. Furthermore, random sampling was carried out across four IT departments at the government level.

A. *Sample size*

Owing to the varying sizes of the IT departments included in this work, the research has implemented a specific distribution technique so as to ensure the accurate representation of the research population. The sample of this work encompasses the four IT departments at the government level included in this work. Table 1 provides an overview of the findings of the sample size, in line with the analysis of 295 completed questionnaires.

TABLE I. The Questionair Collected and Completed Per Governemnt It Department

Sector	Completed questionnaires collected
First government IT department Income tax corporation	27 questionnaires
Second government IT department Jordanian water authority	33 questionnaires
Third government IT department Jordanian Ministry of Water	45 questionnaires
Fourth government IT department Grater Amman Municipality	190 questionnaires
Total	295 questionnaires

V. Data Analysis and Result

A. *Demographic Data Results*

The majority of the staff were males, with only one-fifth (21%) of the sample female. Furthermore, approximately one-third (36.9%) of the staff were aged between 31 and 35 years. In regards the participants' specialisations, approximately 41.1% of the participants were involved in group project management. Furthermore, just over one-quarter (26.4%) of the sample were in the role of System analyst, whilst a similar proportion (24.7%) carried out roles in hardware and software. In addition, more than half (53.9%) were seen to have at least seven years' experience. Table 2 provides an overview of the demographic data.

TABLE II. Demographic Data Results

Description	*Variable*	*Result*	*Percentage*
Gender	Male	233	79.0 %
	Female	62	21.0 %
Age	Less than 25 years	41	13.9 %
	25–30 years	80	27.1 %
	31–35 years	109	36.9 %
	More than 35 years	65	22.0 %
Area of Specialization	Hardware and software	73	24.7 %
	System analyst	78	26.4 %
	Project management	122	41.4 %
	Other	22	7.5 %
Experience	Less than 1 year	4	1.4 %
	2–7 years	83	28.1 %
	7–13 years	159	53.9 %
	More than 13 years	49	16.6 %

B. Validity and Reliability Result

Throughout this work, Smart Partial Least Square (PLS) was applied with the aim of measuring all hypotheses; there was the completion of data analysis through the application of two different phases [31], [32]: the first analysed the overall validity and reliability, whilst the second completed hypotheses testing.

- Path Loadings

Throughout this work, path loadings for all variables incorporated across the model were found to exceed (0.50) through the application of PLS software, thus meaning the acceptance of all variables, as highlighted by Falk & Miller [1992]. The figure below provides an explanation as to the path loadings (factors analysis result) for all variables included in the study model.

"Fig. 2" details six individual elements (Collecting requirements, Defining scope, Verifying scope, Controlling scope, User Involvement, and Successful Project Scope Management). Table 3 below provides an overview of the research constructs' Measure, Item and Factor Loading.

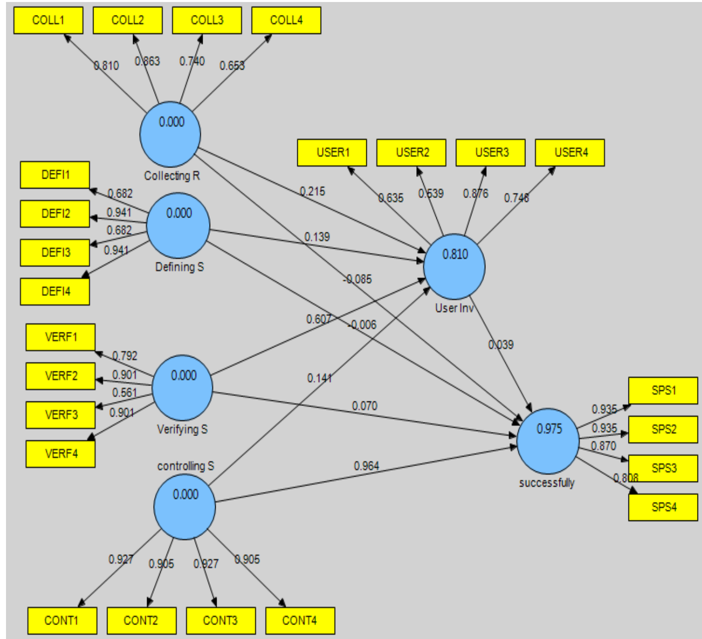


Fig. 2. Path Loading.

TABLE III. CONSTRUCT MESURMENTS

Variables	Item	Factors loading	Measure
Collecting requirements	COLL1	0.81	Collecting requirements is concerned with defining the functions.
	COLL2	0.66	Collecting requirements is concerned with detailing the features
	COLL3	0.74	Collecting requirements in the business setting require top management support
	COLL4	0.65	The requirements management plan provides good practice in regards Collecting requirements
Defining scope	DEFI1	0.68	Preparing project scope statement input includes the project charter
	DEFI2	0.94	The assets of the organisational process are applied in defining scope
	DEFI3	068	The key objectives pertaining to Defining scope determine the project scope statement
	DEFI4	0.94	As time progresses, project scope should become apparent
Verifying scope	VERF1	0.79	The approved project scope statement form the scope baseline
	VERF2	0.90	Scope verification includes stakeholders' acceptance of project scope completion
	VERF3	0.56	Verifying scope relies on project scope quality
	VERF4	0.90	Project managers apply leadership skills in such a way so as to deal with and manage the various obstacles experienced throughout the Verifying scope stage.
Controlling scope	CONT1	0.92	Scope control encompasses project scope change control
	CONT2	0.90	The objective underpinning scope control is to affect the factors underpinning scope changes
	CONT3	0.92	The key outcomes associated with controlling scope include Variance reports
	CONT4	0.90	The IT department is pivotal in achieving controlling scope success
User Involvement	USER1	0.63	User Involvement ensures the scope is kept realistic
	USER2	0.53	User Involvement results in the project selection process being valuable and good
	USER3	0.78	User Involvement throughout the processes of Project Scope Management provides guarantees in terms of flexibility when changing work requirements
	USER4	0.74	User Involvement is recognised as pivotal in line with change requests
Successfully Project Scope Management	SPS1	0.93	The success of Project scope requires performance to be measured throughout the course of satisfying project scope objectives
	SPS2	0.93	Project Scope Management success centres on providing users and sponsors with frequent outcomes
	SPS3	0.87	Project Scope Management success rests on ensuring the creation of a requirements management database
	SPS4	0.80	Changes from a systems perspective should be reviewed if they are to result in Project Scope Management success

- The Measurement Model

Through the use of PLS software, all Cronbach Alpha (CA) and composite reliability (CR) scores were seen to be higher than the suggested value (0.65) [33], which implies that all of the constructs detailed in the model offer good reliability. A commonly practical convergent validity standard is AVE (average variance extracted), as suggested in the work of Fornell & Larcker [1981]; this is seen to span 0.50–0.83, which falls within the scope of the cut-off value of 0.50 or higher. Table 4 below details the reliability AVE and CR for the constructs in this work, with all of them found to exceed the suggested levels.

TABLE IV. MESURMENT MODEL RESULTS

Constructs	Cronbach Alpha (CA)	Average Variance Extracted (AVE)	Composite Reliability (CR)
Collecting requirements	0.77	0.59	0.85
Defining scope	0.85	0.67	0.89
Verifying scope	0.80	0.64	0.87
controlling scope	0.93	0.83	0.95
User Involvement	0.67	0.50	0.79
Successful Project Scope Management	0.90	0.78	0.93

- R (Square) Test

The value of R (Square) coefficient is applied for the central approach to the structural Model's measurement for the suggested model, as highlighted in Table 5.

TABLE V. R (SQUARE) VALUE

Relation	R (Square)
The effects of Project Scope Management processes in line with Project Scope Management success without User Involvement mediation	0.81
The effect of the processes of Project Scope Management on Project Scope Management success with User Involvement mediation	0.97

The effects of Project Scope Management processes in line with Project Scope Management success without User Involvement mediation 0.81

The effect of the processes of Project Scope Management on Project Scope Management success with User Involvement mediation 0.97

Table 5 details that the R-squares for the variables (i.e. Project Scope Management success) without mediation achieves a value of 0.81, which is recognised as exceeding 25%, in line with the suggestion of Hair ,Black, Babin, Anderson and Tatham. [2006], which measures the accepted prediction level across the empirical paper [34]. In contrast, the variable's R (Square) value (i.e. Project Scope Management success) was mediated by User Involvement, with a value of 0.97 achieved, which is seen to exceed 25%; there was an increase in the percentage of Successful Project Scope Management R (Square) by 16% when there was the inclusion of User Involvement as the mediation variable in the link between Project Scope Management processes and Project Scope Management success.

- Latent Variable Correlations Test

There was the application of the Latent Variable Correlations Test in order to determine measurement construct discriminant validity (Collecting requirements, Defining scope, Verifying scope, Controlling scope, User Involvement, and Successful Project Scope Management). Table 6 provided below highlights the discriminant validity across this work.

TABLE VI. DISCRIMINANT VALIDITY

Construct	Collecting requirements	Defining scope	Verifying scope	controlling scope	User Involvement	Successful Project Scope Management
Collecting requirements	1.00					
Defining scope	0.42	1.00				
Verifying scope	0.76	0.55	1.00			
controlling scope	0.81	0.52	0.87	1.00		
User Involvement	-0.01	0.03	0.24	0.16	1.00	
Successful Project Scope Management	-0.01	0.04	0.26	0.19	0.98	1.00

In line with the information detailed in Table 6, discriminant validity was validated across the work, considering that the square root of the AVE achieved from all of the constructs was found to be greater than all other cross-correlations with other constructs.

C. Test Hypotheses

The hypotheses in the model underwent measurement through the completion of T-tests through the application of Bootstrapping in smart PLS to determine the T value. To begin with, the T value for Project Scope Management processes was measured by on Successful Project Scope Management without User Involvement as a mediatory factor. The figure below"Fig 3" provides an overview of this Bootstrapping (T value).

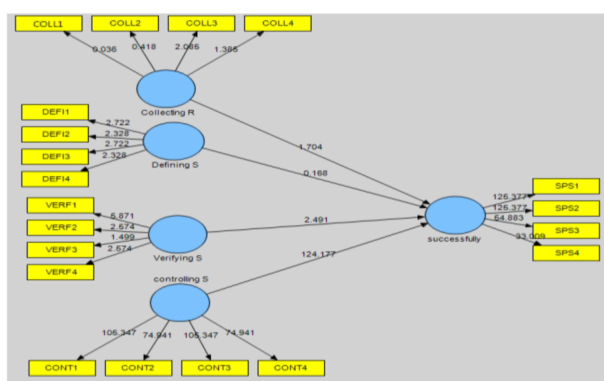


Fig. 3. Bootstrapping (T value) for Project Scope Management processes on Successful Project Scope Management without mediation of User Involvement.

In line with the above “Fig. 3”, the T value was determined by the authors through the application of Smart Partial Least Square (PLS) in order to test the hypotheses associated with Project Scope Management processes (Collecting requirements, Defining scope, Verifying scope, controlling scope) on Project Scope Management success without User Involvement applied as a mediatory factor. Table 7 provides the results.

TABLE VII. TEST OF PROJECT SCOPE MANAGEMENT PROCESS AND SUCCESSFUL PROJECT SCOPE

Relation (direct effect)	T value	Beta value
Collecting requirements and Successful Project Scope Management	1.70	-0.07
Defining scope and Successful Project Scope Management	0.16	0.00
Verifying scope and Successful Project Scope Management	2.49	0.09
Controlling scope and Successful Project Scope Management	124.17	0.96

The T value,in table 7 which is recognised between Collecting requirements and Successful Project Scope Management, was identified as 1.70. This is seen to be significant at the 0.05 level. Further, the Beta value is identified as –0.07. Notably, this value provides the explanation that one degree of change in Collecting Requirements will cause –0.07 degree change in Successful Project Scope Management. As such, these findings mean H1.1 can be accepted.

The T value, which is between the Defining scope and Successful Project Scope Management, is determined at 0.16. This is therefore not significant at the 0.05 level. Furthermore, the Beta is recognised as 0.00, which explains that modification to one degree of Related to the Defining Scope will induce change of 0.00 in Successful Project Scope Management. As such, these findings do not support the acceptance of H.1.2.

The T value, which is recognised as between the Verifying Scope and Successful Project Scope Management, was identified as 2.49. This is seen to be significant at 0.05 level. Further, the Beta value is identified as 0.09, which explains that change to one degree of Verifying Scope will induce change of 0.09 in Successful Project Scope Management. As such, these findings support the acceptance of H.1.3.

The T value, which is recognised between the Controlling scope and Successful Project Scope Management, was identified as 124.17. This is seen to be significant at the 0.05 level. Further, the Beta value is identified as 0.96, which explains that change in one degree of controlling scope will induce change of 0.96 in Successful Project Scope Management. As such, these findings support the acceptance of H.1.4.

In addition, the T value for Project Scope Management processes on Successful Project Scope Management was determined with the inclusion of User Involvement as a mediating factor. The T value for the study model can be seen detailed in the following figure.

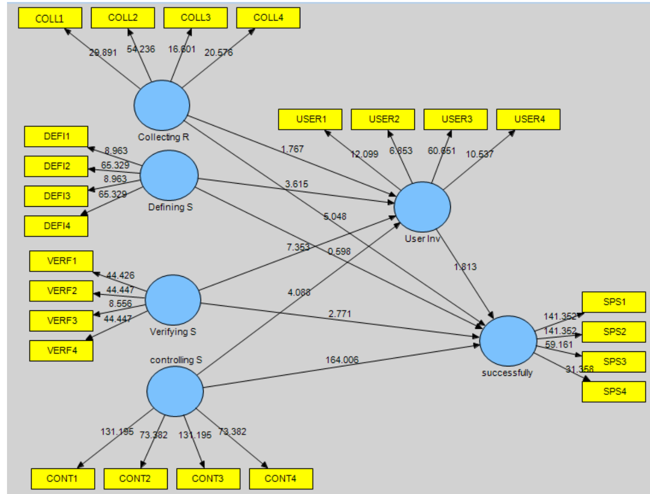


Fig. 4. Bootstrapping (T value) for Project Scope Management processes on Successful Project Scope Management with mediation of User Involvement.

In line with “Fig. 4”, the T value has been

Established through the application of the Smart

Partial Least Square (PLS) in mind of testing all hypotheses associated with Project Scope Management processes (Collecting requirements, defining scope, verifying scope, controlling scope) and User Involvement. The results can be seen summarised in the following table 8.

TABLE VIII. TEST RESULTS FOR PROJECT SCOPE MANAGEMENT PROCESS AND USER INVOLVEMENT

Relation (direct effect)	T value	Beta value
Collecting requirements and User Involvement	1.89	0.21
Defining scope and User Involvement	3.73	0.13
Verifying scope and User Involvement	7.95	0.60
Controlling scope User Involvement	4.54	0.14

Table 8 provide the T value, which is identified between the Collecting requirements and User Involvement, is recognised as being 1.89. Accordingly, it is recognised as significant at the 0.05 level. Further, the Beta value is recognised as 0.21, which explains that change to one degree of collecting requirements will cause change of 0.21 in User Involvement. As such, these findings facilitate the acceptance of H.2.1.

The T value, which is identified between the Defining scope and User Involvement, is recognised as 3.73. Accordingly, it is significant at the 0.05 level. Further, the Beta value is found to be 0.13, which explains that change to one degree of Defining scope will cause change equal to 0.13 in User Involvement. These findings facilitate the acceptance of H.2.2.

The T value, which is recognised between the Verifying scope and User Involvement, is identified as 7.95. Accordingly, this is viewed as significant at the 0.05 level. In addition, the Beta value is determined to be 0.60, which Explains that a change to one degree of Verifying scope will cause change of 0.60 in User Involvement. Such findings support H.2.3 acceptance.

The T value, which is recognised between the Controlling scope and User Involvement, is identified as 4.54. Accordingly, it is significant at the 0.05 level. Further, the Beta value is recognised as 0.14, which explains that change to one degree of controlling scope will induce change amounting to 0.14 in User Involvement. Such findings support H.2.4 acceptance.

In addition, as can be seen from Figure 4, the T value test is applied in the Smart Partial Least Square (PLS) with the aim of analysing the hypotheses linked to User Involvement and project success. Table 9 below provides an overview of the results.

TABLE IX. TEST RESULTS FOR USER INVOLVEMENT AND SUCCESSFUL PROJECT SCOPE MANAGEMENT

Relation (direct effect)	T value	Beta Value
User Involvement and Successful Project Scope Management	1.80	0.03

The T value, which is recognised between User Involvement and Successful Project Scope Management, is identified as 1.80. Accordingly, it is seen to be significant at the 0.05 level. Further, the Beta value is determined to be 0.03, as it is clear in table 9, which explains that change to one degree of User Involvement will induce change of 0.03 in Successful Project Scope Management. These findings provide support for H0.3 acceptance.

TABLE X. TEST RESULTS FOR COLLECTING REQUIREMENTS AND SUCCESSFUL PROJECT SCOPE MANAGEMENT MEDIATING BY USER INVOLVEMENT

Relation	Direct effect	Direct effect	Indirect effect	Total effect	Total effect
	T value	Beta	Beta	T value	Beta
Collecting requirements on User Involvement	1.89	0.21		1.89	0.21
User Involvement in Successful Project Scope Management	1.80	0.03		1.80	0.03
Collecting requirements on Successful Project Scope Management mediating by User Involvement			0.006		
Collecting requirements on Successful Project Scope Management	1.70 Partially mediate	-0.07		5.81	-0.06

Lastly, in the final section, the statistical results emphasise that the T value test result underwent application through PLS to validate whether User Involvement plays a mediatory role in the link between Project Scope Management processes (Collecting requirements, Defining scope, Verifying scope, controlling scope) and Successful Project Scope Management (see tables 10–13).

In relation to the above table 10, the T value identified between the Collecting requirements and User Involvement is recognised as having a value of 1.89. As such, it is seen to be significant at the 0.05 level. Further, the T value between User Involvement and Successful Project Scope Management is determined as 1.80. Accordingly, it was found to be significant at the 0.05 level. In regards the Beta value for Indirect Effect, this is calculated as being 0.006, which highlights that change of one amount in collecting requirements and User Involvement induces a change amounting to 0.006 in Successful Project Scope Management. Such findings underpin the acceptance of H.4.1. As a result, User Involvement is applied in order to partially mediate Collecting requirements and Successful Project Scope Management in government IT departments in Jordan.

In relation to the above table 11, the T value identified between Defining scope and User Involvement is recognised as having a value of 3.73. As such, it is seen to be significant at the 0.05 level. Further, the T value is determined as being 1.80 between User Involvement and Successful Project Scope Management. Accordingly, it was recognised as being significant at the 0.05 level. Further, the Beta value for Indirect Effect is calculated as 0.003, which further highlights that change in one amount in Defining scope and User Involvement will induce alteration amounting to 0.003 in Successful Project Scope Management. Such findings provide support for the acceptance of H.4.2. As such, User Involvement is recognised as being fully mediated between Defining scope and Successful Project Scope Management across the government IT departments in Jordan.

TABLE XI. TEST RESULTS FOR DEFINING SCOPE AND SUSSECCFUL PROJECT SCOPE MANAGEMENT MEDIATING BY USER INVOLVEMENT

Relation	Direct effect	Direct effect	Indirect effect	Total effect	Total effect
	T value	Beta	Beta	T value	Beta
Defining scope on User Involvement	3.73	0.13		3.73	0.13
User Involvement in Successful Project Scope Management	1.80	0.03		1.80	0.03
Defining scope on Successful Project Scope Management mediating by User Involvement			0.003		
Defining scope on Successful Project Scope Management	0.16 Fully mediate	0.00		0.59	0.003

TABLE XII. TEST RESULTS FOR VERIFYING SCOPE AND SUCCESSFUL PROJECT SCOPE MANAGEMENT MEDIATING BY USER INVOLVEMENT

Relation	Direct effect	Direct effect	Indirect effect	Total effect	Total effect
	T value	Beta	Beta	T value	Beta
Verifying scope of User Involvement	7.95	0.60		7.95	0.60
User Involvement in Successful Project Scope Management	1.80	0.03		1.80	0.03
Verifying scope of Successful Project Scope Management mediating by User Involvement			0.018		
Verifying scope of Successful Project Scope Management	2.49 partially mediate	0.09		2.79	0.108

In relation to the above table 12, the T value is identified as 7.95 between the Verifying scope and User Involvement. Accordingly, the value was seen to be significant at the 0.05 level. In addition, the T value between User Involvement and Successful Project Scope Management is calculated as being 1.80. As such, it was viewed as being significant at the 0.05 level. Further, the Beta value for Indirect Effect is highlighted as 0.018, which explains that change in one amount in Verifying scope and User Involvement will subsequently induce change amounting to 0.018 in Successful Project Scope Management. Such findings provide support for H.4.3. As a result, User Involvement is recognised as providing partially mediation between Verifying scope and Successful Project Scope Management across government IT departments in the Jordanian context.

TABLE XIII. TEST RESULTS FOR CONTROLLING SCOPE AND SUCCESSFUL PROJECT SCOPE MANAGEMENT MEDIATING BY USER INVOLVEMENT

Relation	Direct effect	Direct effect	Indirect effect	Total effect	Total effect
	T value	Beta	Beta	T value	Beta
Controlling scope of User Involvement	4.54	0.14		4.54	0.14
User Involvement in Successful Project Scope Management	1.80	0.03		1.80	0.03
Controlling scope of Successful Project Scope Management mediating by User Involvement			0.004		
Controlling scope of Successful Project Scope Management	124.17 partially mediate	0.96		5.81	0.964

In relation to table 13, the T value between Controlling scope and User Involvement is identified as being 4.54. As such, it is considered to be significant at the 0.05 level. Furthermore, the T value between User Involvement and Successful Project Scope Management is established as being 1.80. Accordingly, it was seen to be significant at the 0.05 level. Further, the Beta value for Indirect Effect is calculated as 0.004, which further highlights that change of one amount in Controlling scope and User Involvement will induce change amounting to 0.004 in Successful Project Scope Management. Such findings provide support for the acceptance of H.4.4. As a result, User Involvement is recognised as presenting partially mediation between Controlling scope and Successful Project Scope Management across governmental IT departments in the Jordanian context.

VI. CONCLUSION

Very little is known in the field of Project Scope Management processes and User Involvement, meaning it would be difficult to postulate as to the very best practice in this arena. This paper, however, provides outcomes that present reliable instruments for key factors in the analysis of Project Scope Management processes and User Involvement, with a number of valuable recommendations able to made in line with Successful Project Scope Management. In this work, a number of different factors were highlighted as requiring examination in consideration to their influence on Successful Project Scope Management within IT departments across governmental institutions in the Jordanian context. This work further centred on achieving empirically findings in regards the relative strength of causal relations on User Involvement, which fully mediates the link between defining scope and Successful Project Scope Management across governmental IT departments in the Jordanian context. Moreover, the findings provide insight into the relative strength of causal links on User Involvement, which suggest a partially mediation in regards the link between Collecting requirements, Verifying scope, and Controlling scope, and Successful Project Scope Management across IT departments in governmental institutions in the Jordanian context. It is stated in conclusion that Project Scope Management tools and techniques could undergo adaptation in IT departments in governmental institutions, with the value of such between demonstrated in the creative arena in Jordan.

REFERENCES

[1] Falk, R.F., and Miller, N.B. (1992) A Primer for Soft Modeling. Akron, OH: The University of Akron Press.

[2] Morris, P. W. G. (2010). Research and the future of project management. International Journal of Managing Projects in Business, 3(1), 139-146.

[3] Too,G.T., and Weaver ,P (2014) The management of project management: A conceptual framework for project governance, International Journal of Project Management, 32(8), pp. 1382–1394.

[4] Osorio,P.C., Quelhas,O.L., Zotes,L.P., Shimoda,E., & França,S (2014) Critical Success Factors in Project Management: An Exploratory Study of an Energy Company in Brazil, Global Journal of Management and Business Research: A Administration and Management, 14(10), pp. 38-50.

[5] Fageha,M.K., and Aibinu,A.A(2013). Managing Project Scope Definition to Improve Stakeholders' Participation and Enhance Project Outcome, Procedia - Social and Behavioral Sciences 74 (2013) 154-164.

[6] Ajelabi,I and Tang,Y(2010) The Adoption of Benchmarking Principles for Project Management Performance Improvement, International Journal of Managing Public Sector Information and Communication Technologies (IJMPICT). 1(2),pp.1-8.

[7] Mian Ajmal,M., Helo,P., and Kekale,T(2010) Critical factors for knowledge management in project business, Journal of Knowledge management, 14(1) pp. 156-168.

[8] Kwak,Y.H., and Anbari,F.T (2009) Analyzing project management research: Perspectives from top management journals, International Journal of Project Management 27, 435–446.

[9] Hornstein, H.A (2015) The integration of project management and organizational change management is now a necessity, International Journal of Project Management 33, 291-298.

[10] Travaglini ,A., Radujkovic M., Mancini ,M., (2014) building information modeling (bim) and project management a stakeholder perspective, organization, technology and management in construction · an international journal, 6(2), PP. 1058-106.

[11] Talet, A.N., Karadsheh,L., AL Jarrah,M.A., and Alhawari S (2018) Risk monitoring through better knowledge-based risk processes, Journal of Operational Risk 13(3), pp.1–25.

[12] Alnsour, B.H(2014) The use of virtual project teams for project management in Jordanian corporation, Eurasian Journal of Business and Management, 2(2),50-60.

[13] Alhawari S., Karadsheh L., Talet, A. Nehari Talet., and Mansour,E.,(2012) Knowledge-Based Risk Management Framework for Information Technology Project, International Journal of Information Management,32(1) , pp. 50-65.

[14] Alkhaffaf ,M., AL Jarrah,M.A., Karadsheh L., and Alhawari S (2018) Factors affecting ERP implementation success in Jordanian commercial banks sector, International Journal Economics and Business Research,15(4), pp.417-441.

[15] Al Freidi,S.S (2014) Determinants of the Best Practices for Successful Project Management, International Journal of u- and e- Service, Science and Technology 7(3), pp.173-186.

[16] Mirza,M.N., Pourzolfaghar,Z., and Shahnazari,Z (2013) Significance of Scope in Project Success, Procedia Technology, 9, pp.722-729.

[17] Olsson,N.O (2015). Implementation of pre-defined potential scope reductions in projects, Procedia Computer Science 64, 387-394.

[18] Alhawari S.(2016). Relationship among Project Management Processes and Knowledge Repository for Project Success, International Journal of Enterprise Information Systems (IJEIS), Vol.12 No.4 pp.16-30.

[19] Nikumbh,A.R., and Pimplikar, S.S (2014) Role of Project Management Consultancy in Construction Project, IOSR Journal of Mechanical and Civil Engineering, 10,(6)pp. 14-19.

[20] PMI. (2013) A Guide to the Project Management Body of Knowledge: PMBOK Guide, 5th ed., PMI – Project Management Institute, Pennsylvania, USA.

[21] Sánchez, M.A ., & Schneider,D.E(2014) Project management, strategic management and Sustainable development: a review of the literature, Revista Metropolitana de Sustentabilidade, 4(3), pp. 28-49.

[22] Marinho, M , Sampaio, S Lima, T., and Moura, H.D(2014) A Guide to deal with uncertainty in software Project management, International Journal of Computer Science & Information Technology (IJCSIT) 6,(5), pp. 1-20.

[23] Seresht,M.Z., Akbarijokar, M., Khosravi,S., Afshari,H(2014) Construction Project Success ranking through the Data Envelopment Analysis, Journal of Data Envelopment Analysis and Decision Science, 3(1) pp. 1-13.

[24] Nenni, M.E., Arnone,V., Boccardelli,P., and Iolanda,N(2014) How to Increase the Value of the Project Management Maturity Model as a Business-oriented Framework, International Journal of Engineering Business Management, 6, pp. 1-7.

[25] Mir,F.A and Pinnington,A.H (2014) Exploring the value of project management: Linking Project Management Performance and Project Success, International Journal of Project Management 32, pp. 202–217.

[26] Purna ,S, G. (2012). A model of critical success factors for software projects. Journal of Enterprise Information Management, 25(6), 537-558.

[27] Silvius, A. J. and Schipper, R P. (2014) Sustainability in Project Management Competencies: Analyzing the Competence Gap of Project Managers, Journal of Human Resource and Sustainability Studies, 2, pp. 40-58.

[28] Project Management Institute, I. (2004). A guide to the project management body of knowledge: PMBOK guide (3rd ed.). Project Management Institute. (p. 380).

[29] Saundres, M., Lweis, P., and Thornhil, A. (2009) 'Research Methods for Business Student', (6th ed), prentice hall.

[30] Hair, J.F. Black, B. Babin, B. Anderson, R.E., and Tatham, R.L. (2006) Multivariate Data Analysis, (6th ed), New Jersey: Pearson Prentice Hall.

[31] Fornell, C., and Larcker, D.F. (1981) Evaluating Structural Equations Models with Unobservable Variables and Measurement Error, Journal of Marketing Research, 18(1), pp. 39–50.

[32] Anderson. J., and Gerbing, D. (1988), "Structural equation modeling in practice: A review and recommended two-step approach," Psychological Bulletin, 103(3), 411-423.

[33] Nunnally, J.C. and Bernstein, I.H. (1994). Psychometric Theory. New York, McGraw-Hill.

[34] Gaur, A.S. and Gaur, S.S. (2006) Statistical Methods for Practice and Research: A Guide to Data Analysis using SPSS, (1st ed), Thousand Oaks, California: Sage.

Experimental Evaluation of Security Requirements Engineering Benefits

Jaouad Boutahar[1], Ilham Maskani[2], Souhaïl El Ghazi El Houssaïni[3]
[1,3] Systems, architectures and networks Team
EHTP, Casablanca, Morocco
[2] LISER Laboratory
ENSEM, Hassan II University
Casablanca, Morocco

***Abstract*—Security Requirements Engineering (SRE) approaches are designed to improve information system security by thinking about security requirements at the beginning of the software development lifecycle. This paper is a quantitative evaluation of the benefits of applying such an SRE approach. The followed methodology was to develop two versions of the same web application, with and without using SRE, then comparing the level of security in each version by running different test tools. The subsequent results clearly support the benefits of the early use of SRE with a 38% security improvement in the secure version of the application. This security benefit reaches 67% for high severity vulnerabilities, leaving only non-critical and easy-to-fix vulnerabilities.**

***Keywords*—Software security; security requirements engineering; security evaluation; security testing**

I. Introduction

Security Requirements Engineering (SRE) is the discipline that integrates security to Requirements Engineering, the very first step in the Software Development Life Cycle (SDLC). By adding security requirements to other system requirements during requirements engineering, a big improvement can be made in term of security vulnerabilities, software maintenance efforts and development costs. Moreover, OWASP , the leading organization in web application security, recommends focusing a big part of security flaws detecting efforts on the requirements engineering phase and the design phase[1]. There is related work which proves that it is critical to address security issues at the earliest phase, but few works try to measure just how much improvement can be obtained from applying an SRE approach. The goal of this paper is to make such a quantitative evaluation by developing two versions of the same web application, with and without using an SRE approach and evaluating their levels of security. The SRE approach that will be used is CompaSRE, a proprietary approach detailed in previous work[2]. For evaluation purposes, there's a plethora of testing methods and tools that could be used. A proper benchmark is needed to select the most appropriate. This paper is structured as follows. First, related work is discussed. Then, in the second section, the discipline of SRE is presented, along with definitions of its most important concepts, and the CompaSRE approach is explained. Then, in third section, the followed methodology is explained, along with the scope of the web application that will be developed for tests, and the selected test method and tools. Finally, the testing results and their variables are discussed in the fourth section.

II. Related Work

There is an abundance of security requirements engineering approaches. But when it comes to evaluating their performance, to the best of our knowledge, no source calculates how much is security improved by a certain SRE approach. Magnusson et al. tried to show how IT security investments can create value[3]. They studied models for return on investment on IT security in general. One of the models, developed by MIT, focused on proving the return of investment on secure software development, and showed that the earliest the security is addressed, the highest the benefit. This benefit was estimated at 21%. As reported in another paper [4], finding and fixing a software problem after delivery is often 100 times more expensive than finding and fixing it during the requirements and design phase. This is an evaluation of the financial cost of not thinking about security at the requirements engineering phase, which is their first number one recommendation on how to reduce software defect.

III. Security Requirements Engineering

This section presents the main SRE concepts, their definition and use in SRE. It also presents the CompASRE approach used to elicit and model security requirements for this experiment.

A. SRE approaches

An SRE approach refers to any method or process or framework that sets clear steps in order to elicit security requirements for a system to be at the Requirements Engineering phase. In a previous study, 9 approaches were studied. They go about eliciting requirements from different starting points: goals, users, or risks. But, ultimately, any SRE approach uses a different set of the same concepts. These concepts are drawn from both the fields of security and requirements engineering. All 9 approaches include identifying "goals", 7 of them identify threats, 6 of them identify stakeholders and 4 of them identify assets and risks. Table 1 offers a definition of these concepts, which is based on the ISO/IEC 27000:2016 vocabulary[5].

TABLE I. SRE CONCEPTS DEFINITIONS

Concept	Definition	Alternate labels
Stakeholder	Person or organization that can affect, be affected by, or perceive themselves to be affected by a decision or activity. Some approaches include other systems that have an interest in the IS. Are also included internal software agents to whom a goal will be assigned.	Actor, client, agent
Asset	Anything that has value to the organization, its business operations and their continuity, including Information resources that support the organization's mission (Data).	Information, Resource, Object
Goal	A Security objective that must be achieved by the system to be	Objective
Risk	Potential that threats will exploit vulnerabilities of an information asset or group of information assets and thereby cause harm to an organization	
Requirement	Need or expectation that is stated, generally implied or obligatory. Requirements are low level details of goals.	Goal, objective

B. CompASRE

The CompaSRE approach is the result of a personal previous work. It was designed as a comprehensive approach, incorporating the strengths and best practices found in existing approaches, and filling the gaps between them. It's based on the previous definitions and will be used in this experiment to elicit and model requirements. CompASRE, as illustrated in Fig. 1. below, is structured in five phases, each phase contains a set of activities to perform.

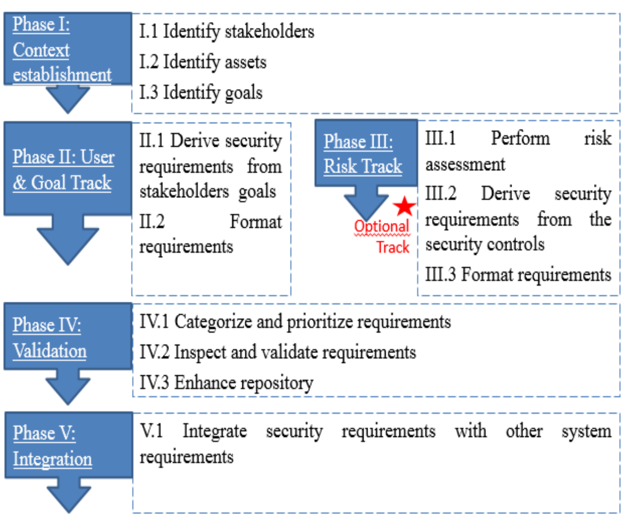


Fig. 1. CompASRE Steps.

The first phase "context establishment" aims to identify all common elements that are necessary to perform security requirements engineering in later phases. Then, the second phase "User & Goal track" aims to elicit requirements from the earlier identified goals. Security goals, as expressed by the stakeholders, are detailed and refined until reaching requirements. The third track is about deriving security requirement by doing a risk assessment, which implies analyzing threats and vulnerabilities. Risk assessment can be time and work consuming and was featured in only 4 out of the 9 studied SRE approaches. But assessing risks leads to thinking about security controls which might lead to new security requirements. Therefore, it was chosen to include this but keep it optional. The choice to perform it or not will depend on the type and size of the project. The more complex a project is, the more necessary it is to conduct a risk assessment. Once requirements are elicited (through phase 2 or 3), they must be modeled. The model created to be used with CompASRE or other SRE approaches is an extension of SysML requirements diagrams[6] and was presented in detail in previous paper [7]. In phase 4, elicited requirements are categorized and prioritized, then inspected for validation to resolve conflicts and eliminate redundancies. When an organization keeps a repository of requirements, this repository is to be updated. Finally, in phase 5, the security requirements are added to all other system requirements to complete the RE phase of the SDLC. Further validation might be necessary by the RE team.

IV. METHODOLOGY & TESTING

In this section, the methodology followed to conduct the study, the web app used as a test subject, the tests that were performed and the tools that were used are presented.

A. Methodology

The aim is to evaluate the positive impact of SRE on reducing system vulnerabilities, using specifically the CompASRE approach to elicit security requirements. To achieve that aim, first, the same web application was developed using 2 different software development lifecycles, resulting in 2 levels of security. The first version of the web app, the "*No Secure*" version, was developed following a classical waterfall lifecycle. This lifecycle was chosen because the app's functional perimeter is relatively small and unchanging. As for security, the way it was incorporated is the it's typically done in software projects where security is either not addressed at all (vulnerabilities will be patched after release) or is only addressed during the test phase. For this case, some minimal testing was done, in addition to correcting for the most obvious vulnerability "SQL injection". For the second *"Secure"* version, the same lifecycle was applied, but will be complemented by CompASRE. It means that, during the "Requirements Engineering" phase which is the first phase, CompASRE will be applied to elicit security requirements. Other later phases will be carried normally. Both versions were developed using the same language (JEE framework/java), same database management system and same development tools. Then, upon development completion, they were hosted on Microsoft's cloud solution Azure.

Finally, once both versions of the web app were hosted, security tests were conducted from a hacker's perspective. Quantitative results were obtained on vulnerabilities found in each version.

Both versions of the application are publicly available for fellow researchers on the following links: https://appgestionschool.azurewebsites.net/ for the secure version; https://appgestionschoolnosecure.azurewebsites.net/ for the no secure version.

B. Test Web Application

The web application that will be used as a test subject is a grade management system for an engineering school. The primary criterion of choice was that the app must be security sensitive, which is the case here since it manages security sensitive information such as students' grades and their personal information. The web app's functional perimeter includes security problematic features such as authentication, filling forms, uploading and downloading files. But the perimeter was kept small on purpose because auditing the two versions of the web app and comparing their security levels is the main goal, not web app's complexity. The app offers 3 menus for:

1) teachers to enter students' grades

2) students to view grades, download records, submit a claim and review individual contractor lecturers

3) administrative staff to manage grades, claims and reviews and upload program's data.

C. Testing Method

Empirically proving SRE benefits relies on obtaining quantitative results on the vulnerabilities found and their levels of severity. To obtain such results, security tests can be conducted in 3 manners:

1) Static Application Security Testing (SAST): It's a white box testing method where the testers have access to the system's code. The code is scanned to systematically detect and eliminate security vulnerabilities

2) Dynamic Application Security Testing (DAST): It's a black box testing method applied on running applications from the outside.

3) Penetration Testing: Manuel conducting of an application penetration scenario to target a specific asset or vulnerability that require human intervention.

Table 2 summarizes the pros and cons of using each test method in relation to this experiment.

TABLE II. COMPARISON OF TESTING METHODS

Testing method	Pros	Cons
SAST	• Finds vulnerabilities sooner during SDLC (before deployment)	• No quantitative report on found vulnerabilities • Tests conducted from the inside • Focus on code
DAST	• Tests conducted from the outside • Automated repetitive tests • Quantitative results • Used post-development • No previous knowledge of the app is needed	• Web app had to be deployed on an internet facing server
Pen Testing	• Allows more targeted tests requiring human intervention • Allows analyzing and exploiting other system components such as OS and hardware	• Tests and results can't be reproduced for both versions • Costly in term of time and human resources

From this comparison, the DAST testing method was chosen because, tests must be conducted from a hacker approach (i.e. a malicious outsider seeking harm), rather than a developer or tester approach (i.e. a development team member seeking to improve security). Furthermore, DAST's automated and repetitive tests will give better quantitative results to compare security in each version such as the number of vulnerabilities.

D. Testing Tools

Many DAST tools are available to conduct tests. These tools work by executing predefined attack scripts that send a request to the web app. The web app's response to the tool is analyzed to determine the existence of a vulnerability. Each tool has its own scripts, and its own parameters to configure security tests[8]. Choosing and using only one tool would give biased data as a result. For this reason, it was decided to use different tools to gather extensive data. The criteria for choosing these tools were:

1) oriented towards application vulnerabilities rather than network vulnerabilities

2) not only targeting a certain type of vulnerabilities

3) detailed results: vulnerability severity, page where found, …

4) available installation and use documentation

5) available user interface

6) Free install or extended free trial

After applying these selection criteria, 3 tools were chosen: OWASP ZAP 2.7.0 [9], Vega 1.0[10] and Acunetix trial version 12.0.180911136 [11]. For each one of these tools, both versions of the web app were tested with the same tool parameters, to guarantee reliable results.

V. RESULTS

In this section, comparison results are discussed as obtained by the testing tools, along with the variables that could influence them. Remaining vulnerabilities are examined.

A. Results Discussion

Fig. 2. shows the number of security alerts reported by each tool. A security alert arises when one vulnerability is detected in a certain location of the web app (a location can be a page, a field within the page, an embedded resource …). So, if the same vulnerability is detected in many locations, it would rise as many alerts as the locations where it was found. For each tool, the benefit was calculated as the percentage of reduction in the number of alerts (1).

$$security\ benefit = 1 - \frac{nbr\ of\ alerts\ in\ secure\ version}{nbr\ of\ alerts\ in\ non\ secure\ version} \quad (1)$$

Every tool reported a decrease in the number of alerts, with a security benefit average of 38%. But benefits varied greatly between tools, with Vega reporting the highest benefit (68%), while Acunetix reporting the lowest (18%). As for the number of alerts, they were quite close to the average with an average number of alerts of 20,66 for the "non-secure" version, and 11,66 for the "secure" version.

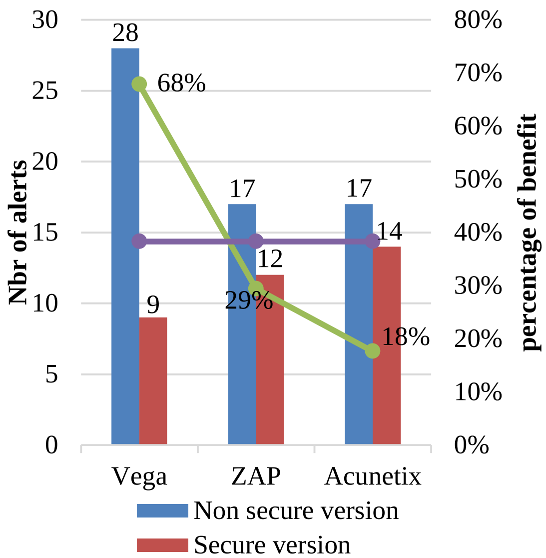


Fig. 2. Number of alerts and benefit per tool.

TABLE III. BENEFIT PER TOOLAND PER SEVERITY

Severity	Vega			ZAP			Acunetix			Average benefit
	No secure	*Secure*	*benefit*	*No secure*	*Secure*	*benefit*	*No secure*	*Secure*	*benefit*	
High	16	4	75 %	4	0	100 %	4	3	25 %	67%
Med	1	1	0%	4	3	25 %	8	8	0%	8%
Low	4	0	100 %	9	9	0%	3	2	33 %	44%
Info	7	4	43 %	0	0		2	1	50 %	46%

To get into the detail of alerts severity, Table 3 shows the number of security alerts, ranked by severity, as reported by each tool. Benefits are calculated by severity level. As shown in the diagram, the best benefits were obtained for high severity alerts with a 69% decrease. Indeed, high severity vulnerabilities are among those targeted early on during the SRE phase. As a result, the secure version of the web app is built with embedded security measures against those vulnerabilities.

As for the nature of the vulnerabilities that were found, table 4 presents the reported vulnerabilities for each version, along with the number of occurrences of each one. In the secure version, some vulnerabilities have disappeared (i.e. cross site scripting XSS), but some persisted, sometimes with fewer occurrences. This persistence of vulnerabilities can be explained, in some cases, by the fact that no requirement has been expressed against that vulnerability. In other cases, a requirement has been expressed against that vulnerability, but wasn't implemented during development (i.e. verbose error output). This is true in software development projects when a requirement is abandoned for time or cost reasons, or because of requirements mismanagement. There are also cases when a requirement is badly implemented, or implemented in only a few locations of the application, leaving some pages vulnerable. Regarding SQL injection, it's a high severity vulnerability that was considered during SRE, and all measures against it had been taken, but it was still reported by one tool in the secure version. Further manual tests could not confirm the vulnerability in the indicated location, so it's considered it as a false positive due to the tool itself.

TABLE IV. NUMBER OF ALERTS PER VULNERABILITY AND PER TOOL

Discovered Vulnerabilities	**No secure version**			**Secure version**		
	Vega	ZAP	Acunetix	Vega	ZAP	Acunetix
XSS Cross-site scripting	1		1			0
Integer Overflow	2					
SQL Injection		4	1			1
Page Fingerprint Differential Detected	12			3		
Session Cookie Without Secure Flag	1			1		
Apache tomcat infromation diclosure			2			2
Verbose Java error output	1		7	1		7
HTLM form without CSRF protection			1			1
String format error		1				
Javascript inter-domain sourcefile inclusion		3			3	
Autocomplete enabled in password field	4					
HTTP cache-control Header not set	1	3	1		3	
X-Content-Type-Options Header not set		3			3	
Lack of pretection against password brute force attack			1			1
Cookie Without Secure Flag	1		1	1		
X-Frame-Options-Header not set	3	3	1	1	3	1
Character Set Not Specified	1			1		
Blank Body Detected	1			1		
Possible sensitive information disclosure			1			1
TOTAL	**28**	**17**	**17**	**9**	**12**	**14**

B. *Remaining Vulnerabilities*

Even after applying the SRE approach, the secure version of the application still has some vulnerabilities. Indeed, no SRE approach claims to be able to eliminate all vulnerabilities. Furthermore, other phases of the SDLC play a big role in how secure a system would be. In the case of this application, the remaining vulnerabilities are not critical and can be corrected with a minimum of effort. More importantly, none of them comes from a design flaw which means that no redesign of the application will be necessary. Some could argue that, since SRE isn't failproof, all vulnerabilities could be left to be discovered and corrected at the end of the SDLC. This could be true for applications with a simple scope. But, for more complex systems, correcting vulnerabilities at the end can either be too costly, too cumbersome (impacting quality) or sometimes downright impossible because of design constraints. So, even if applying SRE has its own cost and isn't failproof, it still delivers better built-in security and quality.

C. Variables

It's noticeable that the nature of the discovered vulnerabilities and the numbers of their occurrences vary a lot from tool to tool. There are many variables to consider when interpreting these results. Any change in these variables would influence the results. The greatest variable is the testing tool itself. It's true that the tools work in a similar way: they crawl the website to find URLs, they attack said URLs with proprietary scripts and malicious input, then they analyze how the web app responds.

But where they differ is: how deep do they crawl? what vulnerabilities are tested for? what script/input is used to detect the vulnerability? The answers to these questions can lead to big differences detection efficiency, leading to false negatives (existing vulnerabilities that go undetected), or false positives (vulnerabilities reported but don't exist)[8]. This analysis by severity is also biased by the fact that the same vulnerability can be considered with different levels of severity (i.e. verbose error output, which is like apache tomcat information disclosure, is high severity for Acunetix and medium severity for Vega). Tools also differ in their settings and parameters. ZAP offered more advanced parameters such as creating contexts, authenticated attacks, adjusting crawling depth… It was found that the same parameters couldn't be applied to all tools, but for each tool, the parameters were the same in each version. Last but not least, if the tests had been done at development phase, prior to deployment, SAST tools would have been used, giving different results.

The second big variable is due to the web application used as a test subject. The more complex an application is, the more locations there are to find vulnerabilities. Security also depends on the technology used for the application. It was noted that applications in truly compiled application languages (i.e. C, C++) are more secure (in terms of (regarding OWASP Top 10) than general-purpose bytecode languages (i.e. Java, .NET) while scripting (i.e. PHP, ASP) are even less secure[12].

Furthermore, the type of application (social network, e-commerce ...) and industry (finance, e-gov ...) also influences what vulnerabilities would be found[13].

Finally, the SRE approach used to elicit requirements is another variable. Each approach has a different set of steps to follow, and various approaches work differently for different projects[14]. The security requirements elicited may also vary, for the same approach, depending on how correctly the approach was applied.

VI. Conclusion & Perspectives

The aim of the paper was to quantify the benefit of using an SRE approach. To achieve that, two versions of the same web application were developed. The contribution of this research is the evaluation of the security level of each version, proving the benefit of SRE. It was found that the second version was 38% more secure than the first. High severity vulnerabilities are the more impacted and were decreased by 67%. Vulnerabilities that persisted were either overlooked during SRE, or mechanisms against them were poorly developed or not developed at all. Remaining vulnerabilities are not critical and easy to correct. These results depend on many variables related to the testing tools, the web application subjected to the tests, and the SRE approach that was applied.

Plans for future work are to further investigate all discovered vulnerabilities to detect false positives and determine how that may have influenced the results. Tests could be done on other types of applications of different technologies to mitigate the effect of these variables. It's also planned to improve CompASRE's efficiency after a detailed study of its results, challenges and lessons learned.

References

[1] "Testing Guide Introduction - OWASP." [Online]. Available: https://www.owasp.org/index.php/Testing_Guide_Introduction. [Accessed: 13-Oct-2016].

[2] I. Maskani, J. Boutahar, and S. El Ghazi El Houssaïni, "Analysis of Security Requirements Engineering: Towards a Comprehensive Approach," IJACSA Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 11, pp. 39–45, Nov. 2016.

[3] H. S. Venter and Information Security South Africa, Eds., Peer-reviewed proceedings of the ISSA 2004 enabling tomorrow conference. ISSA, 2004.

[4] B. Boehm and V. R. Basili, "Top 10 list [software development]," Computer, vol. 34, no. 1, pp. 135–137, 2001.

[5] "ISO/IEC 27000:2016 - Information technology -- Security techniques -- Information security management systems -- Overview and vocabulary," ISO. [Online]. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=66435. [Accessed: 20-Oct-2016].

[6] "What is SysML? | OMG SysML." [Online]. Available: http://www.omgsysml.org/what-is-sysml.htm. [Accessed: 14-Nov-2017].

[7] I. Maskani, J. Boutahar, and S. El Ghazi El Houssaïni, "Modeling Security Requirements: Extending SysML with Security Requirements Engineering Concepts," Int. J. Appl. Inf. Syst., vol. 12, no. 9, pp. 30–36, Dec. 2017.

[8] A. Doupé, M. Cova, and G. Vigna, "Why Johnny can't pentest: An analysis of black-box web vulnerability scanners," in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, 2010, pp. 111–131.

[9] "OWASP Zed Attack Proxy Project - OWASP." [Online]. Available: https://www.owasp.org/index.php/OWASP_Zed_Attack_Proxy_Project. [Accessed: 01-Nov-2018].

[10] "Vega Vulnerability Scanner." [Online]. Available: https://subgraph.com/vega/. [Accessed: 01-Nov-2018].

[11] "Acunetix 14 Day Trial," Acunetix. .

[12] Veracode, "State of Software Security Report, Supplement to Volume 6: Focus on Application Development," 2015.

[13] Veracode, "State of Software Security Report, Volume 6: Focus on Industry Verticals," 2015.

[14] D. Mellado, C. Blanco, L. E. Sánchez, and E. Fernández-Medina, "A systematic review of security requirements engineering," Comput. Stand. Interfaces, vol. 32, no. 4, pp. 153–165, Jun. 2010.

Priority-Aware Virtual Machine Selection Algorithm in Dynamic Consolidation

Hanan A. Nadeem[1], Mai A. Fadel[3]
Computer Science Department
Faculty of Computing & Information Technology
King Abdulaziz University, Jeddah,
Saudi Arabia

Hanan Elazhary[2]
Computer Science Department
Faculty of Computing & Information Technology
University of Jeddah, Jeddah, Saudi Arabia
Computers & Systems Department
Electronics Research Institute, Cairo, Egypt

***Abstract*—In the past few years, many researchers attempted to tackle the problem of decreasing energy consumption in cloud data centers. One of the widely adopted techniques for this purpose is dynamic Virtual Machine (VM) consolidation. Consolidation moves VMs between hosts to decrease energy consumption. However, it has a negative impact on performance leading to Service Level Agreement (SLA) violations. Accordingly, selecting which VM to migrate from one host to another is a challenging task since it can affect performance. Researchers came up with several solutions and policies for efficient VM selection. In this paper, we exploit the fact that many tasks and users may tolerate some performance degradation which means, the tasks running on the VMs can be of different priorities. Accordingly, we propose augmenting consolidation with the priority concept, where low priority tasks are always selected first for migration. Towards this goal, we modified the popular Minimum Migration Time VM selection algorithm using the priority concept. The efficiency of the proposed algorithm is confirmed through extensive simulations using CloudSim toolkit and a real workload. The results show that priority awareness has a positive impact on decreasing energy consumption as well as maximizing SLA obligation.**

Keywords—Cloud computing; energy efficiency; service level agreement; VM consolidation; VM selection

I. Introduction

Virtual Machine (VM) consolidation is a useful technique for enhancing the utilization of the resources of cloud data center and reducing their energy consumption by leveraging the virtualization technology [1]. Virtualization provides the ability to create more than one VM instance on a single Physical Machine (PM) host. Accordingly, it permits more than one application to be allocated on a single PM in order to enhance the overall resource utilization and reduce the overall energy consumption of the data center.

Virtualization also allows live migration of VMs. Dynamic VM consolidation adopts live migration to minimize the number of PMs to which VMs are allocated. This is achieved by shutting down an underutilized host for more energy conservation and migrating its VMs to other PMs. However, this may lead to Service Level Agreement (SLA) violations. SLAs are established between cloud service providers and users to specify the required Quality of Service (QoS). After provisioning QoS, it should be monitored to ensure it is maintained throughout the service duration. QoS provisioning and monitoring are two classical problems in computer science [2] [3]. Unfortunately, when a VM migrates, its primary memory has to be transferred to the destination PM. Unfortunately, during this process, the requested CPU cannot be provided since the VM will be in a transition state which causes performance degradation and leads to SLA violation. Accordingly, enhancement of energy consumption and performance is a trade-off problem. Thus, dynamic VM consolidation techniques need to be designed with ultimate care such that not only power consumption is reduced, but also the requested QoS defined through SLAs is maintained. By carefully choosing which VMs to migrate when needed, consolidation can maintain more obligation of SLAs.

Dynamic VM consolidation is typically broken down into separate sub-problems [4]:

1) Host Overload detection: determining if a host is viewed as an overloaded one calling for a decision to choose one or more VMs to be migrated from this host.

2) Host Underload detection: determining if a host is viewed as an underloaded one calling for migrating all VMs allocated to this host to another, and switching the host to the low power mode.

3) VM Placement: finding a suitable destination host for allocating the migrated VMs from the overloaded and underloaded hosts.

4) VM Selection: a decision of which VMs should migrate from an overloaded host.

Simplifying the VM consolidation technique by diving it into four sub-problems and providing a separate algorithm for each one has the advantage of isolated examination and analysis of each algorithm to find a better approach. This work focuses on enhancement of the VM selection sub-problem.

Our proposed technique is based on the observation that some people might tolerate some performance degradation in services provided by a cloud and accept some SLA violations for cost savings while others cannot. For example, latency insensitive applications can tolerate some delay. From this prospect, we propose a novel approach in which we classify cloud user's tasks into two priority classes and deal with them in two different ways as follows:

- The users with high priority tasks should get a maximum obligation of their SLAs, and the cloud service providers should accept some energy consumption.
- The users with low priority tasks encounter reduced cost at the expense of accepting some SLA violations, while the cloud service providers gain more energy savings.

In other words, we treat users and their tasks differently and attempt to balance energy and performance as much as possible by considering priority when selecting a VM to migrate. It is worth noting that the cloud users' tasks priority will be assigned by the cloud provider as requested by the cloud users.

One of the most popular and effective VM selection algorithms in literature is the Minimum Migration Time (MMT) which picks the VMs with the minimum time required for migration. In this work, we propose priority-aware MMT algorithm to reduce the energy consumption while providing more SLA obligation for users with high priorities. The rest of the paper is organized as follows: Section II discusses related work, Section 3 describes the proposed VM selection algorithm, Section 4 describes the experimental settings and results; and finally, the conclusion will be in Section 5.

II. Related Work

VM selection and VM placement algorithms both comprise a challenging task of choosing a VM for migration and a preferable host for placement respectively. Several algorithms have been proposed in the literature for these purposes. Beloglazov et al. [4] proposed three VM selection algorithms; Random Selection (RS), MMT, and Maximum Correlation (MC). RS randomly chooses any VM for migration without any rules. The idea of MMT migration is to give preference to the VMs that require the minimum time for the whole migration process. Additionally, the VM with the maximum correlation coefficient relative to the other peer VMs on the same PM is the one selected for migration. The correlation is a parameter representing the effect of the VM on overloading the host. Moreover, the authors proposed Power Aware Best Fit Decreasing (PABFD) placement algorithm as a modification of the conventional Best Fit Decreasing (BFD) algorithm. The PABFD algorithm allocates each VM to the host that causes the least increase of power consumption due to placement.

Fu and Zhou [5] proposed a novel VM selection algorithm called Meets Performance (MP). This algorithm finds the host’s utilization deviation over the host overload threshold and compares it with the utilization of VMs on the host. It then selects the VM, whose migration results in shifting the utilization of the host nearer to the upper threshold. This is to reduce the number of migrations needed. Furthermore, the authors proposed a novel VM placement algorithm called Minimum Correlation Coefficient (MCC). This coefficient is used to describe the intense of correlation between the selected VM for migration and the destination host. The higher the correlation, the higher the effect on the performance of the destination host. The algorithm selects a VM with the minimum correlation with the target host to avoid degrading the performance of the other VMs allocated to it.

Rahimi et al. [6] proposed a VM placement algorithm based on priority routing. The main idea is to classify VMs based on their resource utilization, and classify hosts based on their resource availability, then give priority to the resources where CPU has the higher priority compared to the RAM, while the bandwidth has the lowest priority. After that, VMs are placed on the host with the most similar categories by creating a routing path table and considering resource priority. It is worth noting that this idea of priority is totally different from our proposed priority concept of tasks and users.

Farahnakian et al. [7] optimized VM placement by adopting Ant Colony Optimization (ACO) technique and proposed the Ant Colony System-based VM Placement Optimization (ACS-PO) algorithm. The proposed approach uses artificial ants in order to consolidate VMs and allocate them to the smallest number of active hosts based on the present requirements of the resources. What is interesting about those ants is that they work concurrently to develop VM migration plans based on a defined objective function.

Monil and Rahman [8] proposed a fuzzy VM selection algorithm. The fuzzy technique is an approach for tackling intelligent decision-making problems. The authors recognized that there are different VM selection algorithms in the literature offering different advantages; and generated a method which can aggregate the advantages of all of them in a single fuzzy logic tool. The input to the fuzzy tool is MMT and MC discussed above and the output is a VM selected for migration.

As discussed above and to the best of our knowledge, none of the algorithms on the literature classifies the cloud user's tasks based on their priorities. In this paper, such classification is exploited for delivering a better balance between energy consumption and performance in cloud data centers. Specifically, we modify the MMT algorithm using this priority concept as explained in the followings section.

III. Proposed Algorithm

As noted above, cloud service providers can satisfy their requirements of decreasing the cost of energy consumption while optimizing their resource utilization by using dynamic VM consolidation. Typically, the dynamic VM consolidation process sets up a threshold called the utilization threshold. It then monitors all active hosts' utilization ensuring that none of them exceeds this threshold. Whenever such a case is detected, some VM have to be offloaded from the corresponding source host and migrated to another destination host to avoid performance degradation on the source.

As mentioned before, dynamic VM consolidation is typically broken down into separate sub-problems [4]:

1) Host overload Detection: Local Regression Robust (LRR) [4] is one of the most efficient and widely used algorithms to set the utilization threshold and keep the summation of the utilization of all VMs bellow it. If the CPU utilization exceeds the set threshold, the consolidation process invokes the VM selection and VM placement algorithms to take an action.

2) Host Underload Detection: Host with minimum CPU utilization algorithm [4] is one of the popular and successful algorithms for this purpose. The idea is to find the host with the minimum CPU utilization compared to the other hosts. This host is recognized as the underloaded host, and all VMs on it have to be migrated with attention to the destination hosts after placing the VMs to avoid making them overloaded.

3) VM Placement: PABFD [4] discussed above is the most widely known and one of the most effective VM placement algorithms. It allocates each VM to the host which causes the minimum increase of power consumption due to this allocation.

4) VM Selection is a decision of selecting which VM has to be migrated. This is where our paper contributes. Our optimization is based on the priority concept where, as discussed earlier, latency insensitive applications that tolerate delay and users who may accept some performance degradation for price savings are given lower priorities as discussed below.

```
Priority-based Minimum Migration Time Selection algorithm
Input: overloadedHost    OutPut: VM selected for migration
foreach vm in overloadedHost do
  | if vm utilized by low priority
  | | lowPriorityList.add(vm)
foreach vm in lowPriorityList do
  | minMigTime ← MIN
  | selectedVm ← NULL
  | if vm.migrationTime() < minMigTime then
  | | minMigTime ← vm.migrationTime()
  | | selectedVm ← vm
 if selectedVm ≠ NULL
 return selectedVm
```

Fig. 1. Priority-aware MMT VM selection algorithm.

When the total requirements of CPU performance by the VMs exceed the available CPU capacity of the PM, the host is recognized as an overloaded host; the overloaded host may cause an increase in response time and a decrease in throughput. In this case, the cloud users do not get the expected QoS, and some VMs must be migrated from this host to decrease the CPU utilization of the host. Since live VM migration also has a negative impact on performance, low priority tasks will be the ones chosen for migration since those tasks accept some performance degradation due to migration and tolerate some SLA violation. On the other hand, the high priority tasks will be kept in the host saved from performance degradation due to migration.

As shown in Figure 1, we adopt the efficient most well-known MMT [4] VM selection algorithm and modify it using the priority concept. We make the selection decision in two phases. In the first phase, we select all the low priority tasks from an overloaded host and prepare a list for the second phase. The second phase selects from the low priority list the VM with the minimum time required for its migration in comparison to the other peer VMs allocated to the host. The time required for migration is defined [4] as the ratio between the RAM amount used by the VM and the available network bandwidth. After migrating the selected VM, the process is iteratively repeated as long as the host is still overloaded.

IV. Experiments and Results

Since the system of interest is Infrastructure as a Service (IaaS), which is a cloud environment intended to give the users a view of infinite computing resources, it is clear that we need to experiment with and evaluate the proposed VM selection algorithm on a large-scale virtualized data center infrastructure. However, conducting such an experiment in a real environment as a repeatable experiment is very difficult. Thus, simulations were chosen to evaluate the performance of the proposed VM consolidation technique and preserve the repeatability of experiments. CloudSim toolkit [9] is the simulation platform selected because it is a popular framework for simulating cloud computing settings. It can be used for modeling virtualized settings; supporting on-demand resource provisioning management. A recent extension to CloudSim allows energy aware simulations and supports energy-efficient strategies. This is in addition to allowing the simulation of service-oriented applications with dynamic workloads.

A. Experimental Settings

CPUs with dual-core are adequate for evaluating resource management algorithms intended to run on multi-core CPU architectures. In fact, it is essential to simulate a large number of servers to assess the efficiency of the VM consolidation mechanism. Thus, selecting less powerful CPUs for the simulations is beneficial because fewer workloads will overload a server [4]. To evaluate the efficiency of the proposed algorithm, we simulated a datacenter containing 800 heterogeneous PMs with two configurations:

- HP ProLiant ML110 G4 (Intel Xeon 3040, dual-core 1860 MHz, 4 GB, 1 Gbps).
- HP ProLiant ML110 G5 (Intel Xeon 3075, dual-core 2660 MHz, 4 GB, 1 Gbps).

The characteristics of the VM instances are of types identical to those of Amazon EC2 instances except that all VMs are single core. This is because the workload data employed in the simulation comes from the single core:

- Extra-large Instance (2500MIPS, 3.75GB).
- High-CPU Medium Instance (2500 MIPS, 0.85 GB).
- Small Instance (1000 MIPS, 1.7 GB).
- Micro instance (500 MIPS, 613 MB).

B. Performance Metrics

Different performance metrics are used for evaluating the proposed VM selection algorithm. We adopt the same metrics proposed and elaborated by Beloglazov and Buyya [4]:

- Energy consumption in Kwh
- SLATAH (SLA violation Time per Active Host)
- PDM (Performance Degradation due to Migrations)
- SLAV rate (SLA Violation rate), which is the product of SLATAH and PDM

- Number of VMs migrated
- ESV, which is the product of energy consumption and SLAV rate

C. Workload Data

To validate the proposed VM selection algorithm with more applicable simulations, a real workload from a CoMon system which is a monitoring infrastructure for PlanetLab [10] was used. This CPU utilization data is collected from more than thousand VMs from servers distributed over five hundred locations around the world every five minutes. Data is created from ten days randomly chosen during the months of March and April, 2011. The median value is calculated over the ten days and used with each performance metric. The basic features of this data are presented in Table 1.

TABLE I. WORKLOAD CPU UTILIZATION STATISTICS

Date	Number of VMs	Mean	St. Dev	Quartile 1	Median	Quartile 3
03/03/2011	1052	12.31 %	17.09 %	2%	6%	15%
06/03/2011	898	11.44 %	16.83 %	2%	5%	13%
09/03/2011	1061	10.70 %	15.57 %	2%	4%	13%
22/03/2011	1516	9.26%	12.78 %	25	5%	12%
25/03/2011	1078	10.56 %	14.14 %	2%	6%	14%
03/04/2011	1463	12.39 %	16.55 %	2%	6%	17%
09/04/2011	1358	11.12 %	15.09 %	2%	6%	15%
11/04/2011	1233	11.56 %	15.07 %	2%	6%	16%
12/04/2011	1054	11.54 %	15.15 %	2%	6%	16%
20/04/2011	1033	10.43 %	15.21 %	2%	4%	12%

D. Experimental Results

Using the PlanetLab workload data, we compare the original MMT algorithm with the priority awareness optimization. Figure 2 shows the energy consumption due to consolidation in Kwh. The results show that the priority-aware MMT decreases the energy consumption by 13%. Figure 3 shows the percentage SLATAH; the priority-aware MMT decreases the SLATAH metric by 42%. Figure 4 describes the performance degradation due to migration; choosing the low priority tasks that have minimum time to complete migration can provide 37% decrease in degradation. Figure 5 describes the overall SLA violation delivered by the consolidation technique; priority-aware MMT can provide 21% reduction of SLA violation. Figure 6 shows the energy consumption and SLAV rate; the rate is decreased by 31% when using the priority-aware MMT. Finally, Figure 7 shows the number of VMs migrated due to consolidation; they are reduced by 40% in case of the priority-aware MMT. Since live VM migration results in an overhead on the system, the better consolidation mechanism is the one which requires fewer migrations.

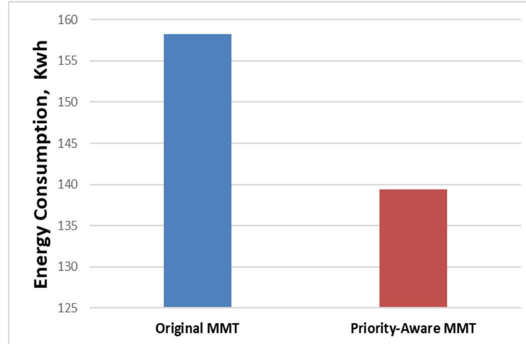


Fig. 2. Energy consumption.

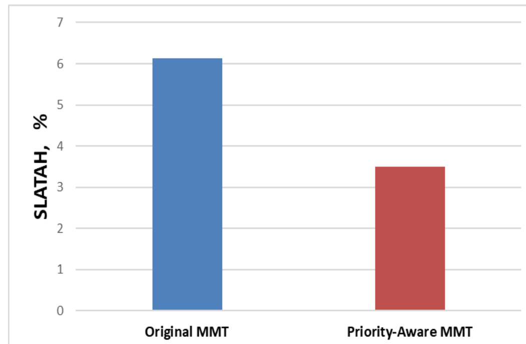


Fig. 3. SLA Time per Active Host.

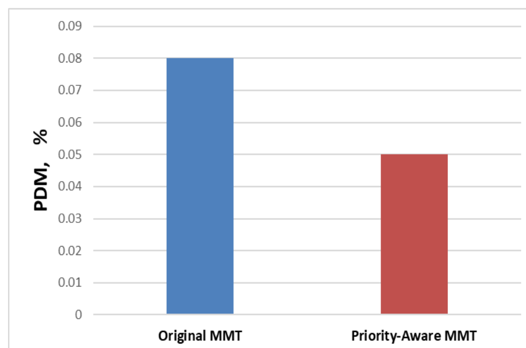


Fig. 4. Pefrormance degradation due to migration.

Based on the results above, it is clear that priority awareness has a considerable positive effect on all performance metrics. In other words, the proposed priority-aware VM selection algorithm is an efficient optimization in comparison to MMT algorithm regarding all metrics.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel priority-aware VM selection algorithm, which takes into consideration the priorities of tasks. This algorithm is original since, to the best of our knowledge, it is the first to exploit the priorities of cloud tasks and users. We selected the widely-used and successful MMT VM selection algorithm and showed that modifying it using priority-awareness has a positive effect on energy consumption and on all performance metrics.

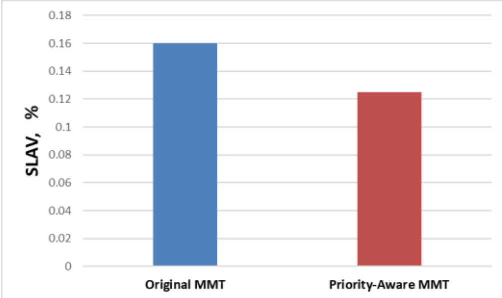


Fig. 5. Overall SLA violation.

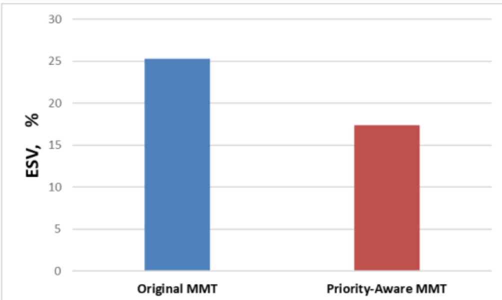


Fig. 6. Energy-SLA violation ratio.

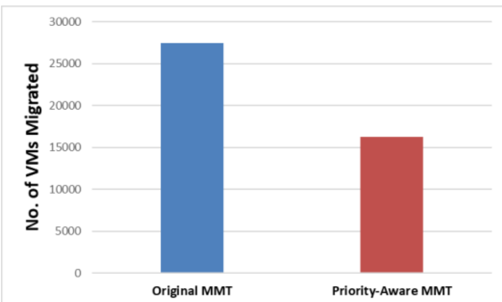


Fig. 7. Number of VMs migrated.

As future work, we intend to apply the proposed priority concept and experiment with different other VM selection algorithms to ensure the reliability of the proposed approach and further assess its effectiveness.

REFERENCES

[1] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. "Xen and the art of virtualization," in Proc. the 19th ACM symposium on Operating Systems Principles, Bolton Landing, NY, USA, 2003, pp.164-177.

[2] H. Elazhary and S. Gokhale. “An integrated approach for QoS provisioning and monitoring,” in Proc. IASTED PDCN, Austria, 2004.

[3] H. Elazhary and S. Gokhale. “Integrating path computation and precomputation for quality-of-service provisioning,” in Proc. ISCC’04, Alexandria, Egypt, 2004.

[4] A. Beloglazov and R. Buyya. "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers," Concurrency and Computation: Practice & Experience, vol. 24, no.13, pp. 1397-1420, 2012.

[5] X. Fu and C. Zhou, " Virtual machine selection and placement for dynamic consolidation in Cloud computing environment", Frontiers of Computer Science Springer Link , vol. 9, no. 2, pp 322–330, 2015.

[6] A. Rahimi, L. Khanli, and S. Pashazadeh, "Energy efficient virtual machine placement algorithm with balanced resource utilization based on priority of resources," ComEngApp Journal, vol. 4, no. 2, 2015.

[7] F. Farahnakian, A. Ashraf, P. Liljeberg, T. Pahikkala, J. Plosila, I. Porres, and H. Tenhunen. " Energy-Aware Dynamic VM Consolidation in Cloud Data Centers Using Ant Colony System" in Proc. 2014 IEEE 7th International Conference on Cloud Computing, Anchorage, AK, USA, 2014.

[8] M. Rahman and R. Monil. "VM consolidation approach based on heuristics, fuzzy logic, and migration control", Journal of Cloud Computing, vol. 5, no. 8, 2016.

[9] R. Calheiros, R. Ranjan, A. Beloglazov, C. De Rose, and R. Buyya. "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," Software: Practice and Experience, vol. 41, no. 1, pp. 23-50, 2011.

[10] K. Park and V. Pai. "CoMon: A mostly-scalable monitoring system for PlanetLab," ACM SIGOPS Operating Systems Review, vol. 40, no. 1, pp.65-74, 2006.

Three-Phase Approach for Developing Suitable Business Models for Exchanging Federated ERP Components as Web Services

Evan Asfoura[1], Mohammad Samir Abdel-Haq[2], Houcine Chatti[3]
Department of Management information system
Dar Al Uloom University
Riyadh /Saudi Arabia

Abstract*—The importance of business models has increased significantly in the last decade, especially in the Internet. The cause of this increase is the effect of Internet and the associated applications and their business processes regarding to the business model. These effects include, for example, the emerging technical and economic aspects of a business model on the Internet, the support and transformation of traditional business models, and the arise of new business ideas based on that technology. One of these new ideas is: how distributed Enterprise Resource Planning (ERP) systems or federated ERP (FERP) systems as web services (WSs) can cover the increasing demands of small and medium sized enterprises (SMEs) for business software covered.This paper aims to provide a derived developing approach with three phases that will lead to three suitable concepts that identify suitable business model for FERP System with different scenarios of value exchanging. *The results of this work will be conceptual models that describe the character, role and revenue models that identify FERP exchanging business model.

Keywords—FERP system; ERP web services; FERP mall; ERP workflow; developing approach

I. Introduction

An enterprise resource planning (ERP) system is a standard software system, which provides functionalities to integrate and automate the business practices associated with the business process of an enterprise. The integration is based on a common data model for all system components and extents to more than one enterprise sectors [2; 27; 36; 38; 49; 4]

The increasing number of the small and medium enterprises' employees extended the need for flexible functionalities in ERP systems. Small and medium sized enterprises (SMEs) are facing different problems when they buy ERP systems, such as [1, 12, 19]

- Not all installed components are needed.
- The usage of licenses, Administration, and maintenance of these products are too expensive.
- Normally, ERP systems are complex and overloaded with stuff, functions and options, therefore it is hard for new user to learn.
- High-end Hardware is required.

Therefore, in the last few years the idea of the Federated ERP-System in the basis of Web-Services has evolved. A federated ERP (FERP) system (see figure 1) is an ERP system, which consists of system components that are distributed within a computer network. The overall functionalities are provided by an ensemble of allied network nodes that all together appear as a single ERP system to the user. Different ERP system components can be developed by different vendors [1; 19; 20; 23]. Through the FERP system, enterprises pay only for components deemed necessary. Also, the needed end-hardware is made available by the service provider, which in turn, reduces costs [19].

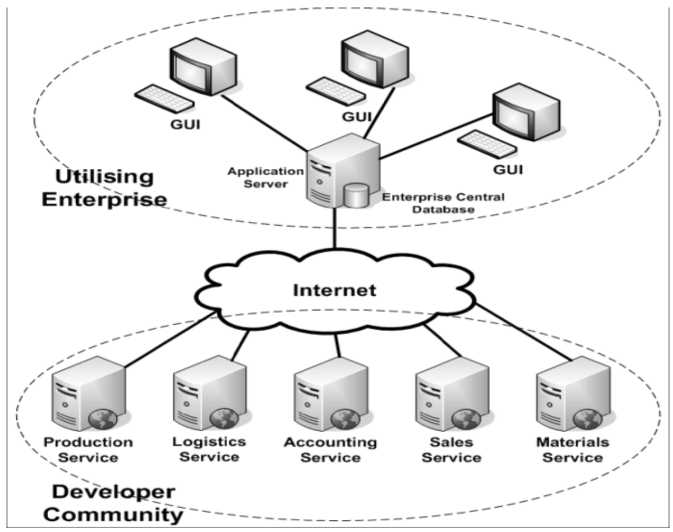


Fig. 1. Architecture of FERP systems [19].

An ERP system component in this case is reusable, closed and marketable software module, which provides services over a well-defined interface. These components can be combined with other components in an unpredictable manner [47] Those components are described, published and used as Web services

A web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other web systems interact with the web services in a manner prescribed by its description using SOAP-messages typically conveyed using HTTP with an XML serialization in conjunction with other web-related standards [53] The search for these services by FERP Systems is covered by the functionality, which is considered as the logical and stable construction stone in ERP system [19].

FERP system is considered as an advanced Software as a service paradigm, which is provided from various, and independent ERP web services providers. As a service paradigm has been widely from the technical side [16;18;21;44, 24, 25]. Therefore, we will focus in this paper on value exchanging and integrating of Software as service through providing logical developing approach of appropriate business model concepts for exchanging FERP components as web services. The realization of FERP scenario are divided into two processes:

- Production process: focus on the isolation of ERP components, description, publishing those components as web services through the web standards and the integration of these web services in FERP workflows.
- Marketing process: focus on the Exchanging of ERP components as web services need through a suitable business model. Therefore, businesses should be adopted to fulfil the new idea's need.

This paper will focus on the marketing process to find out the suitable business model concept for exchanging of FERP components as ERP web services. Therefore, the aim of this work is to present an derived developing approach with three phases that will lead to three suitable concepts that identify suitable business model for FERP System with different scenarios of value exchanging, The results of this work will be conceptual models that describe the character, role and revenue models that identify FERP exchanging business model.

In the second section of this work, we will present the research methods for formulizing the suitable developing approach, which will lead to the targeted business model concepts as results with qualitative and quantitate evaluation. In the third and fourth section, in fifth section we will discuss the result of this work to explain the added value from this scientific work. The last section will includes the conclusion and the work summary.

II. Research Methods and Results

The paper is exploratory and conceptual in nature and builds on existing literature research. Through literature review and crossing the business model definitions by many researchers and experts in business model field (see table 1). Table 1 has been identifies four characteristics through the meeting points of the definitions of business models that can be considered for the characterization of business model types on the Internet [8,9,10,11]. These characteristics and their attributes are: integration in internet economy (IG in IE), targeted business field, products and offered goods and revenue model which includes revenue source and form. The basic model type related to offered goods (see [52]) should be considered in addition to the characteristics that has been shown in Table 1 for the realization of a new business models (see table 2).

There are many business models depending on business field, e.g., business-to-business (B2B), business to customer (B2C), customer-to-customer (C2C), customer to business (C2B), administration-to-administration (A2A), administration to business (A2B), business to administration (B2A), etc. In this paper we will focus on the most popular models which are (B2B), (B2C) and (C2C) (for more details [5, 30]. Not all products or goods can be traded alike through E-markets. There are many factors (related to the vendor or to the customer) that could affect the trading process [43]. These goods are divided into two categories: Material goods and immaterial goods [31; 39; 43, 45]

TABLE I. Cross of Business Model Definitions by Many Researchers and Experts in Business Model [17]

Characteristics / Definitions by	Integration grad in the Internet economy	The Actors And business fields	Products and Offered goods	Cash flow and revenue
[46]		**X**	**X**	**X**
[54]		**X**	**X**	**X**
[55]		**X**		**X**
[56]		**X**	**X**	**X**
[39]	**X**	**X**	**X**	**X**
[3]			**x**	**x**
[26]			**x**	**X**
[32]		**x**	**x**	**X**
[40]			**x**	**X**
[13]		**x**	**x**	**X**
[37]	**x**			**X**
[51]		**x**	**x**	**X**

The degree of integration in the internet's economy depends on the ability to implement the transaction phases electronically, distinguished here: full and partial integration [5]: The relevant 4-basic business models types are "content", "connection", "commerce" and "context" [52] The fact is that most of business models belong to one of the 4-Basics types. a business model could be a hybrid of more than one type. 4-Basic types [35]

TABLE II. The Selected Characteristics and Their Attributes

Characteristics	Attributes					
ID in IE	Full	Partial				
Business field	B2B	B2C	C2C			
Basic business models	Content	Context	Commerce	Connection		
Offered goods	Tangabile Goods	Intangabile Goods				
	Physical Products	Physical Services	Digital Products	Digital Services	Information	Special goods
Revenue sources	Products	contacts	Information			
Revenueform	direct and transaction dependent	direct and transaction independent	indirect and transaction dependent	indirect and transaction independent		

The sources of revenue of business models fall into three categories: Products, contacts and information [42,52] The forms of revenue were classified by Wirtz on one hand according to the players (i.e., buyers and sellers) into direct and indirect revenue and on the other hand according to the pricing conditions into transaction-dependent and transaction-independent [52].

In the next, we will present an approach for developing of appropriate business model for federate ERP systems through logical steps. Through the characteristics of business model which has been shown in the second section we present in the following figure a developing approach which lead us to the appropriate concept for the exchange of FERP components as web services (see figure 2). This developing approach consists of three phases every phase involves one or more of steps with the considering of the existent models, architectures and the special requirements of an FERP system.

The first phase is the characterization phase: In this phase we should answer the following three questions

- Which business model type nowadays is existed and more suitable for FERP as a new Business idea?
- What could be provided in the expected business model?
- To which business fields belongs this business model?

These three questions can be answered and analyzed by using the characteristics “Business Fields”, “Basic Business Models” and “Offered Goods” defined in table 2.

The second phase is the adapting phase: This phase aims to identifying the roles of the actors during the transaction phases and the exchange of services with taking into account-minimizing potential problems and risks in the case of FERP systems.

The third phase is the goal phase: This phase characterizes the revenue model depending on the offers and the actors roles with considering the appropriate pricing models

Then this developing approach in figure 2 is derived from the table 2. Table 2 serves information for classifying the existing business models. In contrary, the approach in figure 2 presents basics phases for developing a new business model for a new business idea.

The targeted results of this developing approach are a character-concept, role-concept and revenue-concept of an suitable business model for exchanging FERP systems as Web services. In the flowing subsections.

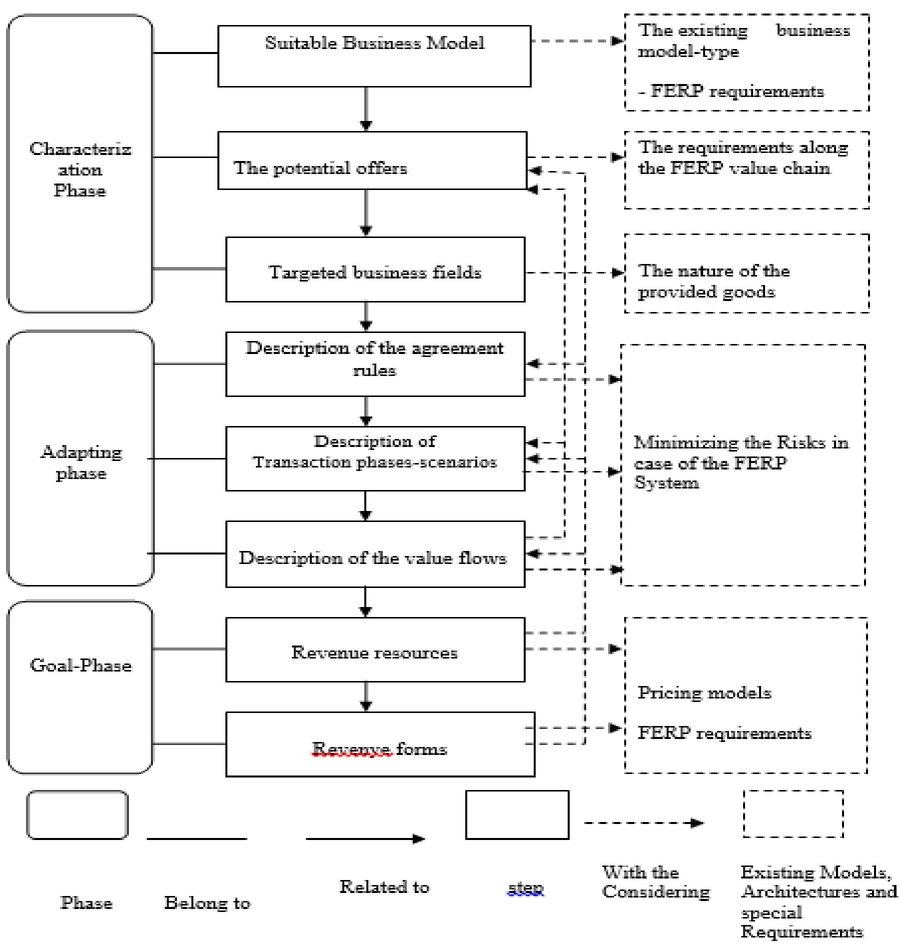


Fig. 2. Eveloping Approach for realization of Appropriate FERP Business model.

A. Characterization Phase

In this part we will discuss the three steps that we explained previously

a) Appropriate Type of Business Model for Marketing of Distributed Erp Systems

The first step is the realization of a suitable business model for marketing of standardized and independent FERP components as WSs. It aims to develop the more suitable business model type, which can be adapted to the requirements of FERP systems Business Model discussed in [5]. The result was that the mediator's business model as a FERP mall is the closest and most appropriate type of business model to cover the lack of pre-defined communication channels and trust gap between the FERP WS providers and the end-user enterprises (SMEs). This intermediary type will be consider as the basis for the next steps by realization of FERP business model

b) The Potential Offers

As second step for characterization of a FERP mall that is identified in first step, is the identifying the expected offers in case of FERP systems. These offers to customers depend on the requirements of a FERP system along the value chain. This steps discussed also in [6]. The possible offers classified in two categories (see figure 3):

Main offers that are necessary subsystems for the realization of the FERP systems. These can be deduced from the FERP value chain. This category includes (FERP Web Services (FERP WSs, WS-publishing services and ERP Workflow-Definitions (WF-Definitions)

Supporting offers. that support the core business of the business model to be developed and strengthen related to FERP-WS providers (developers) and customers (user enterprises). Taking into account the nature of the FERP systems (as WS), This category includes Web Services Developing Environment (WSD-Environment), Testing Services (T-Services) and WS evaluation information (WSE-Information),

In the last step in this phase, we will discuss the business field of this business model according to the nature of the offers.

a) The Targeted Business Fields

Table 3 shows a classification of the goods which are exchanged through the aimed FERP Mall between the actors. , X"-letter in the business field column of this table refers to X developer who develop the web services. We call it "Anonym developer" because the market in the case of FERP system is open for all the developers independent of their personalities.

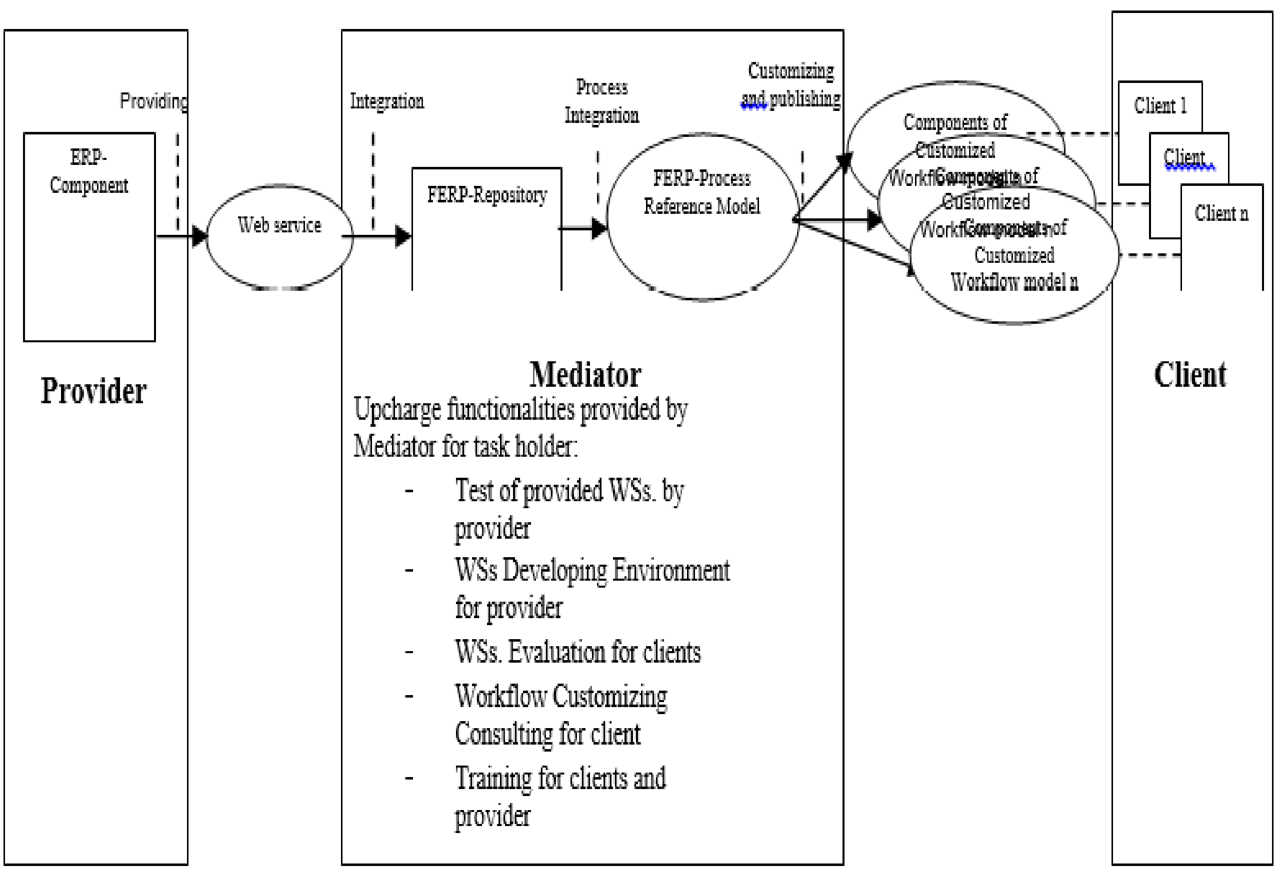


Fig. 3. FERP Scenario (or FERP value Chain).

TABLE III. THE BUSINESS FIELDS OF THE TARGETED BUSINESS MODEL (FERP MALL)

Goods	provider	costumer	Business fiel
FERP WSs	WS Developer	User-Enterprise	X2B
WF Definitions	Mediator	User-Enterprise	B2B
WS Publishing	Mediator	WS Developer	B2X
WSE Environment	Mediator	WS Developer	B2X
T Services	Mediator	User-Enterprise	B2B
WS Evaluations	Mediator	User-Enterprise	B2B
Consulting	Mediator	User-Enterprise	B2B
Training	Mediator	The Employees of the user-Enterprise	B2C

The goals of all the relationships through FERP Mall are business goals. Every costumer invests by making business through buying of the offered goods from the provider. Therefore, all the relationships through this mall belong to B2B field. The relation by the offering of training for the employees of the customer enterprise seems as B2C but no cash flows because the user-enterprise pay for the training curses as part of its investment. Then this FERP mall is absolute B2B business model.

As result of the characterization phase, Figure 4 represents the FERP Mall character. This mall includes several shops offering FERP WSs their functionalities belong to various business sectors.

In addition to intermediation of WSs, this mall also appears as an integrator of FERP-WSs in FERP processes through a workflow reference model that considers all possible scenarios of an enterprise. This integrator is responsible against the user enterprise for the quality of the FERP processes. This integration enables customers to cover their requirements for ERP functionality from one hand. All ERP shops among the FERP Mall look like a single ERP shop which have a single shopping and payment system

In the first phase, we have presented the first three steps from our developing approach. These steps represent the first phase (characterization-phase). Therefore, as feature work, we will discuss the role-concept of the FERP mall as mediator (intermediary) between the other parties and after that, we can present the revenue-concept of this FERP mall as the last part of the presented developing approach for the conception of FERP business model.

B. *Adaption Phase*

The second phase of the developing approach aims to adapting the roles and extending the intermediary's (FERP) mall for minimizing potential risks and increasing stakeholder security. This phase includes the following three steps (see Figure 2):

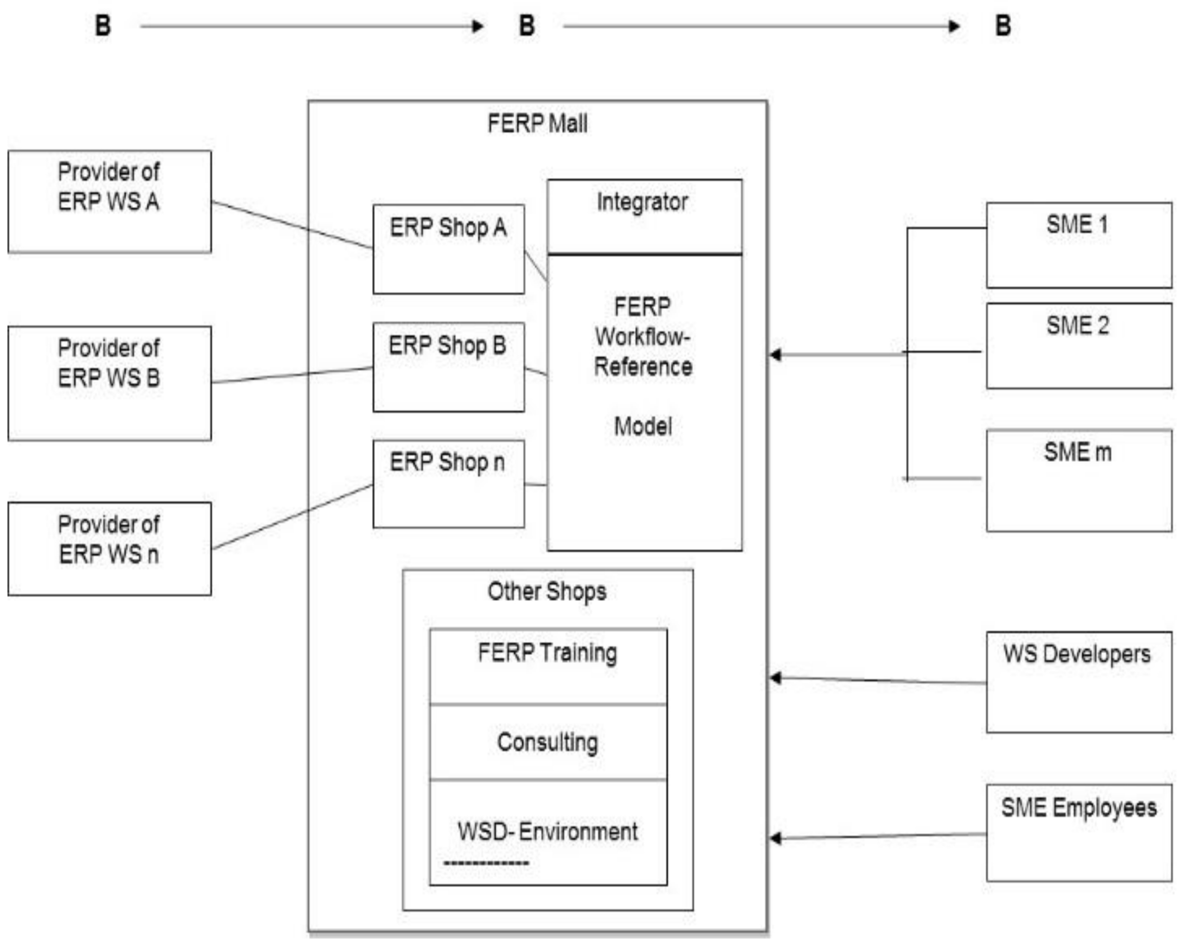


Fig. 4. The character-concept of FERP Mall

- Description of the Service level agreement between the actors (agreement rules).
- Description of the transaction phases scenarios for user enterprises (SMEs) with different requirements.
- Description of the value flows between the actors.

In this phase should the potential to harmful risks be identified and considered by identifying the FERP Mall roles as FERP workflow designer and FERP-WS intermediary between the FERP-WS developers and the user enterprises (SMEs) during the transaction phases. The most important risk that should be considered in this phase especially in case of FERP systems is the difference between the levels of security required by SMEs. The required security level has been classified depending on survey result which has been done among hundred SMEs in Germany.

The result of this survey was: About 18% of the surveyed SMEs are allowed to make external access to their own data only from the mediator as a trusted party; approximately 45.5% of these enterprises should not allow external access to their data due to data protection and 54.5% of the surveyed SMEs confirmed that the decision in this direction is dependent on the related costs and the relevance of data protection. Because the SMEs require different security levels.. For increasing the flexibility against the various user enterprises (SMEs), the security levels required by these user enterprises can be categorized into three categories:

End user enterprise from type A -with lower security level

End user enterprise from type B -with higher security level

End user enterprise from type C -with higher security level and/or aims to implementing external functionality as part of the internal ERP system. The user enterprises from this category use an SOA-based ERP system, which should be adapted for compatibility with the offered FERP-WSs. Lower security level means relatively low compared to the security level required by other user enterprises.

a) The Agreement Rules

To determine the FERP Mall roles, the first step in this phase is to describe the appropriate agreement rules between the FERP WS providers and the end user enterprises for all vendor or customer types, so that this mall (the mediator) will has the different Coordinate activities between the FERP WS providers (developers) and the customers (user enterprises) and it will control security against both. The agreements will be formulated in two sub-steps:

The first is service level agreement (SLA) between the providers and the mediator (FERP Mall) and the second through process level agreement (PLA) between the mediator and the end-user enterprises (for more information about this step see [6;7].

b) Transaction Phases-Scenarios

After formulating the basis for the level agreement between the actors, the second step of the adapting phase describes the possible processing scenarios of the transaction phases in the case of exchanging FERP systems with considering the different requirements of the user enterprises (SMEs).

Due to the digitizable properties of the goods offered by this business model, every transaction phase (information, negotiation and transaction phase) can be carried out online.

Processing the information and negotiation phases is the same for the three possible scenarios, but these scenarios differ in the transaction phase due to the different ways of using the required FERP-WS-Functionality

1) Information phase: FERP WS Publishing begins with the transmission of the ERP service information from the provider to the FERP Mall operator (mediator) and ended with contract signature. The FERP service information describes the static and dynamic properties of each operation of the FERP-WSs and determines the target values of dynamic properties as service-level indicators (SL-indicators). The contract defines the different conditions about the costs and the behaving way towards the customers from different types (A, B or C). The FERP Mall is the trusted party for representing the respective providers against the customer. In this phase, the mediator can check the offers by the appropriate methods.

2) Negotiating phase: The FERP mall operator (mediator) receives the demand from the customer (user enterprise) for a FERP process with the required quality (process level PL). Afterwards, the FERP Mall operator identifies the needed functionality for this process and the required quality and sends the contract including the appropriate providers list, the target values of the process levels and the costs for each provider to the costumer (user enterprise). This phase ends with the returning the contract signed by the customer. In this phase, the mediator will be informed to which type of customer (A, B or C) belongs the user enterprise.

3) Delivery and payment (transaction phase): The mediator behavior during this phase differs depending on the user enterprise type in terms of security levels (A, B and C), that will be described in the following three scenarios modeled with using the sequence diagrams:

Delivery and payment (transaction phase) for user enterprise with security level A

After the agreement between the FERP Mall operator as a representative of the FERP WS providers and the customer (user enterprises), the providers deliver the functionality by calling the FERP service operations by the customer. After the end of the FERP process, the customer returns to the mediator the information about the actual value of the PL and the mediator forwards a copy of that information to the providers, calculates the costs per (operation calls), taking into account the contractual penalty in case of non-compliance with the targeted values of the process levels and sends the resulting invoice to the customer. This phase ends with receiving the payment from the costumer. The Providers receive their shares from the mediator after deduction of the penalties (if required) due to because the providers will take the responsibility in case of non-compliance the contracted PLAs which will be identified by the mediator through monitoring and determining the compliance of the SLAs for each FERP-WS provider (see Figure 5).

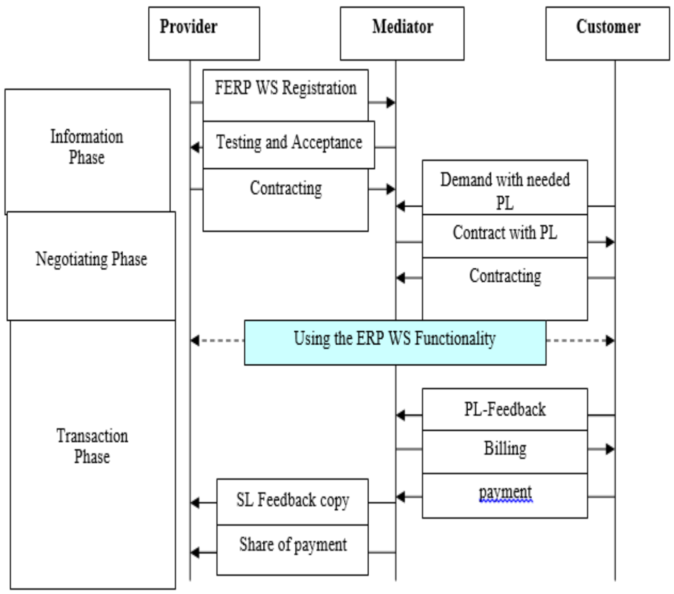


Fig. 5. FERP Mall role during transaction phases for user enterprise with security level A.

This scenario could be considered for small enterprises that don't need a high security level with considering the security solutions that has been proposed by Brehm, Marx Goemez and Rautenstrauch in [14, 22,15] But there are some enterprises that need more security because of their functionality sector (eg. banking sector). Such enterprises don't accept or allow to share their data with X providers. Therefore, in the next subsections there are two proposed scenarios for this type of enterprises

Delivery and payment (transaction phase) for user enterprise with security level type B

After the agreement between the FERP Mall operator as the representative of the provider and the customer from the type B, the FERP-WS providers will be informed about that. Then, the provider delivers the functionality structure to the mediator as a trusted party. The customer (user enterprise) uses the requested functionality by calling the FERP service operations. After the end of the FERP process, the customer returns the information about the PL's actual value to the mediator, which sends a copy of that information to the providers.

This phase ends after the payment of the invoice by the customer and each provider receives its portion of the invoice amount after deducting the resulting contractual penalties, as defined in the contract with the mediator in case of non-compliance of PLAs by monitoring and determining the compliance of the SLAs of each FERP-WS provider (see Figure 6). This scenario has the following disadvantages:

- Higher costs than scenario A because of the payment for the additional storage of the needed functionality structures by the mediator.
- If WS goes down, the replacement of another web service will take longer time than in case of scenario A

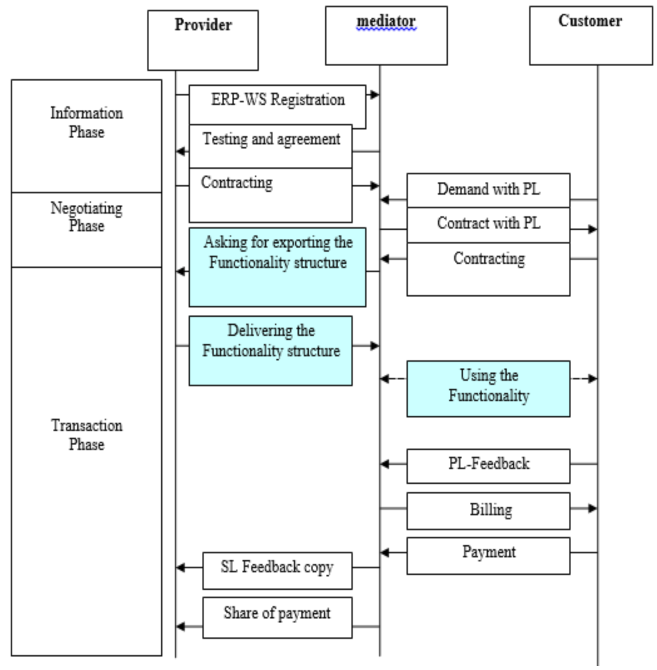


Fig. 6. FERP Mall role during transaction phases for user enterprise with security level B.

Delivery and payment (transaction phase) for user enterprise with security level type C

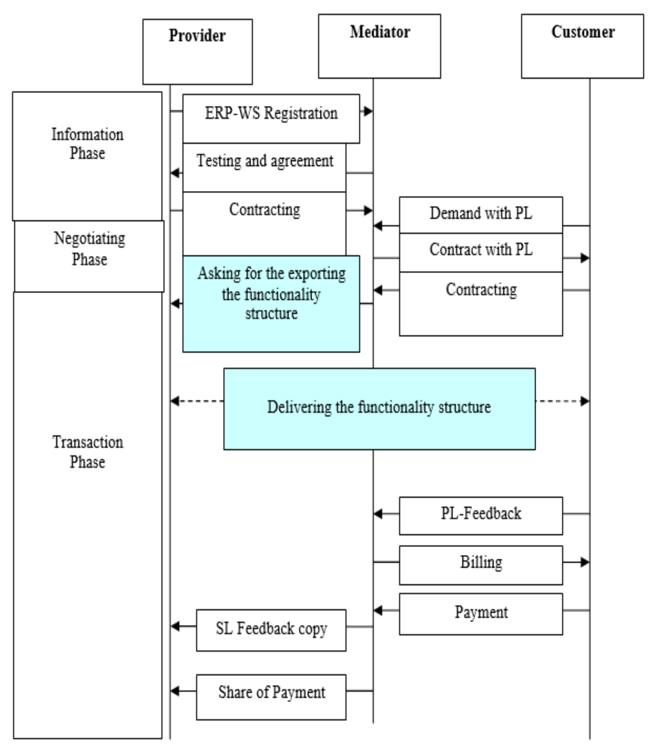


Fig. 7. FERP Mall role during transaction phases for user enterprise with security level C.

After the agreement between the FERP Mall operator and the user enterprise from type C, the FERP-WS providers will be informed about that. Then the providers deliver the functionality structure directly to the user enterprise. In this case, the user enterprise can use the functionality internally by calling the FERP service operations. After the end of the FERP process, the customer returns the information about the actual value of the PL to the mediator and the mediator forwards a copy of that information to the providers. The, calculates the total costs, taking into account the contractual penalty for non-compliance with the targeted values of the process levels and sends the resulting invoice to the customer. In this case, the FERP-WS functionalities will be licensed for internally use. This phase ends after the customer payment and the provider will receive his portion of the invoice after deduction of the contractual penalties in case of non-compliance of agreements (see Figure 7). The costs in this case are higher than the costs in the other scenarios (A and B), because the user enterprises should pay not per operation call, but by the using license and if WS goes down, the replacement of another web service will take longer time than in case of scenario A and B.

Then the end user-enterprise can choose a suitable alternative from these three scenarios depending on the comparison between the costs and the needed security from this enterprise. In the next section, we will describe the last step of adapting phase as the resulted Value flow scenarios.

c) Value Flow Scenarios

We will use use-case maps notations [28]for representing three different value flow networks due to the three types of costumers (end –user enterprise) as **results of the adapting phase depending on previous interactive concepts of transaction phases .**

- Value flow network of FERB business model in Scenario A:

The actors and their value creation activities are described as follows (see Figure 8):

The user enterprises represent the Market segment here, because these enterprises use the FERP systems and supporting services and pay only to the FERP-mall operator as single representative. The payment by the user enterprises will be distributed the supporting services (consulting and training) and the FERP processes (FERP Workflow Definition and FERP WS calls).

The FERP-mall operator is represented as composite actor, because he serves as (consultant, trainer, WSD environment and FERP Process Designer). The FERP-mall operator distributes the payments to the customers participating in the ERP process as providers of FERP-WS functionalities. The provider shares are billed depending on the FERP-WS-calls after the deduction of intermediation fees.

The ERP WS providers are represented as a market segment, because they develop the FERP-WS functionalities for different business sectors using the WSD environment offered by the FERP-mall operator and publish these WSs by this operator against certain fees. There is no direct money flows between the end user enterprises and the FERP-WS-providers because the payment for the WS calls go only through the intermediary (FERP-mall operator).

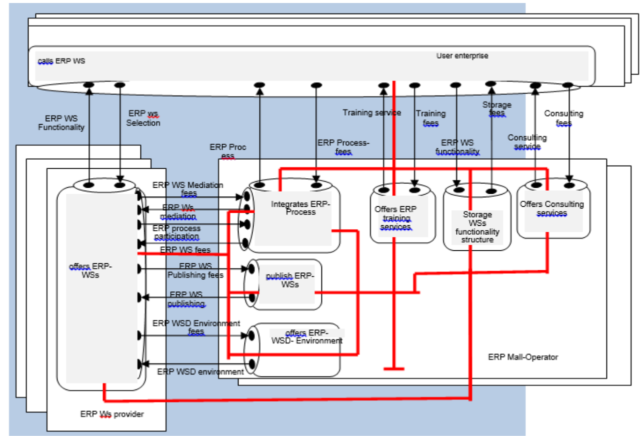


Fig. 8. Value Flows of FERP business model for customer type A.

- Value flow network of FERB business model in Scenario B:

The value flows of this scenario are derived from Scenario A by expanding the FERP-mall operator role for storing the functionality structures of FERP WSs, which are selected by the user enterprises (see Figure 9)..

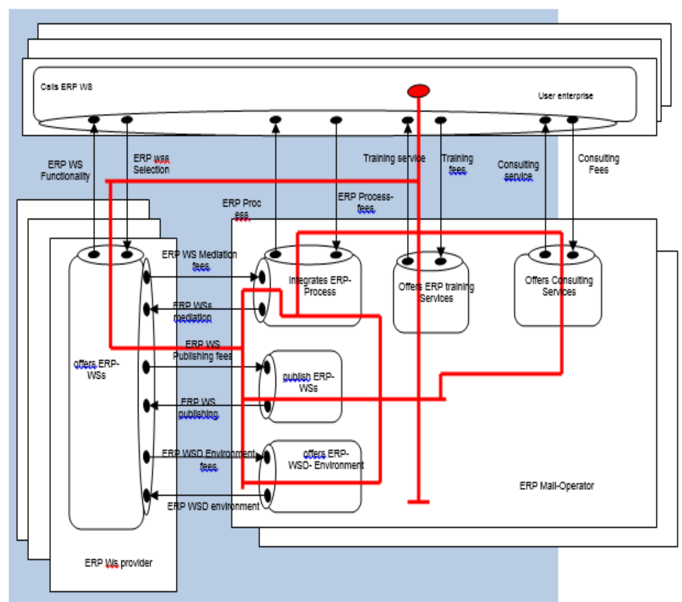


Fig. 9. Value Flows of FERP business model for customer type B.

The storage fee is paid from the use of WS-FERP functionality for increasing the safety for the users (customers).

In this scenario, there is no direct connection between the FERP-WS providers and the user (customers), since the FERP-WS functionalities are used by the mediator (FERP mall).

With the delivery of he WS functionality structures to the mediator (FERP mall), the FERP-WS providers expand their market shares through the users who require a relatively higher level of security .

- Value flow network of FERB business model in Scenario C:

The difference between this scenario and Scenario A is only by the delivery of the selected functionality structures of FERP WSS instead of WS-functionality between FERP-WS providers and user enterprises. The business model in the two scenarios (A and C) has a similar value flow network (see Figure 10). In the case of scenario C there is no direct cash flows between the operating enterprises (costumers) and the FERP-WS-party because the payment well be payed for FERP-WS functionality licensing by the user enterprises only through the FERP mall

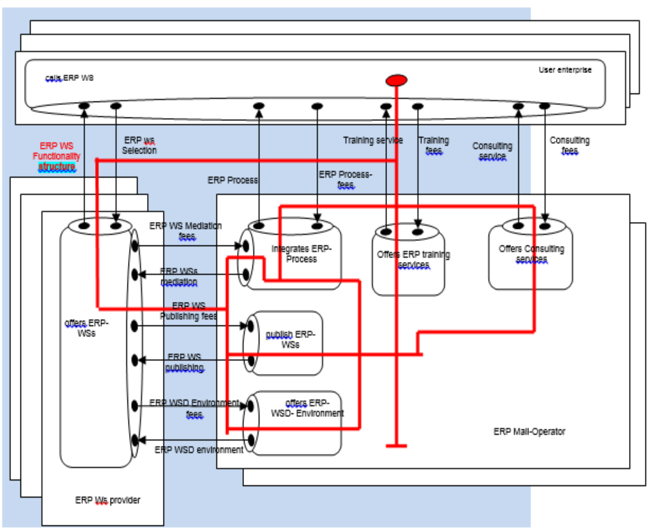


Fig. 10. Value Flows of FERP business model for customer type C.

C. Goal Phase

The last phase of the developing approach is determining the revenue model of the business model **as result concept depending on the characterization and adapting steps**. The goal of any business model is the generation of revenues for increasing the efficiency of the model. The figure 11 shows the concept of possible revenue model due to the determination of the offers and the roles of FERP mall. As the figure 11 shows the revenue will be generated through:

- selling FERP-WSs, the WSD environment (as services), consulting and training services = (Direct and transaction-dependent revenue)
- selling WSs because of FERP-WS-market is open to everyone FERP-WS developer = (Direct and transaction-dependent revenue)
- FERP-WS-publication (WS- advertising) = (indirect and transaction-independent revenue)
- Intermediation of the data storage and transaction phases between FERP-WS providers (developers) and the different types of customers (corporate users) = (Indirect and transaction-dependent revenue)

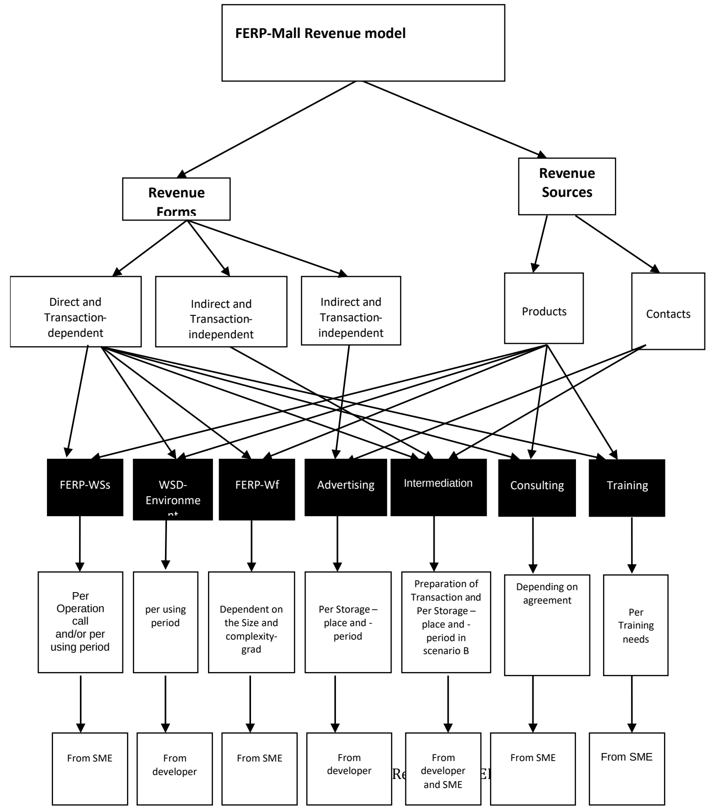


Fig. 11. Revenue of FERP Mall.

III. Qualitative Evaluation

This evaluation focuses on the classification of the various added values that arise from the perspective of the participating actors in the FERP business model. This informational added values are divided into ten categories [36; 48] (please see table 4:

IV. Qualitative Evaluation By Interviewing A Target Group

For the evaluation of the acceptance of the FERP System and FERP Mall idea as a solution for covering the company's requirements for the ERP functionality, we did survey among a target group of these companies, which are working as service providers and / or as manufacturers, and Sellers of physical and digitizable products. For this survey, approximately 120 invitations were sent to the SMEs in Saxony-Anhalt in Germany by fax and e-mail. The invitation included a brief summary of the FERP System and FERP Mall ideas. 33 of the invited companies accepted to participate in the survey. The contacts with the surveyed companies were carried out as follows:

- 73% personally
- 27% by e-mail

TABLE IV. INFORMATIONAL ADDED VALUE

End user enterprise (customer)	**Mediator**	**FERP-WSs provider**
Comparative Added value		
1- The user enterprise may test the generated FERP process by FERP mall before the life time u	1- The FERP mall can test the publishing.	
Integrative Added value		
	1- The FERP Mall offers reusable FERP processes which integrate FERB WSs of various developer	
Organizational Added value		
	1- In the case of ERP systems the FERP mall appears as a virtual enterprise for offering the FERP workflow definitions, the online directory services and support services	1- In the case of the ERP systems. The FERP-WS- developer appears as a enterprise, which offer FERP-WSs according to an respective competence.
Strategic added value		
	1- The FERP mall can provide worldwide market transparency by publishing FERP WSs with their marketing-information, the reviews and the safety rules.	1- The FERP-WS provider can provide more Market transparency by publishing his FERP-WSs and its evaluations.
Innovative Added value		
1- The interactive relationship between the intermediary and the developers helps for offering an adapted FERP WSs, FERP workflows and support services to the individual needs of the user enterprises.	1- Offering ERP functionalities as WSs will open the door for providing new products and services for Supporting FERP customers and FERP-WS developers (e.g. FERP workflow definitions and other services)	1- The provider offers new form of the ERP functionality as WS of which can be combined with other WSs providers with other WSs.
Added values with efficiency effects		
1- Reducing the costs, because the user enterprises pay only for the used functions (pay as you go). 2- ~~reducing~~Reducing the costs through competition between FERP-WS providers of similar FERP-WSs. 3- reducing the cost, because in the case of FERP system less end hardware are needed. 4- The user enterprises can achieve cost advantages through purchasing FERP-WSs with other support services in one package directly from the mediator. 5- The user enterprises can reduce time and cost through faster replacement in case of failure of a Web service. 6- The user enterprises can have the needed FERP System faster because of the online availability of ERP components and the faster adaptability of FERP processes	1- The mediator can adapt in less time the needed FERP workflow by the user-enterprise ERP processes. 2- The mediator generates revenue through the mediation of FERP-WSs, offering FERP workflow definitions and the supporting services.	1- The FERP-WS provider (developer) can reduce development time by using the WSD environment, which is offered by the mediator. 2-The REUSE-applicability of FERP-WSs increases the efficiency because the variable costs are near to zero
Added values with efficiency effect		
1 Selection of the needed ERP functionalities is facilitated by the offered information about the FERP-WSs from the different providers.	1- The FERP mall can increase customer satisfaction because of the safety levels as well as through the supporting services, thereby improving their market shares.	
Flexible Added value		
	1-FERP processes can be cheaper and faster adapted by the user enterprise	1-The online availability of the ERP component shortens the delivery time.
Aesthetic Emotional added value		
Macroeconomic added value		
1- changing the required qualifications of the employees by the user enterprises. 2- changing the IT infrastructure of the user enterprises.		1- A possible change the developing and offering form of information systems in the market.

During the execution of this survey, various problems and difficulties were faced. The most relevant difficulties were:

- Difficulties by gathering the addresses of the surveyed target group due to data protection laws,
- Difficulty in understanding the ideas of the FERP systems and the FERP Mall by the employees of the surveyed companies. Therefore, this questionnaire has been explained by a short presentation.
- The results of the survey are shown below:

1) The surveyed companies has been chosen from different size (number of employees). The ratios of company size are listed in Table 5 in three categories

TABLE V. PROPORTION OF COMPANIES SURVEYED DEPENDING ON THE NUMBER OF EMPLOYEES

number of employees	**1 - 60**	**61- 150**	**> 150**
Proportions	6.63 %	27,.3 %	9,.1%

2) All surveyed companies use more than one information system. 85% of them use 2 to 5 systems that provide functions to cover the different sectors (see table 6).

TABLE VI. FUNCTIONALITY OF THE INFORMATION SYSTEMS USED IN THE SMES

Functionally Sectors	**proportions**
Finance and accounting	63.6%
Customer management	63.6%
Distribution	63.6%
Payroll	54.5%
Cost accounting	45.4%
Procurement	27.3%
warehousing	36.4%
Human resource management	54.5%
Production Planning and Control	36.4%
project management	81.8%
quality management	63.6%
Research and Development	9.1%

- 82% of the surveyed companies have confirmed that the idea of a FERP mall is a good idea. However, 22% of these had additional comments, like

- Whether this idea is also suitable for very small companies,
- And whether the user company can test the new replacement solution.

3) Table 7 explains the answers to the question of the responsibility for defining the appropriate workflows among the user companies. From this table it can be deduced that a favorable definition of the workflow of the user companies can take a market share (about 9.1% + 36.4% = 45.5%) for offering such products (see table 7).

TABLE VII. POSSIBILITIES OF ADOPTING WORKFLOW DEFINITIONS AMONG THE SURVEYED SMES

A definition of the workflow can be created by employees within the company and they take the responsibility for it.	63.6%
A definition of workflow can be made by the FERP Mall itself.	9.1%
That depends on the cost.	36.4%

4) About 55% of the companies have confirmed that training their employees is necessary to introduce a FERP system. 25% of these companies consider the request for such training possible or useful. The estimated market share of these services is about 55% + 27% = 82%.

5) About 73% of the companies have confirmed that consulting is needed in the case of FERP system. 18% of these companies consider the request for such consulting possible and useful. The estimated market share of these services is about 73% + 18% = 91%.

6) Table 8 describes the possible storage location of the company data in case of FERP system from the point of view of the surveyed user companies (SMEs): From this table it can be deduced that the possibility of storing the data by the FERP Mall as a trusted party which Controls costs and data security can be increased.

TABLE VIII. LOCATION OF COMPANY DATA FROM SMES POINT OF VIEW

Storing Place	**Proportions**
By FERP Mall as a trusted party	18.2%
by the user company itself	45.4%
This depends on the associated costs and the relevance of data protection.	54.5%

7) Possible problems in case of FERP system from the interviewed companies point of view are:

- Problems due to the required data security.
- The existing platform in some companies is small and not very suitable for using such system.
- The extra effort duiring the shifting to such system and the necessary training.
- Some of the user companies are in close contact with the suppliers. Therefore, this supplier should also be able to adapt to such systems.
- SMEs are often very specialized - so the software has to be flexible enough to be adapted for " peculiarities " of the SME. z. For example, there are sometimes very significant differences in the interfaces of SMEs belonging to a similar area.

Most of these issues have been taken into account by the processing of the technical and / or economical part of the FERP idea. Regarding to some cases of close connection between the user companies and the suppliers, the suppliers can adapt to the new systems in order to satisfy their customers (user companies).

V. Discussion

The most of existing articles in Business model field aimed mostly to increase knowledge about the theoretical definition and classification aspects and of an business model and its components (see [41; 50, 34. 33, 53]) identified strategic framework for managing business model innovation (BMI) but they didn't identify clearly logical steps that explain: how we can realize suitable business model for innovative business Ideas. It is very important to refer in this context that the previous researches focused on the business models from management side in general without consideration of specific business area especially that are related to the information technology (like as a service paradigm) which need specific background and analysis. Some of the researches that focused on cloud computing as a service paradigm (like [29]) provided business concept for providing Modeling & Simulation (M&S) as a service without to formulize the analytical steps that lead to these concept.

The gap between the aim of this research and the other related research works is that this research focused on deriving and formulizing a suitable developing approach, which provides logical steps that lead significantly to realize the appropriate business model concepts in general and for distributed business information system. The provided approach in this research was as result of business and technical analysis to give framework for developing various type of business ideas with business and it perspectives because it considers a relevant version of ERP System to be provided as a service.

VI. Conclusion

This work presents three-phase- developing approach for designing appropriate business model for exchanging federated ERP systems on basis of WSs. The results of this developing approach are three sub-models (as concepts). These sub-models are, the character, the role model and the revenue model, which together formulate a FERP mall business model. The main type of this business model is a hybrid form as commercial intermediary for covering the lack of pre-defined communication channels and areas of responsibility between the respective developers and the user-enterprise as well as an online shop by offering support services and Workflow definitions. This business model belongs either to content business model type by offering online training services. This Idea also has been quantities and qualitative evaluated.

References

[1] S. Abels, N. Brehm, A. Hahn and J. M. Gómez: Change management issues in federated ERP-systems" An approach for identifying requirements and possible solutions. International Journal of Information Systems and Change Management (IJISCM), 1, pp. 318-335. (2006).

[2] D. Abts, and W. Mülder: Grundkurs Wirtschaftsinformatik", Eine kompakte und praxisorientierte Einführung, 4. Aufl., Braunschweig et al. (2002).

[3] A. Afuah, Business Models: A Strategic Management Approach. New York: McGraw-Hill/(2003).

[4] B. Al-Sawaie and Z. Al-Salti,: `Enterprise Resource Planning (ERP) Adoption: An Exploratory Case Study from Oman`, 26th IBIMA International Conference proceedings, Spain (2015).

[5] E. Asfoura, N. Jamous and W, Salem: The economic classification of E-Learning business models, In: Institute of Electrical and Electronics Engineers: IEEE multidisciplinary engineering education magazine. - New York, NY: IEEE, Bd. 4, 1, S. 8-12 (2009).

[6] E. Asfoura, G. Kassem and R. Dumke:.Characterization of business model for federated ERP systems, in: International Journal of u- and e-Service, Science and Technology Vol. 3, No. 4, December, 2010, S. 19-3

[7] E. Asfoura, G. Kassam and R. Dumke: Kombination von Service, Process Level Agreement und Zahlungsverhalten für sichere FERP Malls , In: Workshop Bewertungs aspekte serviceorientierter Acrhitekturen. BSOA 2009; Darmstadt-Germany.

[8] E. Asfoura, N. Jamous, G. Kassem and R. Dumke, R: FERP mall role in FERP web services marketing In: International Conference on Information Society (i-Society) . - Informations Society,IEEE, ISBN 978-0-9564263-3-8, (London, UK)2010, S. 156-161

[9] E. Asfoura, N. Jamous, G. Kassam and R. Dumke: E-mall as solution for marketing the federated ERP components on the basis of web services, In: International review of business research papers . - Melbourne, Bd. 5. 4, S. 478-490 (2009).

[10] E. Asfoura, G. Kassem, N. Jamous and R. Dumke: Pricing-model for marketing of FERP-Workflow as product. In: fifth International Conference on Digital Information Management . - Piscataway, NJ : IEEE, ISBN 978-1-424-47571-1, S. 321-325Kongress: ICDIM; (Thunder Bay, Canada) 2010 .07.05-08.

[11] E. Asfoura, G. Kassem, and R. Dumke:.Characterization of business model for federated ERP systems, in: International Journal of u- and e-Service, Science and Technology Vol. 3, No. 4, December, 2010, S. 19-3

[12] E. Asfoura, G. Kassam and R. Dumke: Kombination von Service, Process Level Agreement und Zahlungsverhalten für sichere FERP Malls , In: Workshop Bewertungs aspekte serviceorientierter Acrhitekturen. BSOA 2009; Darmstadt-Germany.

[13] F. A. Boons and F. Ludeke-Freund,;. Business ¨ models for sustainable innovation: State-of-the-art and steps towards a research agenda. Journal of Cleaner Production, 45: 9–19 (2013).

[14] N. Brehm, J. M., Gómez and C. Rautenstrauch: " An Open Security Model for Web Service-based ERP Systems. ,, in: 4th Annual Security Conference, Las Vegas, USA, 2005.

[15] N. Brehm and J. M., Gómez : Securing Federated ERP system clients. GITMA World Conference - Global Information Technology Management Association, Orlando, Florida, USA, 2006.

[16] N. Brehm and J. M., Marx Gómez, J : Distribution of ERP system components and security considerations. 17th IRMA Intenational Conference - Managing Modern Organizations with Information Technology. Washington D.C., USA, 2006a.

[17] E. Asfoura, Abdel-Haq MS, and Chatti et al. (2018). Classification of business models with focusing on characterizing "as a service" offers. International Journal of Advanced and Applied Sciences, 5(11): 16-23

[18] N. Brehm and J. M. Gómez : The Web Service-based combination of data and logic integration on Federated ERP systems, 18th IRMA International Conference - Managing Modern Organizations with Information Technology, Vancouver, Canada (2007).

[19] N. Brehm and J. M. Gomez, :Federated ERP-Systems on the basis of Web Services and P2P networks, International Journal of Information Technology and Management (IJITM), pp- 75-89 (2010).

[20] N. Brehm and J. M Gómez: Web Service-basierte Referenzarchitektur für Föderierte ERP-Systeme. In: (Pietsch, T.; Lang, C.V. Hrsg.): Ressourcenmanagement. Erich Schmidt Verlag, Berlin, pp. 125-142.1875 (2007a).

[21] N. Brehm, J. M . Gómez. And H. Strack: Request-Response-Evaluation Infrastructure for trusted Web Service-based ERP systems, In: C. Rautenstrauch (Hrsg.), Die Zukunft der Anwendungssoftware - die Anwendungssoftware der Zukunft. Aachen: Shaker Verlag. pp. 83 – 93 (2007).

[22] N. Brehm and J. M. Gómez.: Sicherheitsprotokoll zur Bewertung von Diensten in SOA-basierten Anwendungssytemen, in: BSOA Workshop - Bewertungsaspekte Serviceorientierter Architekturen, Berlin (2006).

[23] N. Brehm, J. M. Gómez and A. Ziesenitz: Toolunterstützung bei der vermarktungs orientierten Entwicklung von Web Services als Bausteine

komplexer betrieblicher Anwendungssysteme, in: proceeding of Multikonferenz Wirtschaftsinformatik, München/Garching (2008).

[24] D.J Collins and K.P. Lam Selling the Cloud to Smaller Business Organisations", Journal of Cloud Computing , Vol. 2014 (2014), Article ID 586109, DOI: 10.5171/2014.586109

[25] C. R.; Cunha, E, P. Morais; J, P, Sousa and J, P, Gome: The Role of Cloud Computing in the Development of Information Systems for SMEs, in Journal of Cloud Computing http://www.ibimapublishing.com/journals/JCC/jcc.html Vol. 2017 (2017),

[26] D. Debelak: Business model made easy. Wisconsin: CWL Publishing Enterprises (2006)

[27] J, M. Gronau.: Enterprise Resource Planning und Supply Chain Management, Archite-ktur und Funktionen. München (2004).

[28] J. Gordijn, J. M. Akkermans and J. C. van Vliet: What's in an Elec-tronic Business Model. In: Proc Knowledge Engineering and Knowledge Manag-ement,12th Int'l Conf. (EKAW 2000), Springer Verlag,Berlin, 2000, vol. LNAI 1937, pp. 257–273.

[29] Kiss et al : BUSINESS MODELS FOR CLOUD COMPUTING: EXPERIENCES FROM DEVELOPING MODELING & SIMULATION AS A SERVICE APPLICATIONS IN INDUSTRY, in: Proceedings of the 2015 Winter Simulation Conference

[30] V. Kumur and G. Raheja : BUSINESS TO BUSINESS (B2B) AND BUSINESS TO CONSUMER (B2C) MANAGEMENT, in : International Journal of Computers & Technology, Volume 3 No. 3, Nov-Dec, (2012)

[31] R. Maleri: Grundlagen der Dienstleistungsproduktion, 3. vollst. ueberarb. Aufl., pp 50, Berlin (1994).

[32] J. Magretta: Why Business Models Matter. Harvard Business Review on Business Model Innovation. USA: HBR Publishing Corporation (2010).

[33] Malmström et al: Practicing business model management in new ventures, Journal of Business Models, Vol. 5, No. 1, pp. 1-13(2017)

[34] L. Massa, C. L. Tucci and A. Afuah: A Critical Assessment of Business Model Research, in: Academy of Management Annals Vol. 11, No. 1(2017)

[35] M. Merz: E-Commerce und E-Business Marktmodelle, Anwendungen und Technologien ,Kap. 3 dpunkt Verlag, Heidelberg (2001).

[36] C. Rautenstrauch and T. Schulze: Informatik für Wirtschaftswissenschaftler und Wirtschaftsinformatiker, Berlin (2003).

[37] W. Reim, V. Parida and D. Ortqvist. :. Product ¨ –Service Systems (PSS) business models and tactics: A systematic literature review. Journal of Cleaner Production, 97: 61–75 (2015.

[38] D., J. Robey, J Ross and M. Boudreau: Learning to implement enterprise system, An exploratory study of the dialectics of change. Journal of Management Information Systems, pp. 1746 (2002).

[39] C. Scheer and P. Loos: Internetbasierte Geschäftsmodelle ; Neue Möglichkeiten der Wertschöpfungsorganisation in der Internet-Ökonomie" (2001).

[40] S. Slávik: Komparatívna analýza podnikateľských modelov. Ekonomika a manažment, 11(3), 23-43 (2011).

[41] S. Slávik and R. Bednár: Analysis of Business Models. Journal of Competitiveness, 6 (4), 19-40 https://doi.org/10.7441/joc.2014.04.02 (2014).

[42] B. Skiera and A. Lambercht: Erlösmodelle im Internet, in Herrmann, A., Albers, (Hrsg.), pp. 817 (2000).

[43] D. Stelzer.: Digitale Güter und ihre Bedeutung in der Internet-Ökonomi (2000).

[44] S. Y. Stephen and H. G. An: Software Engineering Meets Services and Cloud Computing, in: Published by the IEEE Computer Society OCTOBER (2011)

[45] T. F.; Stillman, F. D. Fincham,; D.V. Kathleen; V. M. N. Lambert and C. Phillips: The material and immaterial in conflict: Spirituality reduces conspicuous consumption, in: Journal of Economic Psychology 33 (2012) 1–7

[46] P. Timmers: Business Models for Electronic Markets". in: Electronic Markets 8(2) (1998).

[47] K. Turowski, K.: Fachkomponenten: Komponentenbasierte betriebliche Anwendungssysteme," Aachen (2003).

[48] Turowski and K. Pousttchi: Mobile Commerce. In: Springer Verlag Berlin.

[49] Usability First, (2016). [online]. [Viewed 17.04.18]. Available from: http://www.usabilityfirst.com/about-usability/web-application-design/erp-systems/

[50] B. Verhoeven and L. W. Johnson,: Business Model Innovation Portfolio Strategy for Growth Under Product Market Configurations, Journal of Business Models, Vol. 5, No. 1, pp. 35-50 (2017)

[51] P. Wells: Economies of scale versus small is beautiful: A business model approach based on architecture, principles and components in the beer industry. Organization & Environment, 29: 36–52 (2016)

[52] . B. W.Wirtz.: electronic Business" ,-243, Wiesbaden (2001).

[53] W3C., 2003. Web service architecture , working Group note, im WWW unter http://www.w3.org/TR/ws-arch /wsa.pdf (15.03.(2008).

[54] R. Amit and Zott, Ch. (2001): Value Chain in E-Business. In: Strategic Management Journal, 22, S. 521.

[55] B. Heinrich and Leist, S. (2000): Bankenarchitekturen im Infor-mationszeitalter – Zur Rolle des Geschäftsmodells. In: Österle, H.; Winter, R. (Hrsg.): Business Engineering – Auf dem Weg zum Unternehmen des Informat-ionszeitalters. Springer-Verlag, Berlin Heidelberg New York, S. 141-165.

[56] Mercer Management Consulting (2002): Herausforderung profitables Wach-stum. Pressemitteilung vom 02.12.2002, http://www.mercermc.de/uploadmaterial/61.zip

Conceptual Modeling of Inventory Management Processes as a Thinging Machine

Sabah Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait

Nourah Al-Huwais
Information Technology Department
Kuwait National Petroleum Company
Kuwait

***Abstract*—A control model is typically classified into three forms: conceptual, mathematical and simulation (computer). This paper analyzes a conceptual modeling application with respect to an inventory management system. Today, most organizations utilize computer systems for inventory control that provide protection when interruptions or breakdowns occur within work processes. Modeling the inventory processes is an active area of research that utilizes many diagrammatic techniques, including data flow diagrams, Universal Modeling Language (UML) diagrams and Integration DEFinition (IDEF). We claim that current conceptual modeling frameworks lack uniform notions and have inability to appeal to designers and analysts. We propose modeling an inventory system as an abstract machine, called a Thinging Machine (TM), with five operations: creation, processing, receiving, releasing and transferring. The paper provides side-by-side contrasts of some existing examples of conceptual modeling methodologies that apply to TM. Additionally, TM is applied in a case study of an actual inventory system that uses IBM Maximo. The resulting conceptual depictions point to the viability of FM as a valuable tool for developing a high-level representation of inventory processes.**

***Keywords*—Conceptual model; diagrammatic representation; inventory control; inventory management; workflow; thinging**

I. Introduction

In general, inventory is defined as items stocked in a store to meet anticipated requests. Inventory management or control is a system to balance product needs with demand to minimize costs that arise from obtaining and holding inventory [1]. There are several schools of thought that view inventory and its functions differently. This paper presents a foundation for inventory processes modeling that views an inventory as an abstract machine, called a Thinging Machine (TM), with five operations that may include infrastructure of submachines. We claim that such modeling, which is based on TM, facilitates understanding and serves as a base for consequent phases of an inventory system's design and development. A model is understood as an abstract view of a portion of reality that enables developers to concentrate on relevant aspects of the system and discount needless complications [2].

Inventory management models are typically classified into three forms [3]:

- A conceptual model that contains text, pictures and diagrams to explain the terms and principles of a particular system's functioning.
- An analytical (purely mathematical) model that uses mathematical concepts and language and contains formulas for analysis and calculations.
- A simulation (computer) model that attempts to simulate an abstract model of a particular system.

The last two modeling techniques are concerned with minimizing the total cost of inventory based on a decision-making process considering the cost of holding the stock, placing an order or encountering a shortage (e.g., insufficient stock to meet demand). This paper is focused on *conceptual* modeling of inventory management systems.

Conceptual modeling pertains to identifying, analyzing and describing the essential concepts and constraints of a domain with the help of (diagrammatic) modeling language that is based on a small set of basic meta-concepts [4]. It helps in understanding and communicating among the stakeholders and serves as a base for consequent phases of a system's development [5]. It should reflect the reality of the organization and its operations; conceptual models are most valued in terms of completeness, faithfulness to the realization of the underlying real system, understandability and susceptibility analysis.

A. Inventory Management

In an inventory management system, several basic notions are recognized, including minimum and maximum stock level, safety and reorder points and timing. The basic function of a management system requires preserving items' quantities and maintains them in such a manner that they do not remain in stock for a long time, which would result in efforts waste (e.g., delays in production, work and maintenance). Inventory cost remains low when the correct quantities of products are ordered at the right price and time.

Control is established by fixing the minimum and maximum levels of stock. These levels are calculated based on historical data, the expected requirements and in view of the inventory's condition. Reaching the minimum level may result in stock running out; reordering occurs in such a way that items are received before the stock volume reaches the minimum level.

Minimum stock occurs when the stock falls below an established critical level at which the enterprise processes may be harmed. In such an event, a warning is sent to management to exert further efforts or extra resources to ensure that the situation is rectified. The maximum point is utilized to avoid any superfluous stocks that may result in halting the flow of items. Designated stock is maintained for safety considerations, especially for events such as a serious stoppage of operations or to maintain the reliability of supply. In most cases, it is equal to minimum stock. The reordering process ensures that items are not out of stock by taking into consideration current stock, lead time and receiving time before reaching a minimum level.

Whatever the adopted inventory system, an organization needs a conceptual description that describes its real-world domain and does not include any information technology aspects. It would be able to serve various levels of granularity and complexity, such as by serving as a guide for the subsequent information systems specification, analysis, design and validation.

B. Problem and Solution

This paper claims that current conceptual models of inventory management systems lack comprehensiveness and completeness. Additionally, a lack of conceptual representation of processes makes it more difficult for end users to understand an existing process or simulate a new one. Accordingly, a high-level conceptual language can contribute by filling some needs and acting as a foundation in this area of research. As a step in this direction, the paper applies the recently developed TM that is based on the *thinging machine* notion and presents a different conceptualization of such. This paper advances the inventory management processes in a holistic way by developing a framework that is sufficiently inclusive. Generally, TM provides a diagrammatic representation of the static, dynamic and control specifications at play in the inventory management processes.

To show the viability of the proposed methodology, we use TM in an actual case study of an inventory management system that is currently being implemented using IBM software without an explicitly documented conceptual description.

C. Appraoch

Apparently, it is very difficult to contrast the involved diagrammatic models because they are based, to a large extent, on factors such as understandability (that pertains to visualization and graph completeness). A straightforward way to accomplish that is to put different diagrammatic depictions side-by-side and judge them accordingly. Therefore, we will give a few examples of current techniques so that, at the end, we are able to observe and contrast different samples.

D. Sample Diagrams

Today, most companies utilize computer systems for inventory control in which withdrawals are recorded and the inventory balance is monitored. Orders are placed and the balance of stock is updated by the computer. Modeling inventory processes is an active area of research that uses many methods, including flow charts, data flow diagrams, Universal Modeling Language (UML) diagrams, role interaction diagrams, Gant charts and Integration DEFinition (IDEF) [6].

In this section, several samples of inventory-related diagrams are presented. The purpose is not intended to give a fair discussion of these examples; rather, the aim is to provide an awareness of the type and nature of conception and depiction upon which this method is built. The samples will also provide the opportunity to contrast the diagramming techniques after presenting TM diagrams of our case study.

Saraste [7] used the flowchart technique, which is “very flexible and easy to use” [7], as shown in Fig. 1. In the inventory control environment, the modeling used by Saraste [7] may not be suitable to model the system in a holistic way through developing a framework that is sufficiently inclusive. In certain situations, the entire enterprise processes may be harmed by local event (e.g., stock goes below an established critical level). Having knowledge of the entire view of the system—something that is missing in flowcharting—can help management employ efforts and/or resources to prevent an adverse effect on a production situation. Another flowchart-based representation is a sample of the Maximo diagram in an IBM Knowledge Center [8], as partially shown in Fig. 2. Flowcharts were the target of many criticisms regarding their value in design and education [9-10] that have nearly led to their elimination. Lately, they have also been revived in the form of a UML activity diagram. According to Storrle and Hausmann [11], “activity diagrams have always been poorly integrated, lacked expressiveness, and did not have an adequate semantics in UML.”

According to Patel [12], Entity Relationship (ER) modeling allows for the formation of high-level conceptual data models that can be used to form a graphical representation for design, as seen in the initial ER diagram in Fig. 3.

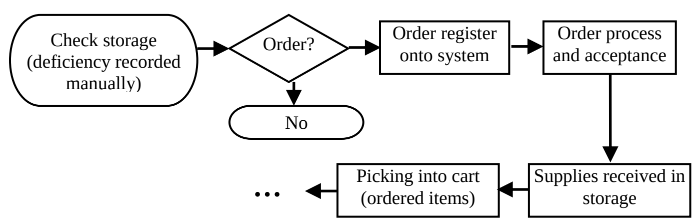


Fig. 1. Order Modeling using a Flowchart (Partially Redrawn from [7]).

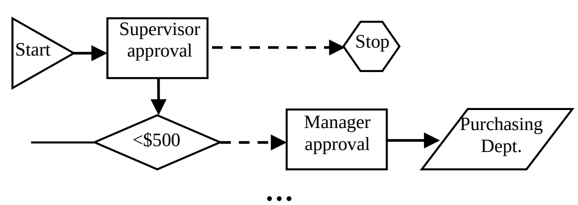


Fig. 2. IBM Knowledge Center diagram that Involves A Flowchart (Partially Redrawn from [8]).

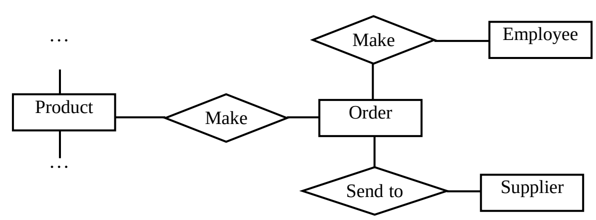


Fig. 3. ER Diagram (Partially Redrawn from [12]).

Nevertheless, difficulties in ER modeling (e.g., temporal aspects) are well known [13]. Rinardwiatma [14] used workflow process diagrams, such as the one shown in Fig. 4, to integrate inventory management in an IBM Maximo-based system.

Most existing conceptual modeling methods use object-oriented methodology that employs UML as a foundation that requires breaking the system structure and behavior into several types of diagrams, then further decomposing them into other diagrams. It is claimed that this approach has many advantages, such as simulating the modeler's way of thinking [15] and contributing to the reduction of complexity in the representation of technical systems and design processes [16]. For example, Tchantchane [17] utilized UML use case, class, state and sequence diagrams in designing a sales ordering system (as shown in Figs. 5-8).

Nevertheless, According to Mordecai [18], there is a "significant inability of common conceptual modeling frameworks to appeal to practicing designers and analysts." These diagrams are completely heterogeneous with several different conceptual bases. The purpose of this heterogeneity is to achieve a wide range of options for expression, depending on the situation. Shoval and Kabeli [19] suggested a merger of the data flow diagram, the ER diagram and object-oriented constructs. The multiplicity of diagrams for the same dilemma in UML is a known problem [20] that contrasts with providing a single, integrated diagrammatic representation that incorporates function, structure and behavior.

The next section introduces TM to be used both as a thinking style and as a vehicle for depicting the capturing inventory processes [21-26].

II. Thinging Machine (TM)

Reality consists of a range of things, such as an externally experienced object, situation, event or action, or a privately experienced sensation, mood, emotion, memory or thought. These things are the content of our *wajood* (existence). Some of these things comprise others or they form an environment or a place for others.

The thing tree takes the thing carbon dioxide from the thing air and gives it the thing oxygen. We say that the tree *receives* carbon dioxide from the air and *releases* and *transfers* oxygen to it. Also, the thing tree *processes* carbon dioxide to *create* oxygen. The tree in this case is a *machine* that releases, transfers, receives, processes and creates things. By the same conceptual manner, the thing human being is a machine that receives the thing oxygen and processes it to create the thing carbon dioxide that is released and transferred to the atmosphere. These things/machines are called TMs. The machine can be conceptualized as a *lived atmosphere* (i.e., environment/space/capsule) of its things where they are created, processed, received, released or transferred. We describe a TM diagrammatically as shown in Fig. 9. The figure is a generalization of the typical input-process-output model that is used in many scientific fields. The machine is constructed from the sub-machines of flows including the machine itself.

Martin Heidegger [27] describes viewing a particular tree as, in our model, a machine. It is rooted in the earth with its trunk rising up and branches splayed out, swaying in the wind. It is inhabited by many tiny creatures, and it responds to the wind currents. The tree is a certain compilation of the threads of life. It is a thing when we see it as a glimpse of life-in-formation, never the same from one moment to the next [28]. It is not only the tree that is a machine, but also people, animals, towns, the sun and clouds, as well as day, night, feelings, numbers, atoms, data, etc.

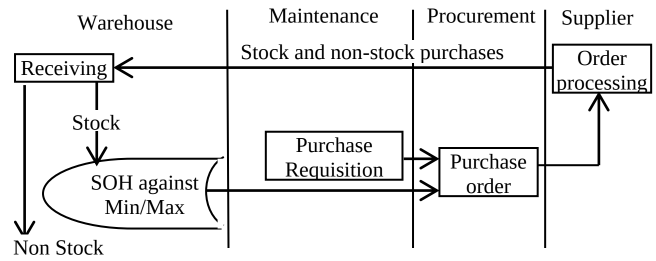


Fig. 4. Inventory Flow Process (Partially Redrawn from [14]).

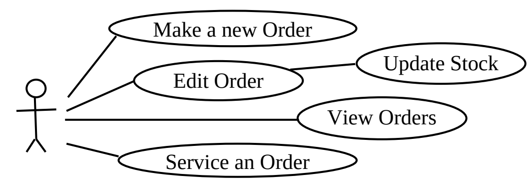


Fig. 5. Use Case Diagram (Partially Redrawn from [17]).

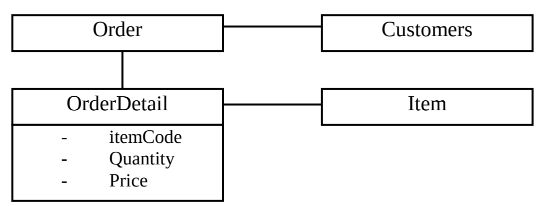


Fig. 6. Class Diagram (Partially Redrawn from [17]).

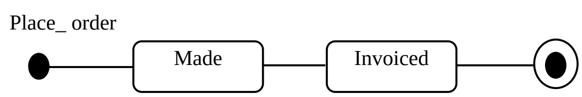


Fig. 7. State Diagram (Partially Redrawn from [17]).

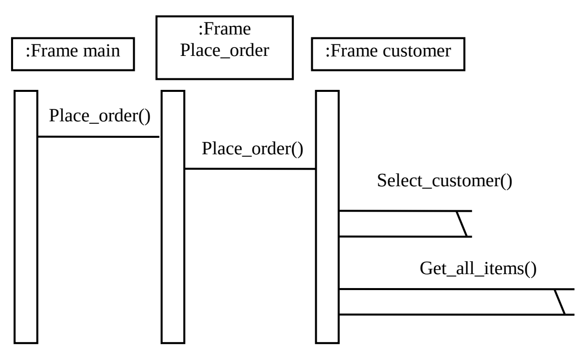


Fig. 8. Sequence Diagram (Partially Redrawn from [17]).

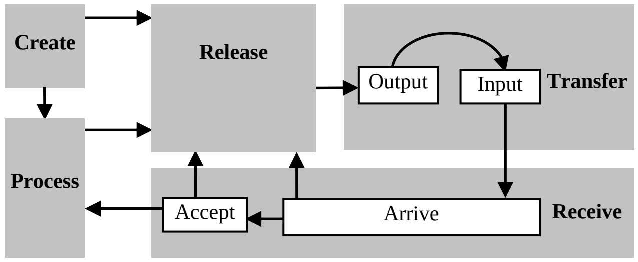


Fig. 9. Thinging Machine.

The (unproven) claim in TM modeling is that the five operations—create, process, release, transfer and receive—are sufficient to represent all activities in a machine. The only justification for this is the diverse modeling of many systems that appear in many publications, including the modeling inventories in this paper.

Things that flow in TM refer to the exclusive (i.e., there is one and only one) conceptual movement among the five operations (stages) shown in Fig. 10. It may be argued that *things* (e.g., goods) can also be *stored* in addition to being created, processed, released, transferred and received. However, because *stored* is not a generic operation, things can be stored after being created, hence becoming *stored created data*, or, after being processed, becoming *stored processed data* and so on. When all arriving things are accepted, then arrive and accept are represented by receive.

Create (emerge) is one starting point of a thing in a machine, in addition to being imported (transfer/receive) from the outside. Through creation and importing the system (machine) becomes aware (recognition) of a new thing in the machine. Additionally, a thing "disappears" from the "radar" of a machine, either when it is *de-created* (e.g., deleted) or by departing (release/transfer). Note that a thing can be released, but not transferred (e.g., finished goods waiting to be shipped when a truck arrives) or a thing being transferred (input), yet not arriving (e.g., an email arriving, but an error preventing the recipient from accessing it).

Process means that the machine changes the thing in a certain way. For example, a doctor machine processes a patient to decide how to treat him/her.

Each type of flow of things is distinguished and separated from other flows. No two streams of flow are mixed, just as lines of electricity and water are separated in buildings' blueprints. However, two types of things can enter a machine of a super type of thing (e.g., integers and real numbers flow to a number machine). A TM does not need to include all of the stages; for example, an archiving system might only use the stages transfer, receive, release and process (i.e., not create).

Multiple machines can interact with each other through flows or by triggering stages. Triggering is a transformation (denoted by a dashed arrow) from one flow to another (e.g., a flow of electricity triggers a flow of air).

III. Inventory in IBM Maximo

This section applies TM to model an actual inventory system. The system extends over physical and digital spheres, where it flows across different machines changing their forms. This paper assesses a case study that uses IBM Maximo, an asset management software system. In Maximo, nodes can represent various points in a business process (e.g., start node, condition node, interaction node, sub-process node, task node and stop node). A workflow process is created by interweaving nodes and connection lines within the workflow. There are many notions: person records, role records, action records, communication templates, notifications, escalations and action groups, etc. Also, there are many actions (e.g., create, change, incident, problem, service request and work order) [29].

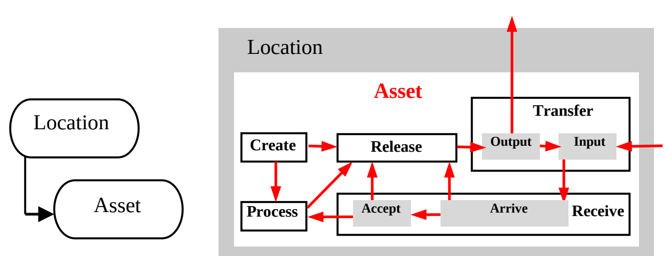
In our case study, Maximo modules are used to manage inventory, including functions such as controlling inventory, making purchases and tracking stock levels and contracts.

Inventory is a central module in Maximo. It functions in a dynamic relationship with the preventive maintenance, work orders, contracts, purchasing and assets modules. Maximo can automate processes that are repetitive or occur at regular intervals; for example, preventive maintenance, periodic inspections or reordering inventory items" [30].

In Maximo, inventory management is an important part of maintaining any asset for which the inventory module in Maximo tracks the required materials (e.g., monitoring reorder points, purchase requisitions and purchase orders; tracking the movement of items into and out of inventory; issuing work orders; etc.).

TM and Maximo have a completely different conceptual view. Maximo-based systems utilize different types of diagrams to document and describe various applications, including inventory management systems. Assuming knowledge of basic Maximo terms, consider a difference in terms of Maximo *location* and *asset* (e.g., a bank as the location). From the standpoint of IBM Maximo, a location is a physical place, an operating position where equipment resides [31]. It is a place where assets are operated, stored or repaired. Fig. 10 contrasts the conceptualization of a location and an asset. In TM, the location is a machine while the asset is a submachine and a thing that flows.

In TM, an inventory system is a machine that serves other machines by managing their input/output flows so that they are near steady states. Ideally, flows in its stewarded machines are never stopped or slowed due to a lack of item supply. The machine also ensures that the supply does not clog them with overstocked items.



(a) Maximo (b) TM

Fig. 10. Maximo and TM Representations of *Location* and *Asset* (The Left Part is a Partially Redrawn Part of a Figure [31]).

IV. Case Study

The TM representation is applied to the process of responding to an approved requested order arriving from its department to the commercial department. Accordingly, the commercial department checks the inventory for the request. Three possible cases are based on the inventory status: the requested items are available, partially available or not available. These cases can be modeled using a TM to produce TMs that involve the inventory system (as shown in Fig. 11).

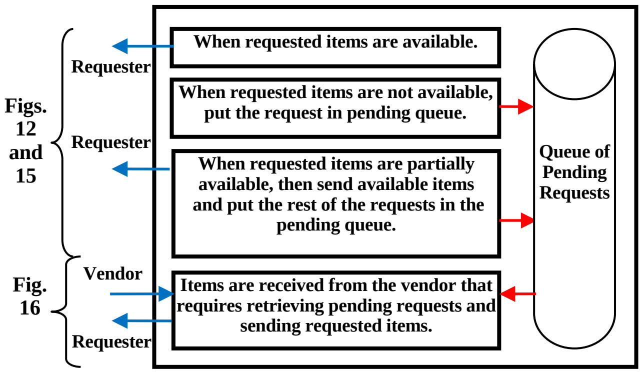


Fig. 11. Different Portions of the Inventory System.

A. A Request Arrives and Current Inventory is Checked

Fig. 12 describes the general process of this case study. It is expanded into three cases that are explained in the following sub-sections. As shown in the figure, an approved request of an item (circle 1) from the requesting department (2) flows (3) to the commercial department system (4) to check the inventory control. The commercial department system, in this case study, is responsible for managing the inventory and ordering items.

Once the approved request is received, the commercial department triggers (5) a check of the inventory status (6) in the inventory machine (7) to examine the current stock (8) of available items that pertain to the request, which is a global variable and is initially set to zero. The current stock is processed and compared to the minimum level of the inventory (9). The minimum level is considered a point of emergency (i.e., when a certain item volume reaches a level that is considered below critical). The availability cases based on this comparison are:

- If the current stock of the request is above the minimum level: The request is processed (11) to extract (12) the requested quantity that flows to a decision machine (13), which also receives the current stock value (14). In the decision machine, the new stock is calculated as the current stock minus the requested quantity (15). The new stock is considered a global variable and is initially set to zero.
- Accordingly

(i). If the new stock is equal or above minimum (16), meaning that there are enough items to be delivered, then the decision machine triggers the release of the requested items (17) to be directly delivered to the requesting department and the replacement of the current stock by the new stock (18).

(ii). If the new stock is less than the minimum (19), then a partial response to the request is possible:

- The available quantity of items is calculated (20) as (the current stock minus the minimum) and released (21) to the requesting department (assuming it accepts that).
- The pending quantity is calculated as the requested quantity minus the available quantity (22). The request is processed (23) to make it an inquiry for pending quantity and is sent to the queue system (24).

- **If the current stock of the request is equal to, or less than, the minimum level**: The request is sent to the queue system (26), where it is processed to increment the number of queued requests (27) and to add the request to the queue (28).

B. Modeling Events

Fig. 11 in the previous sub-section reflects a static structure of distributing orders among the three cases. To model the dynamic behavior of the case when a request arrives and the current inventory is processed, we need the notion of a machine *being-in-time*. In being-in-time, the machine not only creates, processes, receives, releases and/or transfers, but it also *machines* (habitually does these operations again and again). Time in TM is a *thing* that can be created, processed, released, transferred and received. Consider *the request has arrived and received in the commercial department*, which can be represented as shown in Fig. 13.

An event is a machine in a TM that contains at least three submachines: the time, the region and the event itself. The region is where the event *takes place*. The event in Fig. 13 includes the three machines: the region of the event (circle 1), which is a subdiagram of Fig. 12; the (real) time submachine (2); and the event submachine itself (3/green). It was previously stated in section two that the machine is constructed from the submachines of flows, including the *machine itself*. The machine itself is distinct from all of the submachines within. For simplicity's sake, an event will be represented only by its region.

Accordingly, we can identify the events in Fig. 14 as follows.

Event 1 (E_1): A request is received.

Event 2 (E_2): The inventory status is checked.

Event 3 (E_3): The current stock exceeds the minimum.

Event 4 (E_4): The request is sent to the decision procedure.

Event 5 (E_5): The new stock is calculated.

Event 6 (E_6): The new stock is = or > than the minimum.

Event 7 (E_7): Items are sent to the requester.

Event 8 (E_8): The new stock is less than the minimum.

Event 9 (E_9): The available and pending quantities are calculated, and the request is modified and sent to the queue system.

Event 10 (E_{10}): The quantity (circle 25 in Fig. 12) is equal to or less than the minimum.

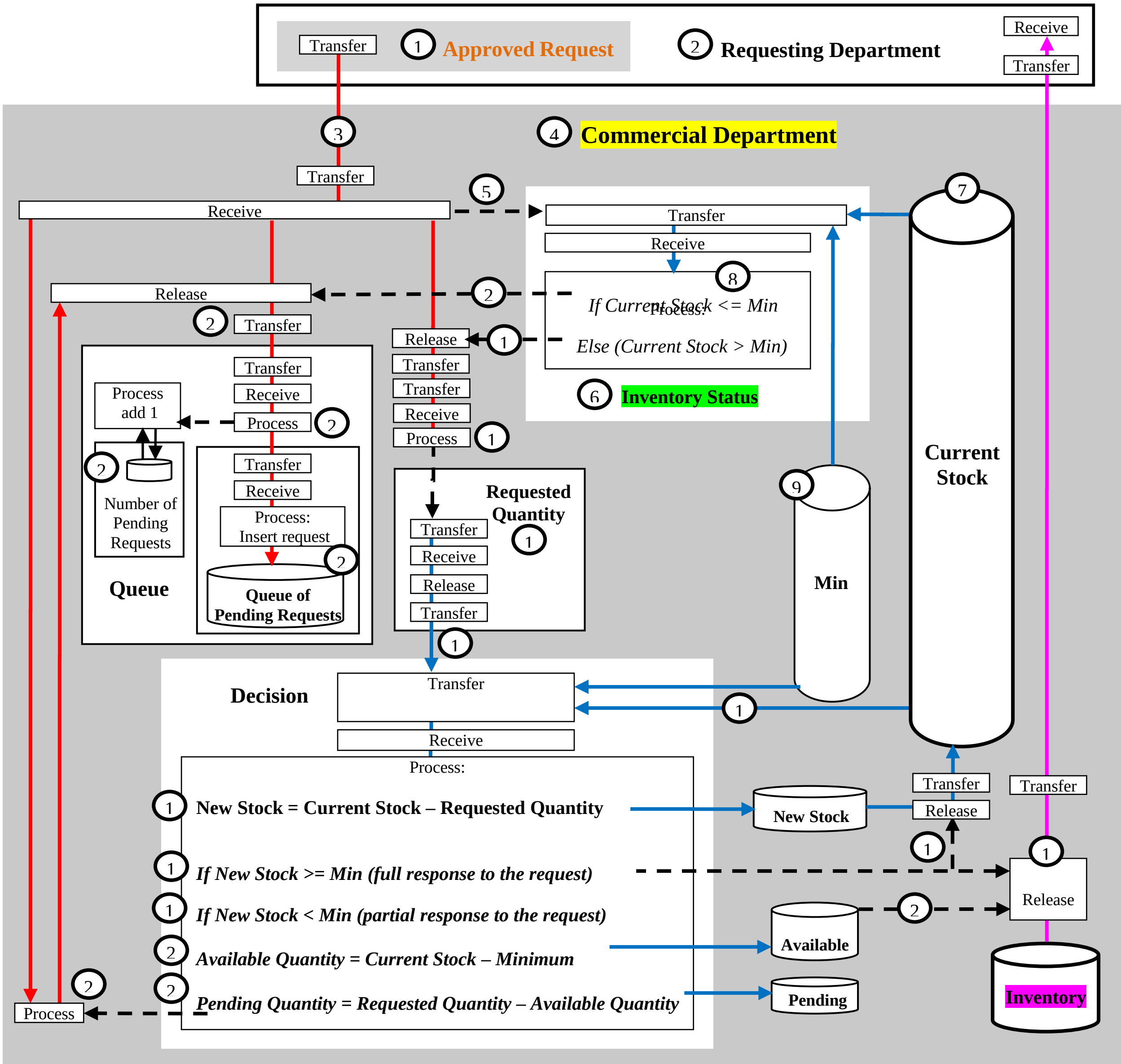


Fig. 12. General TM Representation of the Inventory Control Case Study.

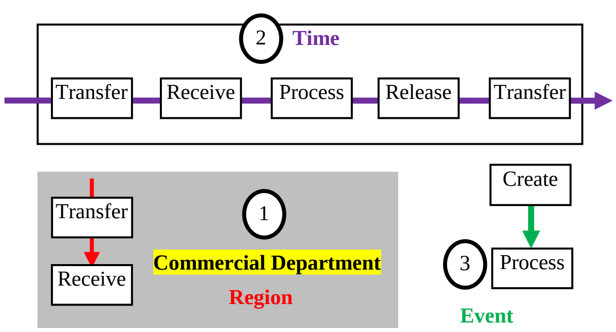


Fig. 13. The Event: The Request has Arrived and Received in the Commercial Department.

Event 11 (E11): The request is added to the queue system.

Fig. 15 shows the chronology of these events. It can be used in the execution and control (event operations) of events as exemplified in the figure. While a machine machines, control is an awareness of this machining that creates second-level machining.

C. If New Stock for the Request Is Below Minimum

Fig. 16 models the machine in which the new calculated stock is less than the minimum (circle 19 in Fig. 12). In this case, as shown in Fig. 16, the quantity of items available is insufficient to be delivered to the requestor. Therefore, the request is partially satisfied by sending the available stock (the current stock minus the minimum) to the requestor (circles 1 and 2). Moreover, the updated current stock (3), which is now equal to the minimum, and the number of pending items (the requested quantity minus available quantity (4) are sent to the supervisor to decide upon making new supply order (5).

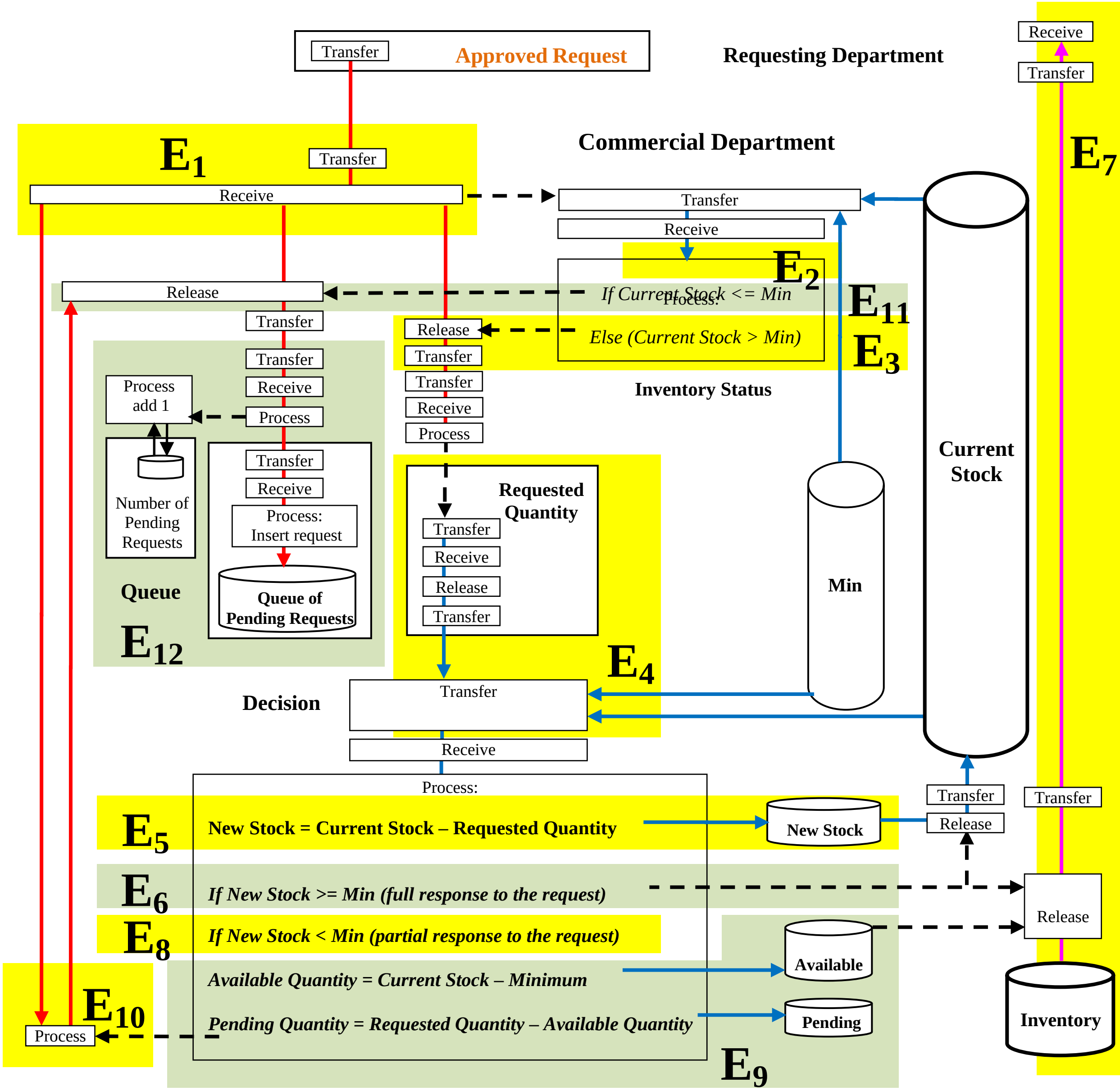


Fig. 14. Events of the FM Representation of the Inventory Control Case Study.

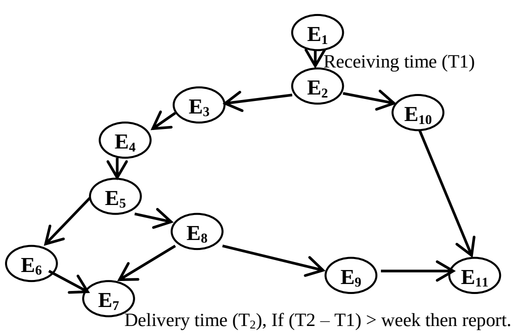


Fig. 15. The Chronology of Events of TM Representation of the Inventory Control Case.

The supervisor applies the company's inventory policy (6), which is based on statistics that pertain to the ordering level, the minimum level and the maximum level. The maximum level is set according to the average of historical data to maintain the number of items typically on hand in recent years. These levels are determined by the commercial manager.

The supervisor (7) decides whether to issue a request for quotation (RFQ) (8). Moreover, the supervisor assigns (8) an employee from the commercial department as the declared buyer (9 –bottom right of the figure) who is tasked with the responsibility of the transaction for this RFQ. The RFQ flows to the team leader (10) to be further processed.

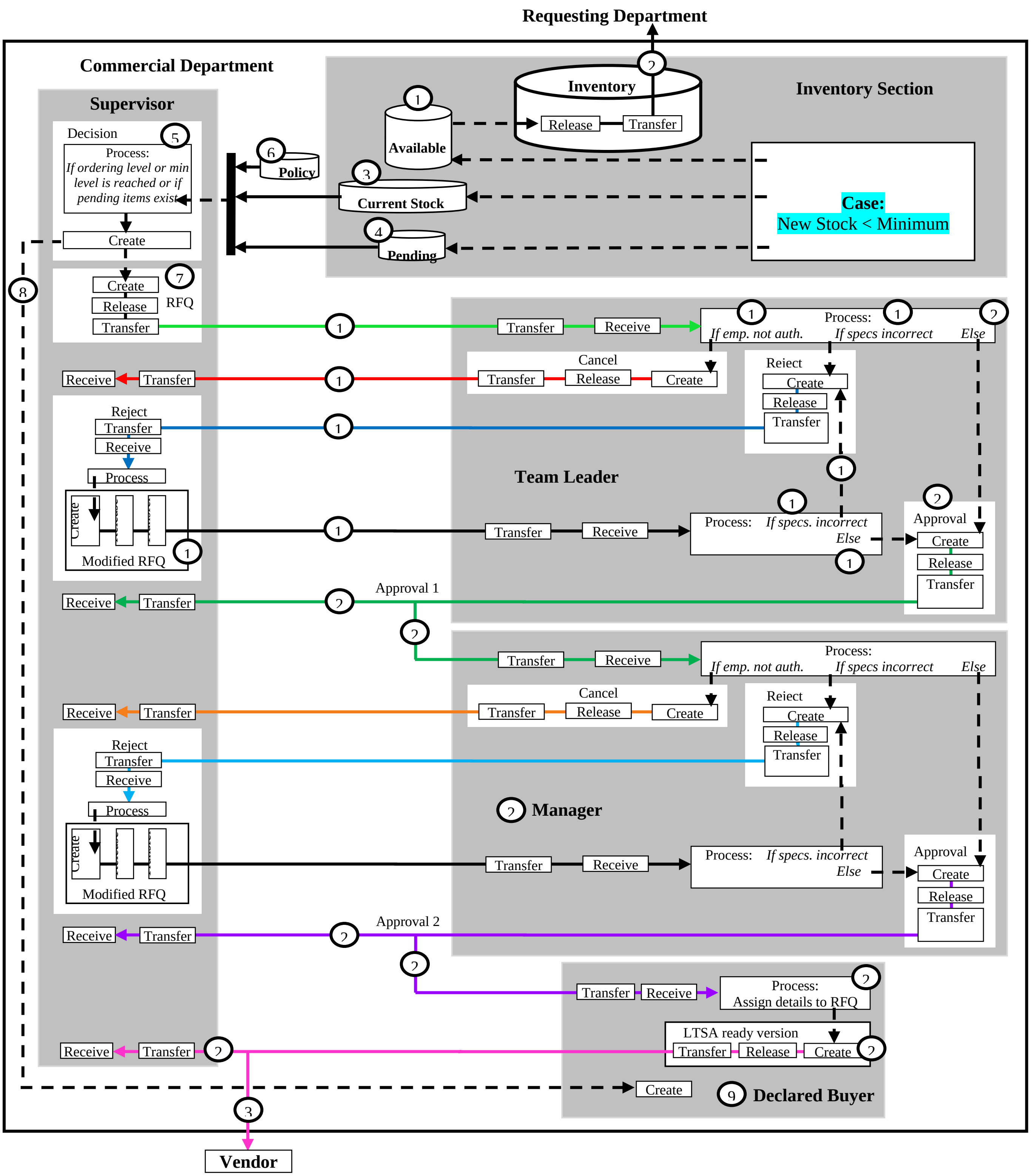


Fig. 16. TM Representation of the Case if the New Stock for the Request is Below the Minimum.

- If the requesting supervisor does not have the authority for such a type of RFQ (11), then it is canceled (12) by the team leader and a cancellation note is created and sent to the supervisor.
- If the specification of the RFQ is incorrect (13), then it is rejected and a rejection note is sent (14) to the supervisor in order to modify the RFQ specifications. Once the RFQ is modified (15), it is sent back (16) to the team leader.

If the specifications are still incorrect (17), then another rejection process is repeated (18). Otherwise, it is approved (19).

- If the RFQ is neither cancelled nor rejected (20), then it is approved (21) by the team leader. Approvals flow to the supervisor (22 - copy) and to the manager (23) for further processing.
- The same cycle of the team leader's actions is repeated for the manager (24) and its description is omitted here. Assuming that the manager approves the RFQ, the approval is sent to the supervisor (25 – copy) and to the declared buyer (26) who was specified by the supervisor.
- The declared buyer is responsible for assigning more details in the RFQ (27), which triggers the creation (28) of a long-term supply agreement (LTSA) (28). The LTSA does not need a bidding process because it is agreed upon with a specific vendor as a single source. It is called long-term because the contract with the vendor states that the price of the requested items shall be fixed for a specific number of years. The LTSA is sent to the supervisor (29 - copy), then the specified vendor (30). The vendor creates his own cycle of preparing and shipping the ordered items according to a specified time limit.

D. TM Representation of Receiving the Ordered Items from the Vendor

Fig. 17 shows the TM representation after receiving the ordered items from the vendor. Once the vendor (1) delivers the items (2), the current stock is updated (3). The pending requests (4) in the queue are released (5), one by one, to be processed to extract the requested quantity (6). Each request is processed to determine its quantity. Furthermore, the quantity is processed (7) to release the corresponding number of items in the queue (8). Additionally, the quantity flows to update the current stock (9) and the total pending items (10). Fig. 18 shows the events after receiving the requested items from the vendor. Fig. 19 displays the control over this sequence.

Note that the loop of events for every request in the queue is represented as a second order control over the events required for each request.

The thick horizontal bar at the bottom of Fig. 17 indicates the possible parallelism of E_5, E_6 and E_7. All of these events should end before starting another round of the loop.

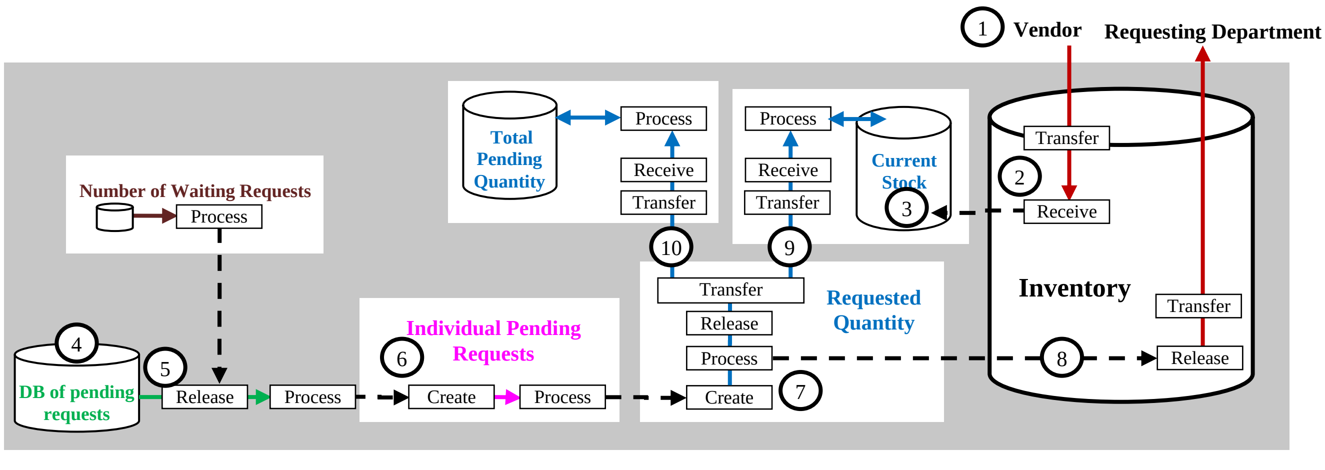


Fig. 17. TM Representation of Receiving the Requested Items from the Vendor.

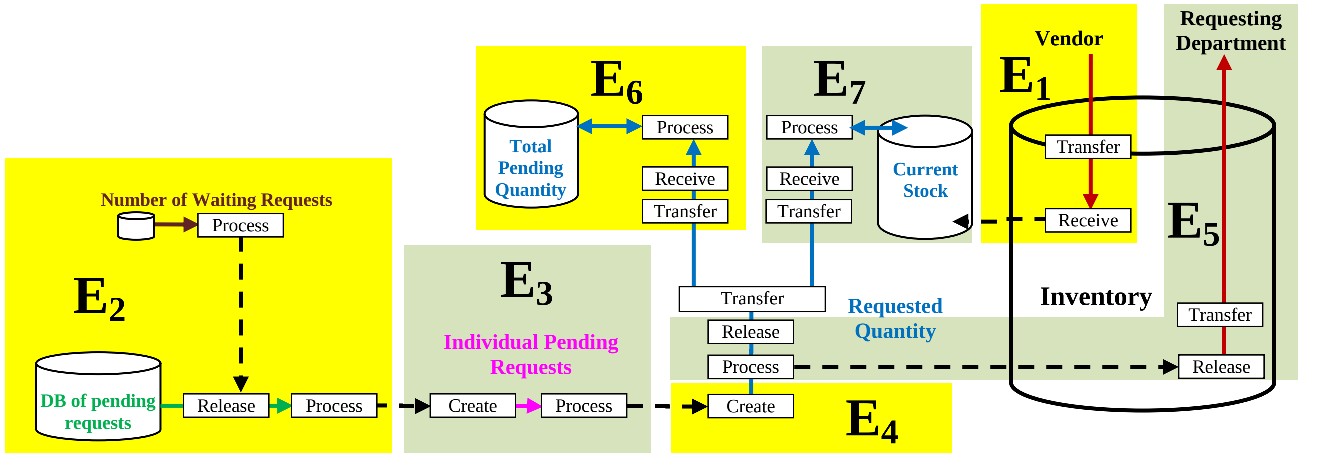


Fig. 18. The Events after Receiving the Requested Items from the Vendor.

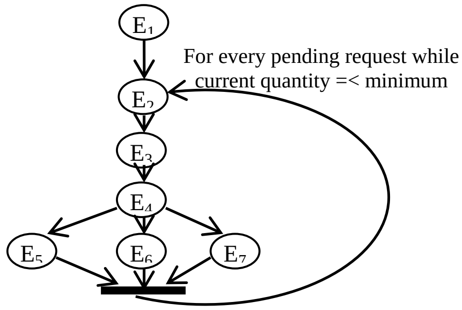


Fig. 19. The Events and their Control after Receiving the Requested Items from the Vendor.

V. Conclusion

This paper demonstrates the viability of TM modeling for inventory management processes. The resultant conceptual model covers the static, dynamic and control of these processes. This feature of notions' uniformity, based on the simplicity of the TM with its five stages, sets the modeling methodology apart from the heterogeneous diagrammatic representations (e.g., UML) that were displayed in section two of the paper. Further research will establish different potential benefits of the TM approach.

References

[1] Encyclopedia of Management, INVENTORY MANAGEMENT, Advameg, Inc., 2018. http://www.referenceforbusiness.com/management/Int-Loc/Inventory-Management.html

[2] J. Mukerji and J. Miller, "MDA guide version 1.0.1.," OMG, 2003. http://www.omg.org/docs/omg/03-06-01.pdf.

[3] A. Muravjovs, "Inventory control system analysis using different simulation modelling paradigms," Doctoral thesis, Transport and Telecommunication Institute, 2015. https://www.tsi.lv/sites/default/files/editor/science/PhD/Muravjovs/muravjovs_promotion_full.pdf

[4] G. Guizzardi, H. Herre, amd G. Wagner, "Towards ontological foundations for UML conceptual models," in R. Meersman and Z. Tari (eds) On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE. OTM 2002. Lecture Notes in Computer Science, vol 2519. Berlin, Heidelberg: Springer, 2002.

[5] I. Hadar and P. Soffer, "Variations in conceptual modeling: Classification and ontological analysis," in Journal of the Association for Information Systems, vol. 7, no. 8, August 2006, pp. 569-593.

[6] "Knowledge based systems, Inc.," http://www.idef.com/IDEF0.htm. IDEF Integrated DEFinition Methods. http://www.idef.com/pdf/idef0.pdf, 2010.

[7] S. Saraste, "A framework for evaluating inventory management in healthcare case: HUS Logistics," Master's thesis, Department of Information and Service Economy, Aalto University, School of Business, 2013.

[8] IBM Knowledge Center, "Example of a purchase requisition business process," Maximo Asset Management 7.6.0. https://www.ibm.com/support/knowledgecenter/en/SSLKT6_7.6.0/com.ibm.mt.doc/gp_wkflow/c_sample_purchase.html

[9] T. Brooks, "The mythical man-month: Essays of software engineering," Reading, MA: Addison Wesley, 1975.

[10] B. Shneiderman, R. Mayer, D. McKay, and P. Heller, "Experimental investigations of the utility of detailed flowcharts in programming." Commun. ACM, vol. 20, no. 6, pp. 373-381, 1977.

[11] H. Storrle and J. H. Hausmann, "Towards a formal semantics of UML 2.0 activities." German software engineering conference, 2005. http://wwwcs.uni-paderborn.de/cs/ag-engels/Papers/2005/SE2005-Stoerrle-Hausmann-ActivityDiagrams.pdf.

[12] D. R. Patel, "Inventory control system for a small business proprietor, project, University of Leeds (School of Computing), 2006. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&ved=0ahUKEwiV34-Ly-TZAhXOJSwKHY3nDfQQFgg8MAQ&url=https%3A%2F%2Fminerva.leeds.ac.uk%2Fbbcswebdav%2Forgs%2FSCH_Computing%2FFYProj%2Freports%2F0506%2FPatelDR.pdf&usg=AOvVaw07Q6nL2rvFvy2tc7ilLRmL

[13] H. Quang and T. Pham, "Expressing a temporal entity-relationship model as a traditional entity-relationship model," in M. Núñez, N. Nguyen, D. Camacho, and B. Trawiński (eds), "Computational collective intelligence. Lecture notes in computer science," vol. 9330. Springer, Cham, 2015.

[14] Y. Rinardwiatma and S. Kocharanchitt, "How IBM Maximo helped this Indonesian mining company save over $70 million on inventory in two years," IBM InterConnect 2017 Conference, March 19 – 23, Las Vegas.

[15] A. Kusiak, E. Szczerbicki, and R. Vujosevic, "Intelligent design synthesis: An object-oriented approach," Int. J. Production Res., vol. 29, pp. 1291–1308, 1991.

[16] Y. P. Khanal, "Object-oriented design methods for human centered engineering," Ph.D. thesis, Ontario, Canada: University of Western Ontario, 2010.

[17] A. Tchantchane, "Case study: Design and implementation of an ordering system using UML, formal specification and Java builder," SETIT 2005, 3rd International Conference: Sciences of Electronic,Technologies of Information and Telecommunications, March 27 – 31, 2005, Tunisia.

[18] Y. Mordecai, "Cyber-physical disruption modeling, analysis, and management: An evolutionary object-process model-based robust systems engineering approach," Ph.D. thesis, Israel Institute of Technology, February 2016.

[19] P. Shoval and J. Kabeli, "FOOM: Functional- and object-oriented analysis & design of information systems: An integrated methodology," Journal of Database Management, vol. 12, no. 1, pp.15 – 25, 2001.

[20] D. Dori, "Why significant UML change is unlikely. Communications of the ACM," vol. 45, no. 11, pp.82 – 85, 2002.

[21] S. Al-Fedaghi and M. BehBehani, "Thinging machine applied to information leakage," Int. J. Adv. Comp. Sci. and Applics. (IJACSA), vol. 9, no. 9, pp. 101–110, 2018.

[22] S. Al-Fedaghi and H. Almutairi, "Diagramming language for process documentation," 15th Int. Conf. on Appl. Comp. (AC 2018), Budapest, Hungary, 21–23 October, 2018.

[23] S. Al-Fedaghi and H. Aljenfawi, "A small company as a thinging machine," 10th Int. Conf. on Info. Mgmt. and Eng. (ICIME), University of Salford, Manchester, England, September 22 – 24, 2018.

[24] S. Al-Fedaghi and M. Alsharah, "Security processes as machines: a case study," Eighth Int. Conf. on Innov. Comp. Technol. (INTECH), London, England, August 15–17, 2018.

[25] S. Al-Fedaghi and R. Al-Azmi, "Control of waste water treatment as a flow machine: a case study," The 24th IEEE Int. Conf. Autom. Comp. (ICAC), pp. 6-7, September 2018, Newcastle University, Newcastle upon Tyne, England.

[26] S. Al-Fedaghi and N. Al-Huwais, "Toward modeling information in asset management: case study using Maximo," 4th Int. Conf. on Info. Mgmt. (ICIM), Oxford, England, May 25–27, 2018.

[27] M. Heidegger, "The thing," in Poetry, language, thought, A. Hofstadter, Trans. New York: Harper & Row, 1975, pp. 161–184.

[28] T. Ingold, "Bringing things to life: Creative entanglements in a world of materials," Realities working papers #15: Bringing things to life, Department of Anthropology, University of Aberdeen, UK, July 2010.

[29] IBM Maximo, Maximo Core Concepts [Online]. Available: https://www.slideshare.net/sunk818/ibm-maximo-keyconceptsv43 IBM, IBM Maximo Asset Management, Version 7 Release 6, Workflow Implementation Guide, IBM Corp., 2008, 2014.

[30] IBM Maximo, "User's guide," Release 6.2.1, January 2007.

[31] R. Potts, "Leveraging the ArcGIS platform to support Maximo spatial, SSP innovations, empowered solutions blog," March 10, 2018. https://sspinnovations.com/blog/leveraging-arcgis-platform-support-maximo-spatial/

Dynamic Tuning and Overload Management of Thread Pool System

Faisal Bahadur, Arif Iqbal Umar, Fahad Khurshid
Department of Information Technology, Hazara University
Mansehra, K.P.K. Pakistan

***Abstract*—Distributed applications have been developed using thread pool system (TPS) in order to improve system performance. The dynamic optimization and overload management of TPS are two crucial factors that affect overall performance of distributed thread pool (DTP). This paper presents a DTP, that is based on central management system, where a central manager forwards client's requests in round robin fashion to available set of TPSs running in servers. The dynamic tuning of each TPS is done based on request rate on the TPS. The overload condition at each TPS is detected by the TPS itself, by throughput decline. The overload condition is resolved by reducing the size of thread pool to previous value, at which it was producing throughput parallel to the request rates. By reducing the size of thread pool on high request rates, the context switches and thread contention overheads are eliminated that enables system resources to be utilized effectively by available threads in the pool. The result of evaluation proved the validity of proposed system.**

Keywords—Distributed system; distributed thread pool; thread pool system; performance; overload management

I. Introduction

The ever-growing expansion of Internet and World Wide Web demands scalable services that must be performance efficient and highly available. The prominent progression of internet's user doubles internet traffics every two or three months. For example, OSN sites such as LinkedIn, Flickr, Myspace, Twitter and Facebook provide facilities to over half a billion users at the same time [1]. The OSN's not only provide basic communication capabilities but also provide other services by third party applications e.g. sharing documents, sending virtual gifts, or gaming. Number of third-party applications run by the Facebook are over 81,000 [1]. There is a profound impact of these third-party applications on the application server's scalability and performance thus results in additional traffic. For example, when Facebook launched its developer platform, the traffic increased by 30% in a week after launching [2], while in case of Twitter the traffic increased by a factor of 20 after launching its API [3]. Also, the variations in demand go to extreme levels in some internet services that cause overload condition and needs special attention to manage server-side resources. In order to deal with these complexities, internet services are provided by distributed application servers [4], that are responsible of providing run time services to applications, where these applications service demands of many concurrent users.

At present, distributed systems have been implemented in almost all domains including telecommunication, defense, industrial automation, financial services, entertainment, government and e-commerce. And that is why, the requirements of complexity management, scaling and overload management are increasing day by day. As discussed earlier, distributed systems handle heavy workloads, where client's requests are incoming from a remote source through some network protocol. These heavy workloads are handled by distributed systems through extremely concurrent design configurations that are implemented as middleware. The performance of distributed systems is dominated by middleware that provide different functionalities, e.g. multithreading policies, remote communication mechanisms, persistence services and transaction management etc. It is the middleware that makes distributed system scalable, highly available and highly performant [5]. Some remarkable examples of middleware services for distributed systems are middleware of Distributed Object Computing (such as CORBA, SOAP, RMI) Component middleware (such as .NET, Java Beans), Message Oriented Middleware (such as Java Message Queue, BEA's WebLogic MessageQ) etc.

One of the most important performance related feature of any middleware service in distributed systems is concurrency control that handles multiple concurrent requests. Two most commonly used concurrency models are Thread Pool System (TPS) and event driven model (EDM).

As compared to TPS, EDM is more performance efficient, but at the same time it is much complicated and challenging to implement than TPS [6]. The most challenging task in EDM is to handle scheduling and assembling of events [7]. Moreover, EDM leads to enormous cascading callback chains [8]. As compared to EDM, TPS offers more solid structuring constructs for concurrent servers by means of threads that are light weight and represent work from the perception of the task itself [9,10]. Moreover, TPS avoids resource thrashing and overheads of thread creation and destruction [11]. Some examples of TPS in middleware for distributed systems include .NET thread pool [12], Java Message Queue Thread Pool [13].

A typical TPS contains a request queue, a pool of threads (workers) and dynamic optimization algorithm (DOA) that optimizes pool size, as shown in Fig.1.

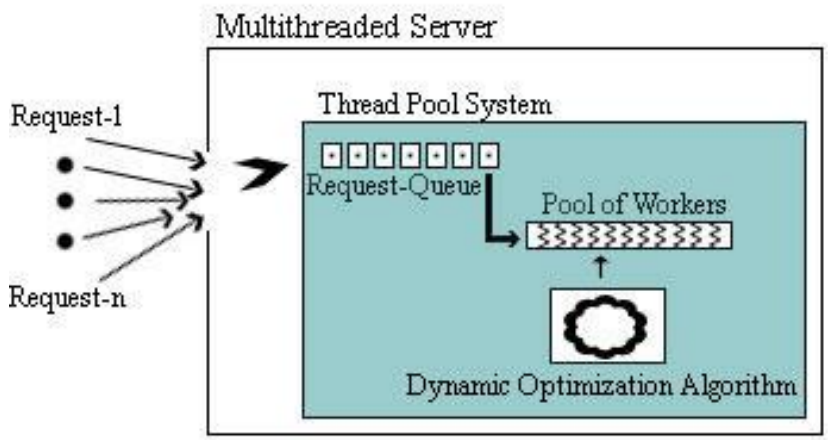


Fig. 1. Conceptual Model of Thread Pool System Embedded in a Server.

Request queue stores incoming client's requests. Worker threads in the pool fetches and executes these requests. These worker threads in the pool are recycled (instead of being destroy) to process next client's request from queue. The re-spawning and recycling of worker threads avoids thread creation and destruction costs, but, under a heavy load scenario, additional threads must be dynamically created and inserted inside pool to cope with the load. The DOA component of TPS is responsible to decide the quantity of extra threads. It is a challenging task of DOA to maintain an optimal pool size on run time, in order to produce better response times and maximum throughput so that quality of service can be maintained. If thread pool size is beyond an optimal limit, then it increases thread context switches and thread contention (on shared resource), that ultimately provides poor performance. On the other hand, if pool size is smaller than an optimal limit then it results in poor response time and throughput. Handling this tradeoff on run time is essential to achieve best performance. Optimizing thread pool size by DOA is not an exact science and it can be performed on the basis number of parameters and factors.

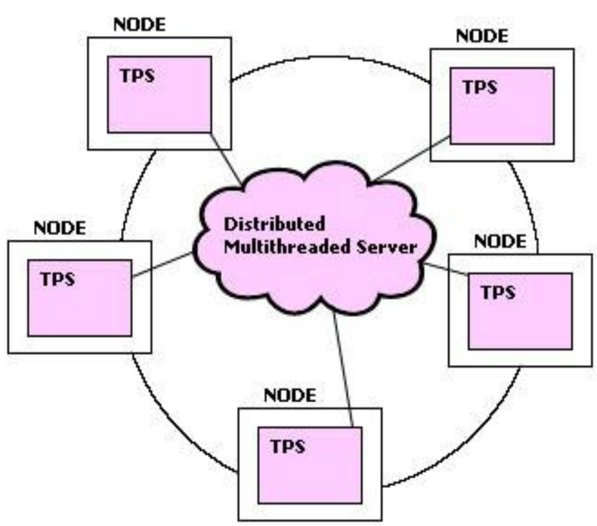


Fig. 2. Conceptual Model of Distributed Thread Pool System.

The variety of target servers where TPSs are installed makes DOA more challenging, as there are varied characteristics of the deployment system with a diverse nature of tasks. Because of this reason, TPS has been evolved from single-pool to multi-pool and from multi-pool to DTP. DTPs are designed for distributed systems where they are horizontally scaled over number of nodes available on the network as shown in Fig. 2.

In DTPs, overload monitoring of each TPS running in a server node is a crucial factor to avoid unsuitable large quantity of threads in the pools, so that overheads of thread contention and context switches may be reduced. Moreover, metric for detecting overload condition must be selected carefully in order to gain maximum performance.

This work is based on our previous work [14], in which, we have presented a distributed framework of TPS called distributed frequency based thread pool (DFBTP), where each server node has its own TPS, that is tuned on the basis of request arrival rate, and the load on each node is balanced by a round robin strategy in order to fairly distribute the load among available TPSs. In this paper, we have extended DFBTP by presenting overload control based distributed thread pool (OCBDTP), whose overload control mechanism tackles overload condition by detecting throughput fall, in case of high request frequencies. In such a case, thread pool size is restored to previous appropriate size, where it was running normally.

The rest of the paper is organized as follows. Section 2 presents literature review. The design of is presented in section 3. The validation of proposed system is detailed in section 4, and finally conclusion is given in section 5.

II. Related Work

A mathematical model for dynamic optimization of TPS is presented in [15], that forms a relationship among system load, pool size and the costs associated with thread creation, destruction and maintenance. However, the estimated optimal thread pool size might be inaccurate, since accurately estimating the time for thread creation and thread context switching is difficult in practice.

CPU-utilization based TPS is presented in [16], where dynamic optimization scheme increases pool size when CPU utilization decreases and vice versa. It defines a lower threshold and upper threshold variables for CPU utilization, and an update function repeatedly optimizes pool size on the basis of these threshold variables. This scheme however can't be used for I/O bound applications.

TPS presented in [17] performed dynamic optimization by calculating average idle time (AIT) of queued requests. This scheme increases pool size if AIT is increasing. However, the uses of too many thresholds in dynamic tuning algorithm have no justification and can affect the performance.

A prediction-based scheme is utilized in [18], in order to set a pool size in advance for future use by Gaussian distribution. These predictions might be inaccurate due to synchronization overhead.

Exponential moving averages were utilized in [19], in order to predict pool size in advance. This scheme suffered from creating redundant threads.

Thread borrow scheme was used in an application server [20], that utilized multiple thread pools, but management of multiple pools is itself a cost-effective operation.

A model fuzzing approach was used in [21], to optimize pool size. Number of parameters and constraints were utilized for optimizing pool size which was too complex to quickly make a conclusion, hence it is not suitable for the system having frequent state change.

TPS presented in [22], used response coefficient to optimize pool size. However, this metric is normally affected by different run time parameters.

By extending the work of [18], the trends of time series in exponential moving averages were analysed in [23], in order to avoid redundant threads. This scheme however suffered from creating lacking threads.

A theoretical framework of distributed thread pool was presented in [24], that is governed by software agents that can dynamically add and remove threads in the pool based on load conditions.

A design of hierarchical thread pool executor based on non-blocking queue was presented in [25]. This TPS was presented for java-based DSMs running on cluster. This TPS served as an outclass alternative to Jackal's thread pool executor which is based on blocking queue. This design was targeted to only DSM systems.

The dynamic optimization of TPS in [26] is performed by performance data of threads associated with system resource usage. The calculated value is gradually updated and compared with old values over time. This trend is analysed over time to perform dynamic optimization.

It is argued in [27], that response times of a TPS-based application suffer when it utilizes more threads than required, as it increases context switching overhead. And that is why he established an inverse proportional relationship of response time to pool size. This TPS gradually decreases pool size on high response times, until system's stability.

A divide and conquer strategy is used in [28] that divides the tasks into subtasks by a pipeline based technology that allows the sub-tasks to run in parallel that reduced the computational cost. But dynamic division of tasks into subtasks was challenging.

Thread pool system presented in [29] is optimized based on application level metric called thread utilization. On high thread utilization, thread pool size is increased and vice versa.

A multiple pool approach is used in [30], where each pool is reserved to process requests having specific service time. In this way requests having large service times are separated from requests having small service times, hence avoiding large requests to block small ones in order to occupy all threads in the pool. However, a large variation in service times of incoming requests results in large number of pools that increases pool management overhead.

In [31], a middleware for CPS-systems is presented which utilized a linear approach to tune its thread pool. Thread pool is resized on the basis of increase in request rate. The system uses unbounded DTS that is only suitable for CPS systems.

A dynamic framework is offered in [32] for n-tire application that is running in the cloud-based system. At each tier, thread pool is tuned by queuing laws and system-level metrics.

In our previous work [14], we have presented DFBTP that was designed for distributed applications, where each node has its own thread pool. The size of each thread pool is optimized by frequency of incoming requests and load is equally distributed among all nodes by a round robin scheme. This scheme creates threads in the pool based on request arrival rate that can be very large on heavy load situations that may result in very large pool size that effects performance.

III. Motivation

Each TPS in DFBTP [14] keeps pool size equal to the request arrival rate. In case of high request arrival rate, DFBTP creates a very large pool size that effects performance. Fig. 3 is an illustrative diagram, that demonstrates the problem of throughput degradation due to worse pool size on heavy load.

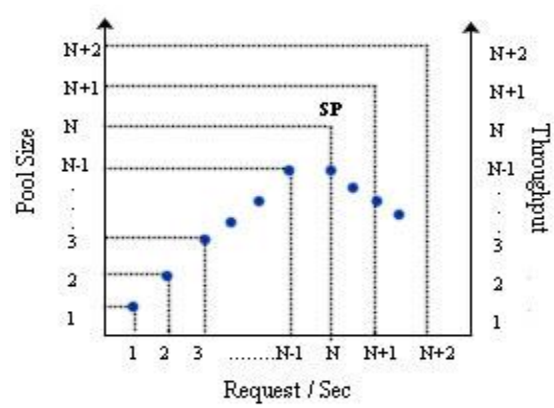


Fig. 3. Throughput Starts Falling on Saturation Point.

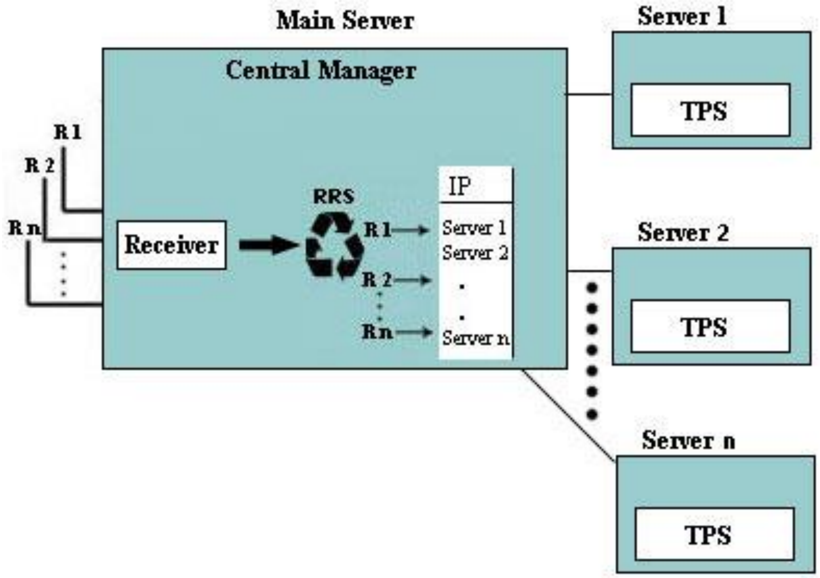


Fig. 4. Architecture of Distributed System having Frequency-based Thread Pools on Slave Servers.

X-axis represents request arrival rate, whereas y-axis consists of multiple axis. PoolSize axis represents number of threads in the thread pool, and Throughput axis represents number of requests completed per second. Throughput is represented by blue dots. We can see in Fig.1, that pool size is equal to the request arrival rate at every moment. Also, throughput is gradually increasing for some moments and equal to the request arrival rate. When request arrival rate is N, the system kept pool size equal to N also, however, throughput decreased and turned into N-1. This point is called saturation point (i.e. overload condition), where throughput started falling, because, DFBTP continuously increasing pool size on next successions and due to very large pool size, system is busy in context switch overhead and thread management overhead. System resources are busy in managing thread context switches and thread contention, instead of doing useful work. The motive of proposed scheme is to tackle saturation point and react accordingly.

IV. Material and Method

In this section we discuss in detail, the design and implementation of OCBDTP. First, we will discuss the distributed architecture of our proposed system, and next we will describe the design of proposed TPS, followed by overload control mechanism.

A. Proposed Distributed System Architecture

OCBDTP consists of a central manager (CM) component that is running in the main server, and one or more TPS running in server nodes, as shown in Fig. 4. System initialization starts from CM that runs in the main server first and waits for the TPS to connect with it. When a TPS starts on a server node in the network, it connects to the CM, that in turn stores the IP address of corresponding server in a list and starts its Receiver thread that will accept client's requests and handover to round robin scheduler (RRS). RSS is a load balancer component that equally distributes the load on all available TPS by iterating over the list that contains IP addresses of servers. CM is now ready to accept client's request. On request arrivals, Receiver will receive requests and forward it to the RRS, that in turn distribute these requests to TPSs. Each TPS is responsible to accept requests and process these requests.

B. System Architecture of TPS

The architecture of proposed TPS is shown in Fig. 5. When a TPS starts in the server node, it initializes three detector threads named Request Rate Detector (RRD), Throughput Detector (TD) and Saturation Detector (SD). Next, it connects to the CM through Connector component. On arrival of first request, TPS starts detector threads that perform their tasks after every second until TPS is running.

A counter is incremented on every request arrival. Each request arrived at TPS is first inserted into request queue and picked up by any available thread inside pool that will process request and store it as a response in response queue. RRD activates after every second, reads the value of counter, stores it as a current request frequency and sets the value of counter again to zero in order to detect the next request arrival rate. RRD maintains request rates of current and previous phases. These request rates are later read by SD in order to adjust pool size equal to the request frequency if throughput is not falling. TD activates after every second, counts and de-queues all the responses from response queue. The counts of responses are stored in Throughput object. The responses are sent back to the CM, that in turn sends these responses back to the client. TD maintains throughputs of current and previous phases.

SD also activates after every second that performs two jobs. First, it tunes thread pool size based on request rates. Second, it periodically measures throughput of TPS in order to detect overload condition. In case of overload condition SD reacts accordingly that is discussed in the next section.

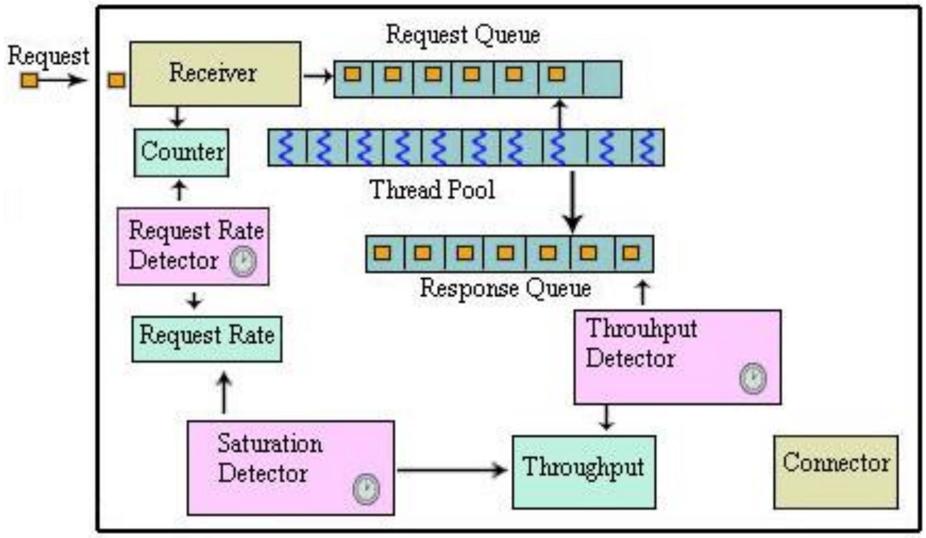


Fig. 5. Architecture of Proposed Thread Pool System.

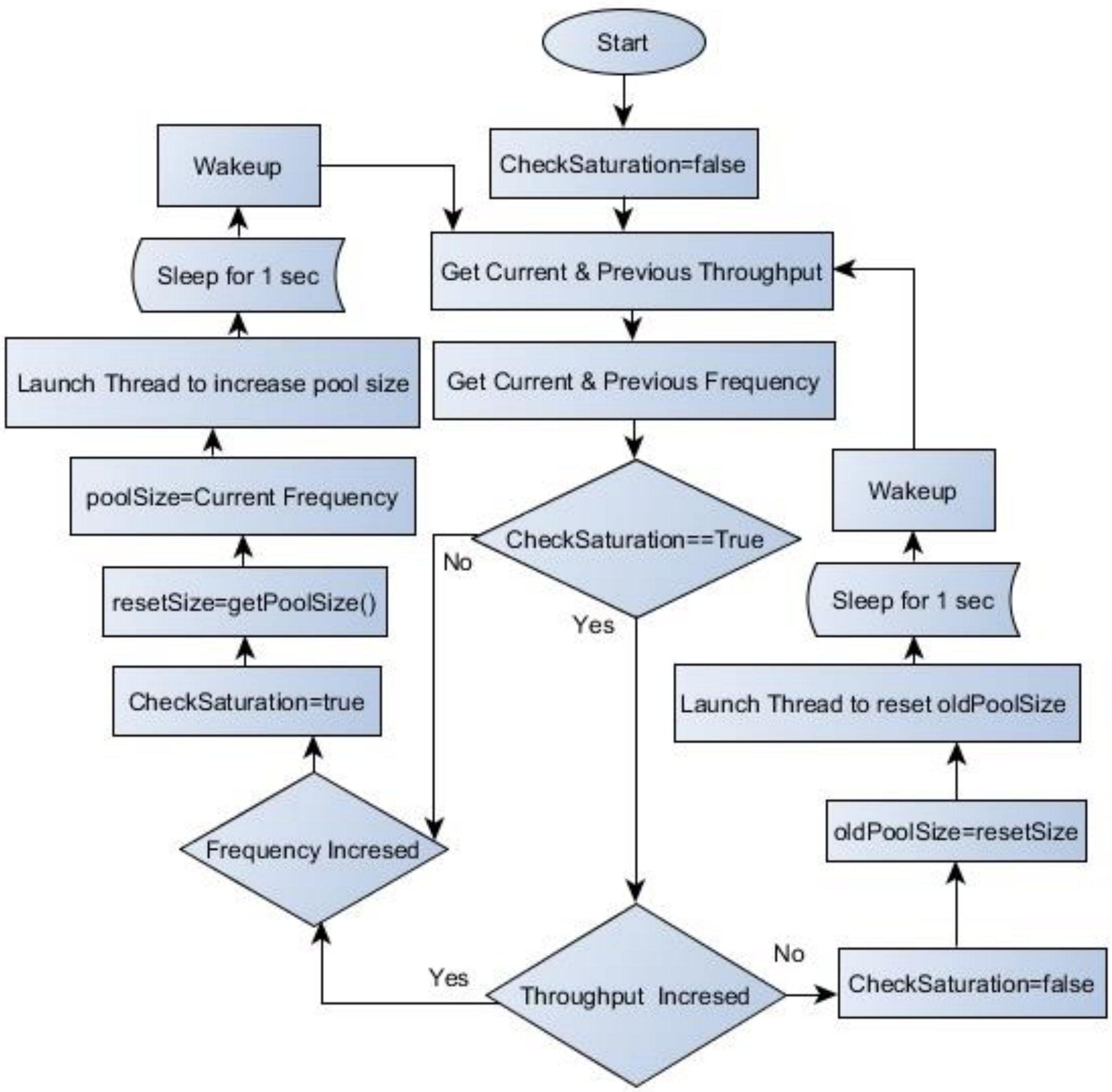


Fig. 6. Flowchart Diagram of Saturation Detection and Avoidance Mechanism.

C. Overload Detection and Control

Fig. 6 is a working flow chart diagram of SD, that dynamically sets an appropriate pool size of thread pool that can avoid thread context switch and contention overheads. And this is done by detecting an inappropriate pool size at which throughput starts falling. SD algorithm first collects values of throughout and frequency regarding current and previous phases. Then, it checks for increase of request rate; if it is increasing, then it sets pool size equal to the request rate by running a separate thread in the background and repeats itself after one second. In the next pass it first checks for throughput gain. Since quantity of threads was increased in the last phase, that means more throughput should be there, and throughput should be equal to the size of thread pool per second. If throughput is improved and equals to the number of threads in the pool, then the size of thread pool is considered to be appropriate and it is saved for later use in case of overload, so that, an inappropriate size can be restored to saved one. Next, it again checks for current request rate; performs corresponding steps; and repeat after one second. In case throughput is not improved, it means that pool size was not set to an optimal value in the last phase, so it resets pool size again to the saved one (by running a separate thread in the background) where TPS was running normally, i.e. giving throughput equal to the request rate. On high request rate, SD first keeps pool size high, and watches over throughput. However, due to large number of threads in the pool beyond the capacity of system, the system turns into a state, where it is busy in managing thread context switches and thread contention, instead of doing useful work by threads in the pool. And that is why, its throughput starts falling initially. However, SD detects this overload condition, and resets pool size to the previous phase's size where there were no overheads of thread context switches and contention. Restoring pool size eliminates overheads of thread context switches and contention and system resources are effectively utilized by available threads that again restores system performance.

V. Results and Discussion

In order to validate proposed thread pool system, we simulated it using a java-based toolkit named Pool-Runner[1] that can embed a DTP inside its server tier for performance testing. Its client tier has a load generation engine that displays results in form of graphs. We performed simulation on a network with three machines. First machine has a client tier of toolkit, second machine is used as a main server running CM and third machine is used as a slave running a TPS.

We use static workload for this test. The static workload is simulated by StaticTask object (available in toolkit) that simulates 1kb file by sleeping for 100 milliseconds (approximately). Load is first generated by poisson distribution with the rate of 1000 request/sec as shown in Fig. 7. After every minute we increased the load. As shown in Fig. 7 that the load is 2000 request/sec in second minute, and 3000 request /sec in third minute.

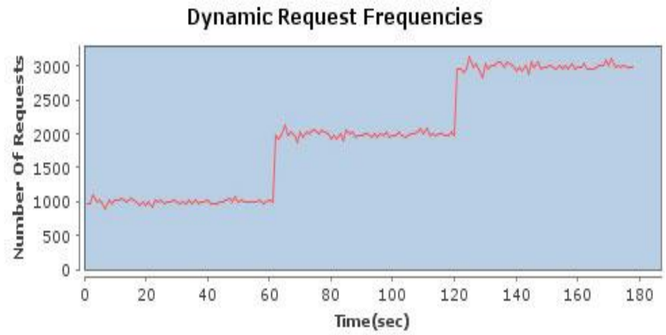


Fig. 7. System Load Used for Simulation.

Fig. 8 shows comparative analysis of thread pool size between DFBTP and OCBDTP, where DFBTP kept its pool size equal to request rate all the time, whereas OCBDTP kept its pool size equal to the request rate till 2 minutes, however, pool size is restored to previous appropriate size that was in the second minute. And this is because of throughput falling in 3rd minute. When request rate turned into 3000 request/sec in 3rd minute, OCBDTP detected saturation in the system, that can be seen in Fig.9 that shows that throughput started falling initially but raised up again when pool size is set to an optimal level.

Fig. 9 shows throughput per second, where DFBTP sustained its throughput in first two minutes, as average throughput is equal to the request rate, however in 3rd minute, its throughput is continuously falling because its pool size is very large that is creating thread management (context switches and contention) overheads. OCBDTP also sustained its throughput till 120 seconds, however, when request rate raised up to 3000 requests/sec, its throughput started falling initially, that is recovered by SD, and pool size is set to an optimal value that caused throughput to sustain again.

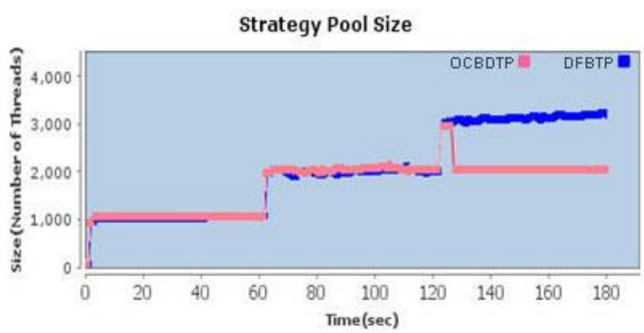


Fig. 8. Comparative Analysis of Pool Size.

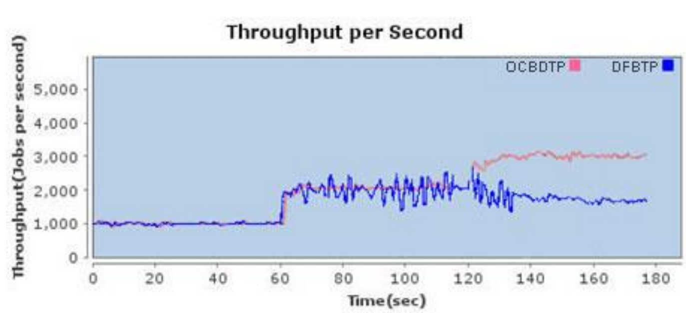


Fig. 9. Comparative Analysis of Throughput Per Second.

[1] http://www.jpoolrunner.net

VI. Conclusion

This paper has presented a distributed thread pool system, that is based on central manager, that forwards client's requests (in round robin fashion) to all instances of thread pools running in the distributed environment. The overload condition at each TPS is detected by throughput decline and resolved by reducing and restoring pool size to previous appropriate value, at which the system was stable. The overload control (by reducing pool size) eliminates thread management overhead, that enables system resources to be used effectively by threads in the pool. Proposed system prevents pool size to increase beyond an appropriate level, hence avoids overheads of thread context switching and contention, hence increases system performance. Proposed system results in more throughput gain. In the future, we will tune thread pools by software agents having artificial intelligence capabilities that would be used on networks and clusters for automatic control of distributed system resources.

References

[1] A. Nazir, S. Raza, D. Gupta, C-N. Chuah, B. Krishnamurthy, "Network level footprints of facebook applications, "In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement, pp. 63-75, 2009.

[2] A. Nazir, S. Raza, C.-N. Chuah, "Unveiling facebook: A measurement study of social network based applications, "In Proc. Internet Measurement Conference (IMC), 2008.

[3] B. Krishnamurthy, P. Gill, M. Arlitt, "A few chirps about twitter," In Workshop on Online Social Networks, 2008.

[4] N.Aghdaie , Y.Tamir, " Fast transparent failover for reliable web service," InInternational Conference on Parallel and Distributed Computing and Systems, pp. 757-762, 2003.

[5] G. Denaro, A. Polini, W. Emmerich, "Performance testing of distributed component architectures," In Testing Commercial-off-the-Shelf Components and Systems, Springer, Berlin, Heidelberg, pp. 293-314, 2005.

[6] M. Andreolini, V. Cardellini, M. Colajanni, " Benchmarking models and tools for distributed web-server systems," In: IFIP International Symposium on Computer Performance Modeling, Measurement and Evaluation, pp. 208-235, September, 2002.

[7] M. Welsh, D. Culler, E. Brewer, "SEDA: an architecture for well-conditioned, scalable internet services," ACM SIGOPS Operating Systems Review, 35(5), pp. 230-243, 2001.

[8] A. Gustaffson, "Threads without the Pain," Queue, Vol. 3, No. 9, pp. 34-41, 2005.

[9] R.V. Behren, J. Condit, E. Brewer, "Why events are a bad idea (for high-concurrency servers)," In: Proceedings of the 9th conference on Hot Topics in Operating Systems, pp.4-4, 2003.

[10] R.V. Behren, J. Condit, F. Zhou, G.C. Necula, E. Brewer, "Capriccio: scalable threads for internet services," ACM SIGOPS Operating Systems Review, Vol. 37, No.5, pp. 268-281, 2003.

[11] T. Peierls, B. Goetz, J. Bloch, J. Bowbeer, D. Lea, D. Holmes, "Java Concurrency in Practice," Addison-Wesley Professional, 2005.

[12] The Managed Thread Pool https://msdn.microsoft.com/enus/library/0ka9477y(v=vs.110).aspx (accessed on 10 July 2018).

[13] ORACLE, Java System Message Queue 4.3 Technical Overview https://docs.oracle.com/cd/E19316-01/820-6424/aerck/index.html (accessed on 10 July 2018).

[14] S. Ahmad, F. Bahadur, F. Kanwal, R. Shah, "Load balancing in distributed framework for frequency based thread pools," Computational Ecology and Software, Vol. 6, No. 4, pp. 150-164, 2016.

[15] Y. Ling , T. Mullen , X. Lin, "Analysis of optimal thread pool size," ACM SIGOPS Operating Systems Review, Vol. 34, No. 2, pp.42-55, 2000.

[16] D. Robsman, "Thread optimization," US6477561 B1, 2002.

[17] D. Xu , B. Bode, "Performance Study and Dynamic Optimization Design for Thread Pool System," In Proc. of the Int. Conf. on Computing Communications and Control Technologies, pp.167-174, 2004.

[18] J. Kim , S. Han , H. Ko , H. Youn, "Prediction- based Dynamic Thread Pool Management of Agent Platform for Ubiquitous Computing," In Proceedings of UIC, pp. 1098-1107, 2007.

[19] D. Kang , S. Han , S. Yoo , S. Park, "Prediction based Dynamic Thread Pool Scheme for Efficient Resource Usage," In Proc. of the IEEE 8th Int. Conf. on Computer and Information Technology Workshop, IEEE Computer Society, pp. 159-164, 2008.

[20] T. Ogasawara, "Dynamic Thread Count Adaptation for Multiple Services in SMP Environments," IEEE International Conference on Web Services, pp. 585-592, 2008.

[21] J. Hellerstein, "Configuring resource managers using model fuzzing: A case study of the .NET thread pool," IFIP/IEEE International Symposium on Integrated Network Management (IM '09), pp. 1-8, 2009.

[22] N. Chen, P. Lin, "A Dynamic Adjustment Mechanism with Heuristic for Thread Pool in Middleware," 3rd Int. Joint Conf. on Computational Science and Optimization. IEEE Computer Society, pp. 324-336, 2010.

[23] K. Lee , H. Pham , H. Kim , H. Youn , O. Song, "A novel predictive and self-adaptive dynamic thread pool management," In: Proceedings - 9th IEEE International Symposium on Parallel and Distributed Processing with Applications, pp. 93–98, 2011.

[24] P. Martin , A. Brown , W. Powley , J. Luis , V. Poletti, "Autonomic management of elastic services in the cloud," In: Proceedings - IEEE Symposium on Computers and Communications, pp. 135–140, 2011.

[25] S. Ramisetti , R. Wankar, "Design of hierarchical thread pool executor for DSM. Second International Conference on Intelligent Systems Modelling and Simulation (ISMS), pp. 284-288, 2011.

[26] F. Muscarella, " Method and apparatus for dynamically adjusting thread pool," US8185906 B2, 2012.

[27] R.T. Gowda, "Dynamic thread pool management," US 8381216 B2, 2013.

[28] M. Chen, Y. Lu, G. Zhang, W. Zou, "Real-Time Simulation in Dynamic Electromagnetic Scenes Using Pipeline Thread Pool," 10th International Conference on Computational Intelligence and Security (CIS), pp. 770-774, 2014.

[29] S. Sahin, "C-stream: A coroutine-based elastic stream processing engine", Doctoral dissertation, bilkent university, 2015.

[30] J. Mace, P. Bodikz, M. Musuvathiz, R. Fonseca, K. Varadarajanz, " 2DFQ: Two- Dimensional Fair Queueing for Multi-Tenant Cloud Services," In Proceedings of ACM SIGCOMM Conference, ACM, pp. 144-159, 2016.

[31] M. García-Valls, C. Calva-Urrego, A. Juan, A. Alonso, "Adjusting middleware knobs to assess scalability limits of distributed cyber-physical systems," Computer Standards & Interfaces, Volume 51, pp.95-103, 2017.

[32] H. Chen, Q. Wang, B. Palanisamy, P. Xiong, "DCM: Dynamic Concurrency Management for Scaling n-Tier Applications in Cloud," In IEEE 37th International Conference on Distributed Computing Systems (ICDCS), pp. 2097-2104, 2017.

Development of Interactive Ophthalmology Hologram

Sarni Suhaila Rahim[1], Nazreen Abdullasim[2], Wan Sazli Nasaruddin Saifudin[3]
Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

Raja Norliza Raja Omar[4]
Department of Ophthalmology, Hospital Melaka, Jalan Mufti Haji Khalil, 75400 Melaka, Malaysia

***Abstract*—Ophthalmology is one of the medical branches that deal with the eye. This field is associated with the anatomy, physiology and diseases of the eye. The main objective of this paper is to develop a novel interactive simulation of eye anatomy in three-dimensional (3D) by using holography environment approach in order to create better visualization of the structures of the eyes. Currently, public can access the information on medical through some conventional methods, such as brochure, pamphlet, booklet, in addition to some other better ways, for example 2D and 3D video. However, these methods do not offer an interactivity visualization of the medical information that help creating much engaging presentation to users. Moreover, the medical doctors are unable to show and explain in details about the diseases occurrence through these methods. An interactive method, therefore is required to assist the doctors to convey the disease information effectively. Since the human eye is the most complex organ in our body, thus an advanced technology, i.e., hologram should be used to visualize the visual system in an effective and interactive method, also to produce an effective eye disease explanation. Hologram is in a three-dimensional form besides an interactivity elements, where the proposed technology will help doctors examining vital organs using 3D displays. It is envisaged that the proposed interactive holography technique for clinical purpose would greatly contribute to and assist the management of the eye diseases and other diseases as well. It is hoped that the developed interactive holography technique for eye will assist clinicians in delivering the disease information efficiently and attractively.**

Keywords—3D animation; hologram; ophthalmology; interactive; eye

I. Introduction

Holography was invented by Denis Gabor, a Hungarian physicist who found basic principles of microscopy while exploring to improve the transmission electron microscope efficiency in 1948 [1]. Recently, holography is a popular research area and many researchers focus on and contribute to the advancement of this study area. Most researchers focus on finding and proposing the technique or system through exploring the holography. Although there have been immense advancements in this area of research, there are still lacunae or spaces for improvement. The proposed techniques or systems in this research will most notably benefit the realm of holography in a number of areas or ways.

A study on the recent improvements in creating three-dimensional images and videos by holographic techniques and the potential of holography in future is presented by Abbasi et al. [2]. Among the advantage of holography is the real 3D display without use of any other viewing aids and the quality of the science-art will be improved and as a result, it is impossible to differentiate between holographic images and the real objects [2]. In addition, in the field of art, Oliveira and Richardson [3] claim that the holographic technologies help change the cultural perception of artists and art institutions to hold holographic science. Moreover, the holographic technologies enable to create new dimensions and the likelihood to display a subject in 3D and beyond [3]. For example, one of the holographic features in Hologram Table [4], provides the display of digital models of cities or buildings as miniatures, with the ability to zoom in down to single blades of grass or even smaller.

This research aims to create an interactive hologram specifically for one of the medical branches, which is ophthalmology. There is a lack of interactivity visualization of the medical information through the conventional methods. In addition, through the conventional methods, such as brochure, pamphlet, booklet, or even 2D and 3D video, the presentation and explanation of diseases in details could not be conveyed effectively. Moreover, ophthalmology which deals with eyes, is consider an innovative and fast-moving field, where the human eye is the most complex organ in our body. Therefore, an advanced technology such as hologram should be used to visualize the visual system and produce eye disease explanation in an effective and interactive method.

The objectives of this paper are to study the holography method, to design and develop a novel 3D interactive simulation of eye anatomy by using holography environment approach in order to create better visualization of the structures of the eyes. Target users of this product are from different fields, such as for medical field, it can be used by the medical doctors, ophthalmologist and optometrists. In addition, for training field, the medical trainers or practitioners, nurses and paramedics are benefited from this research product. Besides, in education field, the medical lecturers, medical students, multimedia lecturers and multimedia students are benefited. Also, the public and patients can use this product as well. Currently, the research product focusing on one important module, which is Eye Anatomy. Methodology used for this research product is Multimedia Production Process which consists of three stages which are pre-production, production and post-production. The software used to create this product is Unreal. The outcome of this research is the interactive application of ophthalmology with holographic method. Besides, the research product provides the novel content verification for eye anatomy features by the Consultant Ophthalmologist, new verified human eye model and a novel voice-over recording for eye anatomy pronunciation by the Consultant Ophthalmologist. In addition, the research also received support, particularly in the contribution of expertise, from the Department of Ophthalmology at the Hospital Melaka, Malaysia.

This paper is organized as follows. Section 2 presents previous related work on hologram and interactive hologram system. Section 3 describes in details the overall methodology of this research. The discussion of the research paper is presented in Section 4. Finally, Section 5 presents some conclusions and future work.

II. Literature Review

The research product, Interactive Ophthalmology Hologram is an application developed using an engine called Interactive Hologram System. Interactive Hologram (i-HO) is an interactive system for viewing real-time 3D content on a holographic pyramid panel. This system consists of two parts, which are i-HO engine and i-HO projection panel. The novel development of an interactive hologram engine successfully overcome the limitation of traditional hologram, such as linear and video-based, no interaction and limited content representation. The developed system opens the possibility for creating an exciting 3D interactive content which can be used in various field, like games, education, architecture, engineering, medical and others. The interactive hologram architecture and process is presented in details in [5]. Fig. 1 shows the architecture of Interactive Hologram System, representing how input devices, i-HO engine and the projection panel are being implemented.

The basic i-HO architecture consists of capturing images from real-time 3D content, mapping and projecting phases [5]. Hologram engine enable the user to interact in real-time with 3D content. The capturing part is a capturing process of a real-time 3D object from four different perspectives, which are frontal, back, left and right. Next, the captured images will be mapped into a single plane accordingly.

Interactivity is the main feature of i-HO as it enables the developer to customize the engine for 3D game development and other 3D interactive contents. This important feature is different compared to many conventional hologram pyramid where the video or images were pre-rendered, linear and not interactive. Interactive hologram implemented provides interactive and immersive holography environment, in addition to real-time and interactive content in order to create better visualization. Therefore, with these capabilities offered by interactive hologram system, the developed Interactive Ophthalmology Hologram is able to produce better visualization of the eye model to show from different perspective includes frontal, back, left and right. In addition, it helps the model interacts with input devices, like keyboard, mouse, joystick or gamepad, sensor, motion capturing and motion controller. Moreover, the feature such as zoom in and out is provided to increase the usability efficiency of the product.

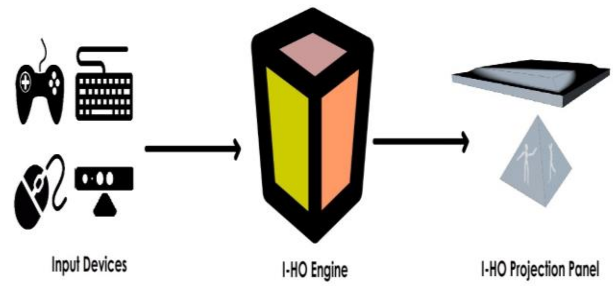


Fig. 1. Interactive Hologram System Architecture [5].

There are numerous medical application developed using holography. Holography in medicine and biology was presented earlier by Bally [6], comprising of holography in orthopaedic, radiology, ophthalmology, urology, dentistry, otology and others. Mehta [7] summarized the success of medical applications of holography shows a strong potential for holography to appear as a powerful tool for medical applications. Among the holographic techniques in medical are imaging through tissues, ophthalmology, dentistry, urology, otology, pathology and orthopaedics. For example, X-ray holography able to examine the samples in aqueous solution with high resolution without the sample preparation which end up with structural alternations [7]. Meanwhile, endoscopic holography provides a potential to record a 3D large focal depth and high resolution image of internal organs and tissues, hence enhances the capability of detection [7]. Besides, there are internal and external holographic endoscope which have been used for early detection of cancerous indurations in the wall of urinary bladder. Moreover, the multiplexed holography can be used for medical tomography which able to display the three-dimensional tomographic medical data. In addition, for imaging thorough human tissue, the holographic light-in-flight recording method can be adapted. In the ophthalmology field, the recording of a three dimensional eye image application was initially developed.

Recently in ophthalmology field, holography has potential to investigate corneal endothermal morphology, changes on the cornea, crystalline lens changes and the nerve head and the retina surface characteristics [7]. Potcoava summarized various holographic techniques and application on the imaging of ophthalmic tissue, fingerprints and microsphere samples [8]. Meanwhile, holographic in orthopaedics offers a benefit for the orthopaedic structures, especially external fixtures in order to reveal and measure strains on fixation pins and rods. Kim in [9] reviewed the digital holography focusing on the applications in biomedical microscopy, in addition to the related methods, current issues, trends and potential of holographic.

Based on holographic applications in medical presented above, it can be concluded that the developed application able to provide as a tool for visualizing patient data while training students and surgeons. It produces 3D real images and no special viewing devices or glasses are needed. Holographic applications in medical also allow the viewer to examine different organs or body parts, such as brain, liver, lungs, heart, skeleton, vascular system, nerves and muscles. The existing applications have covered different type of modules, different medical branches and different techniques. However, these existing systems do not offer an interactive content for development of the applications. Interactivity between the application and the user is the main issue or feature of holography. Therefore, the proposed research presents a novel interactive content for ophthalmology hologram, besides novel content verification for eye anatomy features, new verified Asian eye model and a novel voice-over eye anatomy pronunciation by the Consultant Ophthalmologist.

III. Methodology

The methodology used in this research is Multimedia Production Process, consists of three main phases, which are pre-production, production and post-production. There are six phases involved in these three main phases, which are analysis, design, implementation, testing, evaluation and publishing. In this section, all requirements will be further analysed in detail.

A. Analysis

In order to gather important information which is the content of the research, an interview session is conducted. The interview and discussion is conducted with the Consultant Ophthalmologist, Department of Ophthalmology, Hospital Melaka, Malaysia. The medical information is required in order to ensure the information provided is valid and sufficient for the user.

The data collection process can be divided into three phases. The first and second processes which are the content verification for eye features and verification on 3D eye model focusing on Asian eye model are conducted at the Eye Clinic, Department of Ophthalmology, Hospital Melaka, Malaysia. Meanwhile, the third process which is the eye anatomy voice-over recording in order to record the correct pronunciation of the eye anatomy for the research product is conducted at the Multimedia Studio, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka. The recording is conducted in the studio due to the facility and the environment provided which suitable for the audio recording in order to produce the best output. Table I shows the details of data collection process. Meanwhile Table II, Table III and Table IV present the data collection summary for the first, second and third content verification process with the expert, respectively.

TABLE I. Content verification

Verification	Subject Matter Expert
Contents for eye anatomy features	Consultant Ophthalmologist, Hospital Melaka
3D eye model –focusing on Asian	
Voice-over recording for eye anatomy pronunciation	

TABLE II. Content verification SESSION 1

Features	Description
Location	Eye Clinic, Department of Ophthalmology, Hospital Melaka
Date/Duration	2 November 2017
Time	3.00 pm – 5.00 pm
Data Collection Method	Content Verification Session 1 Discussion with the Consultant Ophthalmologist
Discussion	• Present 3D human eye model version 1 for expert verification • Explanation on the structures of eyes • Expert suggestions: - Once the eye model rotates, the labels of the eye anatomy appear - In addition to the glaucoma and diabetic retinopathy, age related macular degeneration can be added as another eye disease

TABLE III. Content verification SESSION 2

Features	Description
Location	Eye Clinic, Department of Ophthalmology, Hospital Melaka
Date/Duration	11 November 2017
Time	5.30 pm – 7.00 pm
Data Collection Method	Content Verification Session 2 Discussion with the Consultant Ophthalmologist
Discussion	• Present 3D human eye model version 2 for expert verification • Show the eye anatomy labelling • Expert suggestions: - Make the label line a bit tiny - Add important anatomy, such as optic disc, ciliary body, suspensory ligament - Make a correction color for these anatomy: Sclera (white) Choroid (pink) Retina (yellow) - Follow the correct measurement for each of the anatomy, must be proportional - Make a correction on the design of lens (bigger) and iris (a little bit curve) - In addition to superior and inferior views, add lateral view

TABLE IV. Content verification SESSION 3

Features	Description
Location	Multimedia Studio, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka
Date/Duration	3 April 2018
Time	3.00 pm – 5.00 pm
Data Collection Method	Content Verification Session 3 Discussion with the Consultant Ophthalmologist
Discussion	• Audio recording for the Interactive Ophthalmology Hologram introduction • Eye anatomy pronunciation recording

B. Design

During the design phase, the product is designed to satisfy the requirements that identified in the previous analysis phase. In this phase, the human eye is modelled in 3D. The physical eye model is used as a guideline in order to model an accurate 3D eye model. In addition, the input from the Consultant Ophthalmologist also is considered throughout the design process of the 3D human eye model. The 3D human eye model is designed and developed according to the Asian human eye. This is another contribution of this research as other 3D human eye model is based on Caucasian eye. Fig. 2 shows the physical human eye model as a referral for designing the 3D human eye model. Meanwhile Fig. 3 shows the 3D human eye model referred. Fig. 4 to Fig 10 show some of the interfaces of this research product. Fig. 4 shows the interface of the eye anatomy module. Meanwhile, the interfaces of the eye anatomy which includes choroid and sclera are presented in Fig. 5 and Fig. 6 respectively. The interfaces of the eye cross section from four different views are shown in Fig. 7 to Fig. 10.

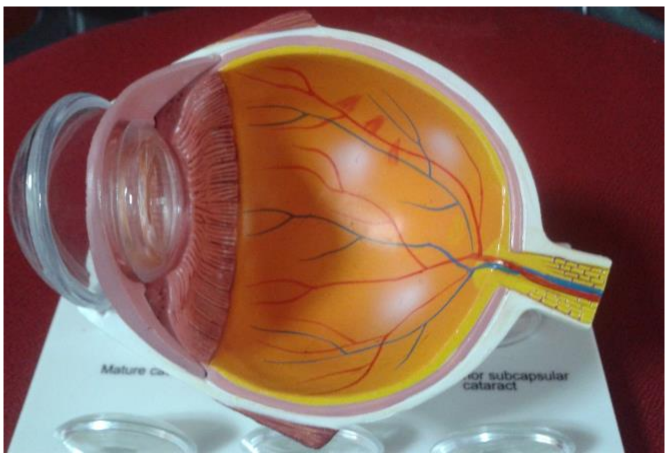


Fig. 2. Human Eye Model.

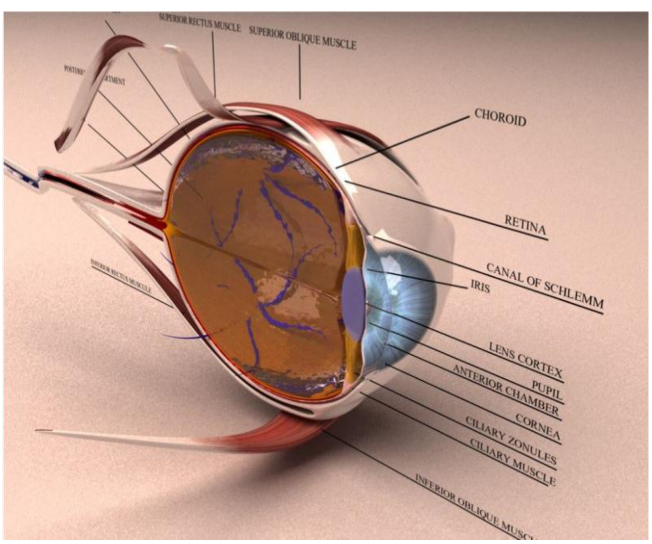


Fig. 3. 3D human Eye Model [10].

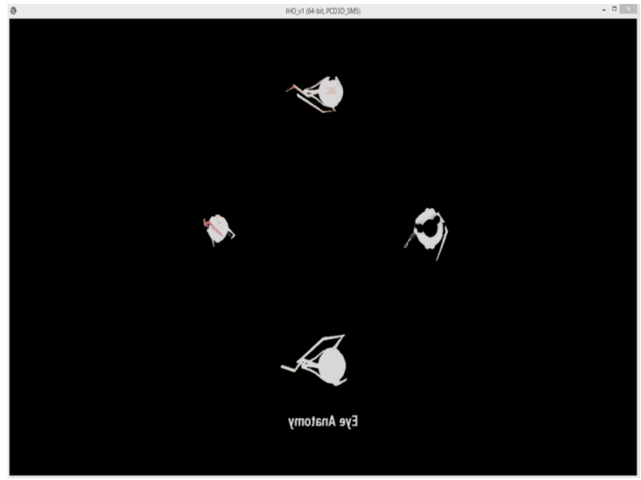


Fig. 4. Interface of Eye Anatomy Module.

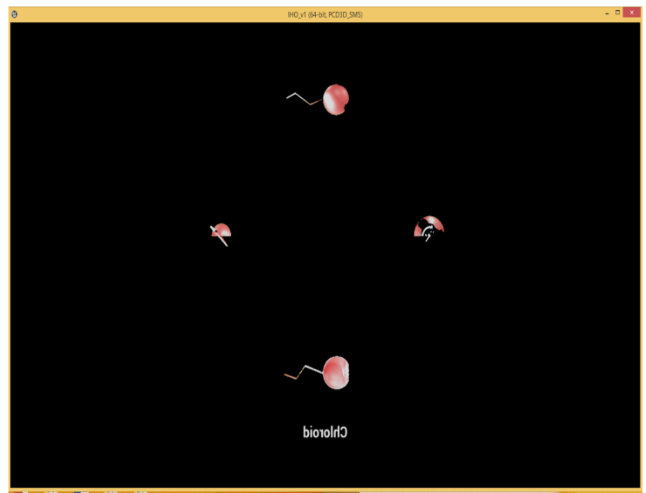


Fig. 5. Interface of Eye Anatomy Module – Choroid.

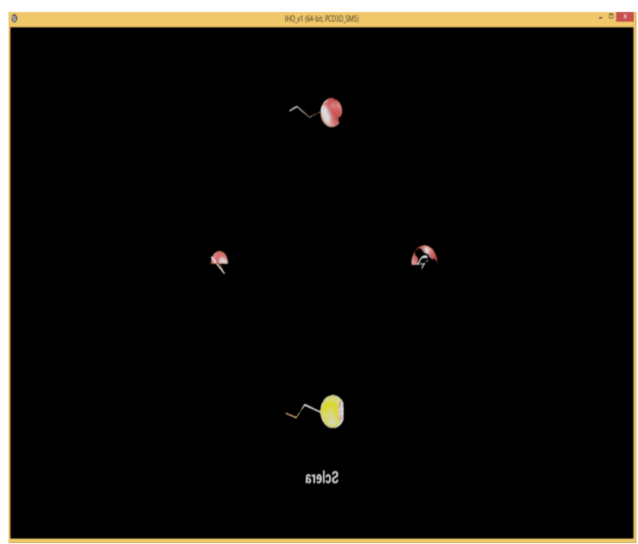


Fig. 6. Interface of Eye Anatomy Module – Sclera.

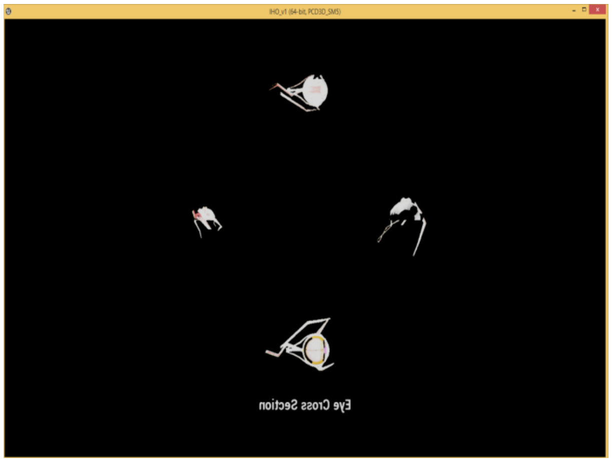


Fig. 7. Interface of Eye Cross Section View 1.

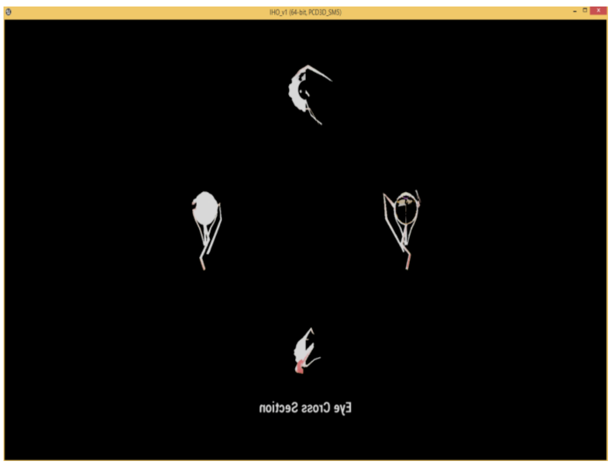


Fig. 8. Interface of Eye Cross Section view 2.

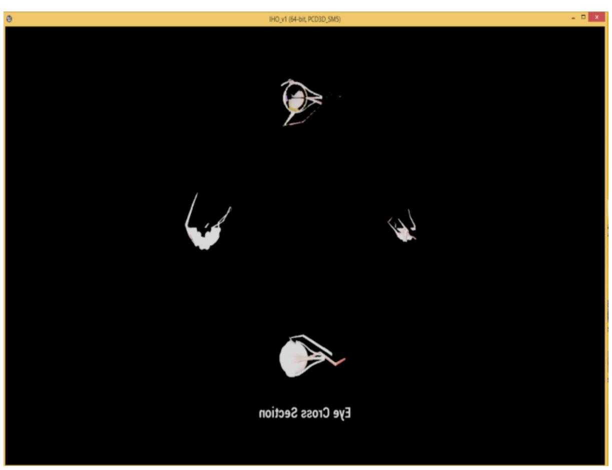


Fig. 9. Interface of Eye Cross Section View 3.

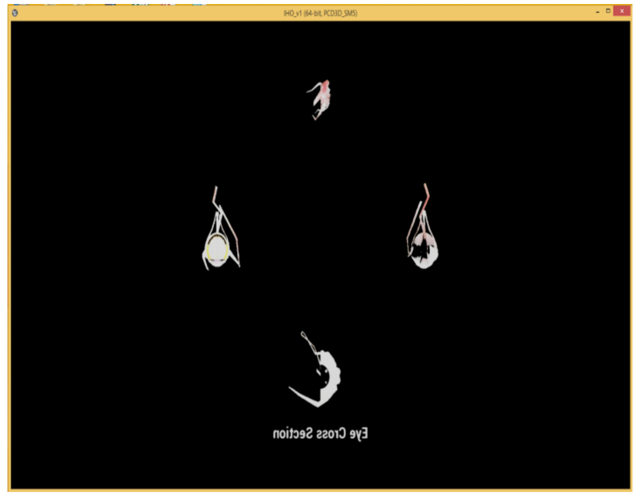


Fig. 10. Interface of Eye Cross Section View 4.

C. Development

Interactive Ophthalmology Hologram (O-iHO) has been developed by integrating a game engine with interactivity features. The product is capable to view the human eye anatomy from different perspectives, including the front, back, left and right sides. The human eye model also can be viewed in different cross section views. Besides representing the eye model in real time 3D graphic visualization, the product also able to provide real time interaction capability. The product can be interacted with devices such as keyboard, mouse, joystick and other input devices. Moreover, hand gesture and motion sensor are possible to interact with in order to produce an exciting experiential learning. The prototype of the product is presented in Fig. 11. Meanwhile, the product features developed are presented in Fig. 12 to Fig. 15. Fig. 12 shows the introduction screen of the Interactive Ophthalmology Hologram. Meanwhile, Fig. 13 presents one of the product features, which is the real time interaction capability. In addition, the other features which are interaction with hand gesture or motion sensor and also real time 3D graphic visualization are shown in Fig. 14 and Fig. 15 respectively.

The graphics or images used in this product are in 3D form to provide better visualization and details as the human eye is complex. The 3D image is designed by using Cinema 4D software. The audio is recorded by using audio recording equipment and then edited. The audio of the first page mentioning the title and the purpose of the product is provided once the user starts using the product. Meanwhile, the user can listen to the eye anatomy pronunciation for each of the eye structures every time when the eye features presented. The video of the proposed product which shows the implementation and details of the final output is presented in [11-12].

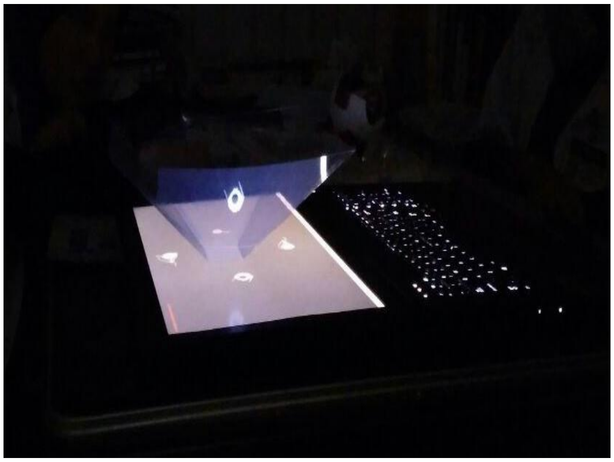


Fig. 11. Prototype Demonstration.

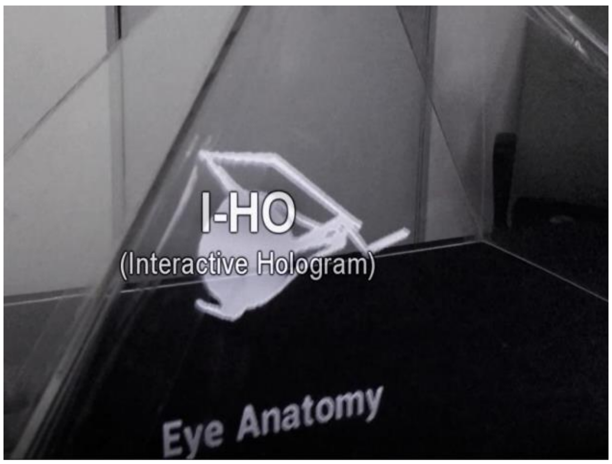


Fig. 12. Interactive Ophthalmology Hologram Introduction Screen.

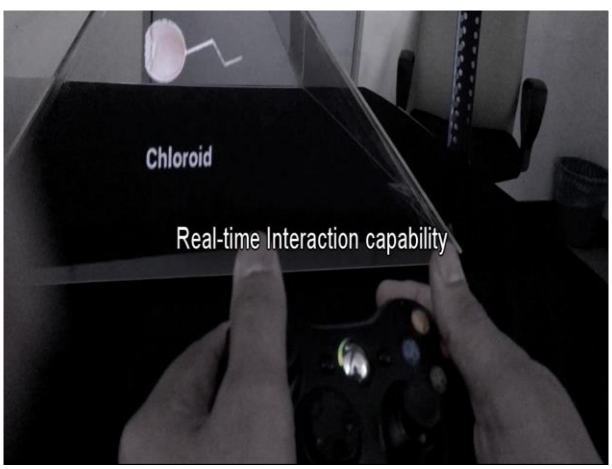


Fig. 13. Real Time Interaction Capability Feature.

Fig. 14. Interaction Feature.

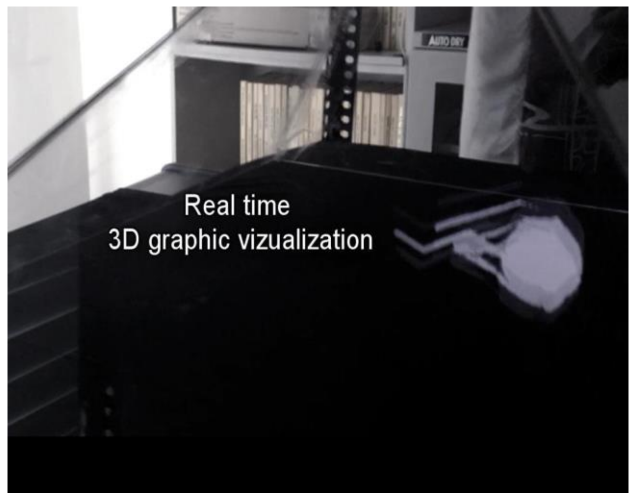


Fig. 15. Real Time 3D Graphic Visualization Feature.

D. Evaluation

Interactive Ophthalmology Hologram has been developed according to the objectives underlined. The product has been evaluated by the expert continuously. For every verification process, the product has been amended and improved based on the expert suggestions and comments in order to come up with a good output. The product has been presented to a big group of ophthalmologists during the National Glaucoma Update 2018 launching ceremony organized by the Department of Ophthalmology, Hospital Melaka, Malaysia on 27 April 2018 [12]. It was a great honoured to present the product to the target user and the responses received were positive and impressive. The product will be enhanced with other future improvements in order to provide a complete ophthalmology hologram.

IV. Discussion

This research product is attractive as it consists of many multimedia elements and interactive capabilities, which able to attract the user attention. In addition, the characters of the animation are well designed for presenting the eye anatomy information. Moreover, the procedure of the product is designed nicely to make it ease of use for the users. The users can easily understand the information presented and how to interact with. The most important issue is the information presented is accurate since it is verified by the Subject Matter Expert which is the Consultant Ophthalmologist.

However, this research initially covers only one important module in ophthalmology field, which is the eye anatomy. By adding more modules, users can interact more with the application and their understanding towards ophthalmology such as eye anatomy and eye diseases can be increased. The explanation of each of the eye structures which have been verified by the expert will be added in order to provide more knowledge and understanding to the user. Moreover, users can easily understand each of the eye structures if the explanation is simple and understandable. The eye diseases module should be added for a complete content for ophthalmology hologram. Glaucoma, diabetic retinopathy and age related macular degeneration occurrence can be added as they are among the common eye disease. The holographic approach will help the medical doctors in conveying the information of the eye diseases, especially the complex signs of the eye diseases. For example, diabetic retinopathy is one of major complications of diabetes mellitus which causes blindness. Diabetic retinopathy is composed of a several of lesions found in the retina of individuals having diabetes. Therefore, the interest of this holographic approach could substitute current explanation procedures, eventually contribute to producing a more reliable and understandable explanation to the patients.

V. Conclusion

In conclusion, the interactive hologram for ophthalmology has developed successfully. The objectives have been achieved where the interactive content development provides the main ophthalmology information to the users. It is envisaged that the developed interactive hologram would greatly contribute to increase the understanding and awareness to the public especially patients. It is hoped that the developed application will assist the medical doctors, medical students and community in delivering the medical information efficiently.

This research contributes to the medical team, especially medical lecturers and medical students by increasing their understanding on eye anatomy and eye diseases. Furthermore, the product is not merely produced for the medical field, but also for education. Since holographic is a new advanced technology, the multimedia lecturers and students are benefited in exploring and learning this new technology. Moreover, this research provides a significant contribution to knowledge in the area of multimedia system. This is achieved by proposing a novel interactive hologram for ophthalmology.

Acknowledgment

This research is a collaboration between Universiti Teknikal Malaysia Melaka and Department of Ophthalmology, Hospital Melaka, Malaysia. This product won Gold Award Winner in the Innovation Carnival: UTeMEX 2017 organized by Universiti Teknikal Malaysia Melaka and Bronze Medal at the 29th International Invention & Innovation Exhibition (ITEX) 2018.

The product “Interactive Medical Hologram : Ophthalmology” is Universiti Teknikal Malaysia Melaka copyrighted (LY2018000965).

References

[1] D. Gabor, “A New Microscopic Principle”, *Nature* vol. 161, no. 4098, pp. 777-778, 1948

[2] H. Abbasi, T. Zalie, N. J. Farahani and A. G. Rad, “Studying the recent improvements in holograms for three-dimensional display”, *International Journal of Optics*, vol. 2014, pp. 1-7, 2014

[3] S. Oliveira and M. Richardson, “The future of holographic technologies and their use by artists”, *J. Phys.: Conf. Ser.*, vol. 415, pp. 1-5, 2013.

[4] Euclideon Holographics, World’s First Multi-User Hologram Table, available online: http://www.euclideonholographics.com, last visit 27.03.2018

[5] N. Abdullasim, W. S. N. Saifudin and S. S. Rahim, ‘Development of interactive hologram system’, Proceedings of Innovative Research and Industrial Dialogue 2018, 2018

[6] V. Bally, Holography on medicine and biology, Proceedings of the International Workshop, Munster, Fd. Rep. Of Germany, March 14-15, 1979, Springer-Verlag Berlin Heidelberg, 1979

[7] Integraf, Medical Applications of Holography, available online: https://www.integraf.com/resources/articles/a-medical-applications-of-holography, last visit: 27.10.2018

[8] M. C. Potcoava, Digital holography applications in ophthalmology, biometry, and optical trapping characterization, BiblioBazaar, South Carolina, United States, 198 pages, 2011

[9] M. K. Kim, “Applications of digital holography in medical microscopy”, *Journal of the Optical Society of Korea*, vol. 14, no. 2, pp. 77-89, 2010

[10] TurboSquid, Human anatomy - eye section, available online: https://www.turbosquid.com/3d-models/3d-model-human-eye-section--/677870, last visit: 1.11..2017

[11] Youtube, iHO: Interactive ophthalmology hologram, available online: https://www.youtube.com/watch?v=ysMPcXiaM8Q, last visit: 25.7.2018

[12] Youtube, iHO: Interactive ophthalmology hologram gimmick, available online: https://www.youtube.com/watch?v=ehV9IWxgDS4, last visit: 25.7.2018.

Efficient Page Collection Scheme for QLC NAND Flash Memory using Cache

Seok-Bin Seo[1], Wanil Kim[2], Se Jin Kwon[3]
Department of Computer Engineering
Kangwon National University
Samcheok, South Korea

***Abstract*—Recently, semiconductor companies such as Samsung, Hynix, and Micron, have focused on quad-level cell (QLC) NAND flash memory chips, because of the increase in the capacity of storage systems. The QLC NAND flash memory chip stores 4 bits per cell. A page in the QLC NAND flash memory consists of 16 sectors, which is two to four times larger than that of conventional triple-level cell flash NAND flash memory. Because of its large page size, when the QLC NAND flash memory is applied to the current storage system directly, each page space is not efficiently used, resulting in low space utilization in overall storage systems. To solve this problem, an efficient page collection scheme using cache for QLC NAND flash memory (PCS) is proposed. The main role of PCS is managing the data transmitted from the file system efficiently (according to the data pattern and size), and reducing the number of unnecessary write operations. The efficiency of PCS was evaluated using SNIA IOTTA NEXUS5 trace-driven simulation on QLC NAND flash memory. According to close observation, PCS significantly reduces 50% of write operations compared with previous page collection algorithms, by efficiently collecting the small data into a page. Furthermore, a cache idle-time determination algorithm is proposed to further increase the space utilization of each page, thereby reducing the overall number of write operations on the QLC flash memory.**

Keywords—Solid state drive; storage systems; cache; flash translation layer

I. Introduction

In recent years, the solid-state drive (SSD) market has grown rapidly owing to the capacity increase of storage systems. Accordingly, semiconductor companies such as Samsung, Hynix, Micron, and Toshiba, are carrying out research on low-cost, high-capacity quad-level cell (QLC) NAND flash memory chips. Single-level cell (SLC) NAND memory stores a single bit in a single cell, and multilevel cell (MLC) NAND flash memory stores two bits in a single cell. However, QLC NAND flash memory stores four bits in a single cell; thus, the capacity can be increased more compared with SLC/MLC/tri-level cell (TLC) NAND flash memories NAND flash memories commonly consist of a "page" (a set of sectors) in read/write operation units, and several pages are gathered and compose a "block," which is a unit of erase operation. Furthermore, they have an "erase-before-write" property, i.e., an erase operation must be performed on a block prior to performing data renewal of a page [1].

This work was supported by Basic Science Research through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3B04031440). This study was also supported by 2017 Research Grant from Kangwon National University (No. 000000000).

The biggest differences between QLC NAND flash memory and SLC/MLC/TLC NAND flash memories are the page/block size and performance. In a QLC NAND flash memory, a page consists of 16 sectors, which is larger than that of conventional SLC/MLC/TLC NAND flash memory. Consequently, the read/write/erase operations of QLC NAND flash memory are slow compared with those of SLC/MLC/TLC NAND flash memory. Therefore, a page of QLC NAND flash memory is larger than that of SLC/MLC/TLC NAND flash memory, and its overall performance is reduced. Consequently, when the read/write pattern traces of the file system used in conventional operation systems are applied to a QLC NAND flash memory, the space utilization is decreased, and the number of write operations is increased, and this can be confirmed through a performance evaluation.

In this research, the problem of unnecessary write operations occurring in a QLC NAND flash memory is solved by implementing a page collection method in a flash translation layer [2]. When the file system issues a read/write requests along with its logical addresses, Page Collection Scheme stores the data in a temporary page storage ("cache" or "register") according to its page collection algorithm. Then, the collected data in the cache is transferred to the address mapping layer to be written onto the flash memory. The main issue of this paper is to solve the inefficiency caused by the large page size of QLC NAND flash memories. Therefore, the scope is limited to Page Collection Scheme. However, please notice that Page Collection Scheme can co-exist with any address mapping algorithms, since its main role is to increase interchangeability between the address mapping layer and QLC NAND flash memory.

The proposed Page Collection Scheme was tested through SNIA IOTTA's Nexus 5 Smartphone Traces-based simulation, and the results showed a large increase in the number of pages that had good space utilization and a decrease in the number of write operations, as discussed in Section 3.The rest of this work is structured as follows. Section 2 shows the characteristics of flash memories, the result of using data patterns for the read/write requests of a conventional host in the NAND flash memories that emerged before QLC NAND flash memory, and the problem occurring when the same patterns are applied to QLC NAND flash memory. In Section 3, the proposed Page Collection Scheme algorithm is described. In Section 4, the comparative performance evaluation performed through the SNIA IOTTA Nexus5 Smartphone Traces simulation based on the proposed Page Collection Algorithm is described. In Section 4, a future study is outlined, and a conclusion is provided.

TABLE I. NAND FLASH MEMORY WITHOUT PAGE COLLECTION SCHEME ALGORITHMS

Device / Request	*SLC NAND Flash Memory*	*MLC NAND Flash Memory*	*QLC NAND Flash Memory*
Number of pages with space utilization greater than 50%	234,543 Pages	234,544 Pages	52,620 Pages
Number of pages with space utilization of 50% or less	38,024 Pages	38,052 Pages	219,448 Pages

II. PROBLEM DEFINITION

SLC NAND flash memory has the characteristics that an operation of 1 bit per cell is performed, and the 2 KB data area exists per page. In an MLC NAND flash memory, an operation of 2 bits per cell is performed, and a 4- KB data area exists per page [3] [4]. In a QLC NAND flash memory, an operation of 4 bits per cell is performed, and there is an 8 KB data area per page.

Experiments were performed through SNIA IOTTA Nexus5 Smartphone Traces simulation without applying the Page Collection Scheme to the SLC and MLC NAND flash memories. In the results, the percentage of pages that had 50% or lower space utilization in a page was 14% (38,024 and 38,025, respectively), as shown in Table 1, and there was no occurrence of problems caused by space utilization in each page. In the results of performing the simulation for QLC NAND flash memory by using the same method, each page showed a low space utilization, as shown in Table 1. Among 272,568 pages, the number of pages having 50% or lower space utilization was 219,448 (80%).

Algorithm 1. Page Collection Scheme Algorithms

INPUT: write(LPN, data, SIZE);
1.WAIT: IF 'write request' IS EQUAL 0 GOTO WAIT
IF 'write *date* size' IS BIGGER THAN 6KB THEN
IF 'LPN' IS EQUAL 'Register' THEN
'*data* write in flash memory' AND
'clear register'
5. ELSE '*data* write in flash memory'
6. ELSE '*data* write in Register
7. GOTO WAIT

The detailed comparative analysis in Table 1 revealed the reason why a page could not be filled with data by exceeding 50% space utilization in the QLC NAND flash memory, compared with the SLC and MLC NAND flash memories. Even when the data was applicable for an update, or when a write command requested from the file system was of 16 or more sectors, the starting or ending part with respect to the data size of each sector command did not fully fill a page. To increase the low space utilization of these pages that are operated inefficiently, a page collection method using a cache is proposed.

III. PAGE COLLECTION SCHEME

The page collection method using a cache proposed in this paper is assumed to use the double cache structure described in Fig. 1. Moreover, a random in/out method is used to store and modify data in the resister of a double-cache structure [6]. Furthermore, because the QLC NAND flash memory is based on the double-cache structure, it is assumed that the size of the cache or resister is the size of one page of QLC NAND flash memory.

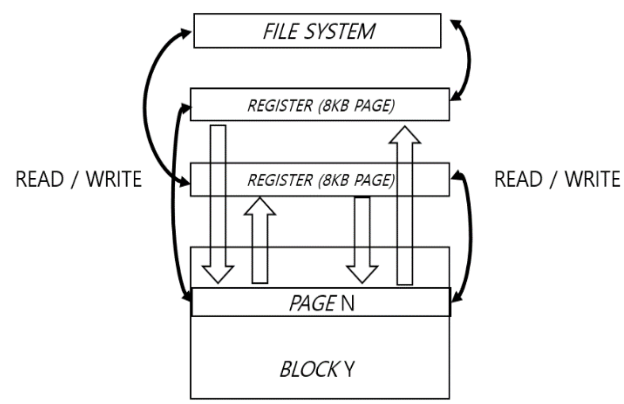


Fig. 1. Double Cache Structure [5].

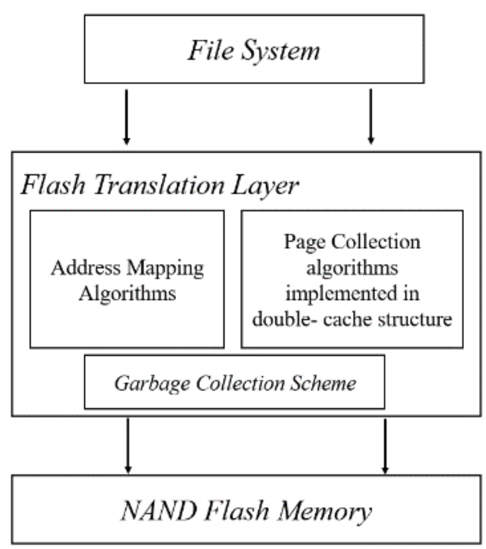


Fig. 2. FTL Structure with Double-Cache Structure.

In addition, it is assumed that the Page Collection Scheme method and the FTL structure are included in the Page Collection Scheme algorithm having a double-cache structure, as shown in Fig. 2. Based on the above assumption, Algorithm 1 is proposed as follows: when a write request of the OS file system is mage in FTL, the size is determined, and a data write request smaller than a certain size always resides in the cache, ensuring the size of a page will be as close to 8 KB (i.e., 16 sectors per page based on the 512-byte size) as possible.

First, when a write command of the file system is called, the data amount of one page that can be accommodated by the resister is assumed to be less than 6 KB (Lines 1 and 2 of the algorithm). If data of 6 KB or larger come in, they are checked to determine whether a corresponding resister exists, and, if there is a corresponding resister, the write operation of data to the flash memory is performed, and the corresponding resister is emptied for a new write.

IV. Performance Evaluation

The proposed Page Collection Scheme method was implemented based on Algorithm 1 proposed above, and then evaluated. As an outcome of using all the 31 traces provided by IOTTA's Nexus5 Smartphone Traces, Table 2 shows the results of the experiment for the pages that had an increased space utilization compared with the conventional method. In the results of the experiment, when the Page Collection Scheme was not applied, there were 219,948 pages that had low space utilization, as shown in Table 1, but when the proposed Page Collection Scheme method was applied, the number of pages showing low utilization was 110,936, displaying the decrease effect from approximately 80% to 50%.

However, despite the existence of the cache, many pages still had a problem regarding the space utilization. A register exists in a wait state continuously when the resister is not applicable for an update. Accordingly, an algorithm that decides the idle time of the cache is proposed to solve the above problem.

TABLE II. QLC NAND flash memory with Page Collection Scheme Algorithms

Algorithms / *Request*	*QLC NAND Flash Memory Without PCS Algorithms*	*QLC NAND Flash Memory With PCS Algorithms*
Number of Pages over 6 KB	52,620 Pages	126,836 Pages
Number of pages less than 6 KB	219,448 Pages	110,936 Pages

V. Future Research Activities

To overcome the limitations of the proposed page collection method using cache, the idle time of the OS of the system is decided. Further, when it is in the category of idle time, the data in the cache are moved to the NAND flash memory, and the corresponding cache is emptied. This way, it is expected that the number of pages having more active and higher space utilization will be increased compared with the conventional methods.

The page collection method that decides the idle time of the cache is an algorithm that decides the idle time of the device or decides the idle time by counting the number of times that the cache was not updated after the write command of the file system in the OS. The algorithms constructed are shown in Algorithm 2 and Algorithm 2-1. In Algorithm 2, a variable was added to Algorithm 1 to check the update status of the cache (Line 5 of Algorithm 2). After the conditions of Algorithm 2 are satisfied, Algorithm 2-1 is performed. If the number of "not-updating the cache" is larger than a certain value or the device is in an idle-time state, the data existing in the cache are moved to the flash memory, and the corresponding memory is emptied for a new write (Lines 1-4 of Algorithm 2-1).

The cache idle-time decision algorithm is different from the conventional algorithm, in which a resister is continuously remaining in the wait state to obtain an expected result even when the page will no longer be updated. The algorithm that decides the idle time of the cache leads to a more efficient use of QLC NAND flash memory overall by moving data to the NAND flash memory according to the criteria, instead of waiting for an update of the resister indefinitely. To prove that the above results of the experiment are significant results, an experiment will be conducted on the current embedded board in future.

Algorithm 2. Cache idle-time decision algorithms
1. WAIT: IF 'write request' IS EQUAL 0 GOTO WAIT 2. IF 'write *data* size' IS BIGGER THAN 6 KB THEN 3. IF 'LPN' IS EQUAL 'register' THEN '*data* write in flash memory' AND 'check register()' AND 'clear register' 7. ELSE '*data* write in flash memory' 8. ELSE '*data* write in register' 9. GOTO WAIT

Algorithm 2-1. Cache idle time decision algorithms
<CHECK REGISTER()> IF 'register non-update' IS MORE THAN $\mathcal{I}$ OR 'device' IS EQUAL 'idle-time' THEN '*data* write in flash memory AND 'clear register' AND RETURN

VI. Conclusion

In this work, a page collection method that uses cache was investigated for QLC NAND flash memory. In the SLC and

MLC NAND flash memories, the problem of space utilization in each page was not found with respect to the write requests of the conventional file system. However, in the QLC NAND flash memory, problems occurred in many pages because of low space utilization. The proposed page collection method using the cache ensures appropriate space utilization by using a resister, and it performs the write request to the NAND flash memory instead of sending the data directly to the NAND flash memory according to the write request of the file system. In the results of trace-based simulation, the QLC NAND flash memory that applied the proposed method exhibited a decrease in the number of total write operations and an increase in the number of pages having high space utilization. Furthermore, in the future, a study is planned on the page collection method considering the cache idle-time decision, and more pages of QLC NAND flash memory are expected to have high space utilization.

References

[1] M.A. Dongzhe FENG, Jianhua; LI, Guoliang, "A survey of address translation technologies for flash memories," *ACM Computing Surveys* (CSUR), 2014.

[2] E. Gal and S. Toledo, "Algorithms and Data Structures for Flash Memories," *ACM Computing Surveys*, 37(2): 138-163, 2005.

[3] MICRON Electronics, "NAND Flash Memory MT29F4G08AAAWP, MT29F8G08BAAWP, MT29F16G08FAAWP,"MICRON Electronics Data sheet, 2018.

[4] TOSHIBA Memory, "Toshiba Mos Digital Integrated Circuit Silicon Gate CMOS, TH58NVG4S0HTAK0,"TOSHIBA Memory Data Sheet, 2017.

[5] MICRON Electronics, "NAND Flash Status Register Response in Cache Programming Operations," MICRON Electronics Technical Notes, 2018.

[6] Se Jin Kwon, Arun Ranjitkar, Young-Bae Ko, Tae-Sun Chung, "FTL algorithms for NAND-type flash memories," Design Autom. for Emb. Sys. 15(3-4), pp. 191-224, 2011.

Edge Detection on DICOM Image using Triangular Norms in Type-2 Fuzzy

D. Nagarajan[1], M.Lathamaheswari[2]
Department of Mathematics
Hindustan Institute of Technology & Science, Chennai-603 103, India

R.Sujatha[3]
Department of Mathematics
S.S.N College of Engineering
Chennai-603 110, India.

J.Kavikumar[4]
Department of Mathematics and Statistics
Universiti Tun Hussein Onn, Malaysia

***Abstract*—In image processing, edge detection is an important venture. Fuzzy logic plays a vital role in image processing to deal with lacking in quality of an image or imprecise in nature. This present study contributes an authentic method of fuzzy edge detection through image segmentation. Gradient of the image is done by triangular norms to extract the information. Triangular norms (T norms) and triangular conorms (T conorms) are specialized in dealing uncertainty. Therefore triangular norms are chosen with minimum and maximum operators for the purpose of morphological operations. Also, mathematical properties of aggregation operator to represent the role of morphological operations using Triangular Interval Type-2 Fuzzy Yager Weighted Geometric (TIT2FYWG) and Triangular Interval Type-2 Fuzzy Yager Weighted Arithmetic (TIT2FYWA) operators are derived. These properties represent the components of image processing. Here Edge detection is done for DICOM image by converting into 2D gray scale image, using Type-2 fuzzy MATLAB and which is the novelty of this work.**

Keywords*—*Aggregation operators; T norm; T conorm; triangular interval type-2 fuzzy number (TIT2FN); fuzzy morphology; gray scale Image; medical image processing

I. Introduction

In the field of optimization problems in Mathematics, Statistics, Economics and Information Science, the max and min operators are very useful for any dimension. Uncertainty convoluted in most of the real world problems. Fuzzy theory has been developed as an efficient and powerful mechanism in mathematical design of many engineering and objective phenomena [1-5].To deal uncertainty in any field one needs an effective and predictable incentive. Usually incomplete data and errors in the analyzing stage will be the reason for getting vague situation and this can be dealt with fuzzy theory. Mathematical devices may figure out an impreciseness. The largest and the smallest elements of a precise set of real numbers is the maximum and the minimum and so Yager triangular norm is chosen for this work [6-10]. We are facing many problems to add, melt and synthesize the datum from different sources to get a conclusion. The operators may be chosen according to the characteristic properties and then the operations for minimum and maximum can be applied [11-14].

The triangular norms with maximum and minimum operators could be used for an image processing since these norms play as the synthesize operators for which these maximum and minimum operators are just an exclusive choice [15-17]. A Fuzzy Set (FS) is defined from a universal set to [0, 1] and the membership values (MVs) of every element is a crisp value between 0 and 1. This kind of system is called Type-1 Fuzzy Set (T1FS) system. In many of the real world problems it is necessary to have a MV itself fuzzy instead of crisp value which is called T2FS [18-19]. The generalization of union and intersection operators are triangular norms. Though the general case is important there is an equal important for the particular cases which provide efficient algorithm and more understanding missing in the general case [20].

T2FS is used when T1FS is blurred. In T2FS, the MVs lies in an interval so it is useful in image processing as many of the images are not properly visible. The parameter η in Yager triangular norms, accepts for tuning the norm between the other norms [21]. Yager norms covers all the continuous norms by changing the parameter where as other major norms can't do the same and have more time complexity [22]. In automation, visual sense, remotely second scene analysis and bio medical image processing, Fuzzy image analysis has been applied. When the images with low brightness, the structure will not be evidently visible. In this situation, the sets which have better and naturally include different types of uncertainties might be useful for image analysis in any field.

To deal this complication Fuzzy Sets and their advanced extensions like T2FS Sets are suitable since it handles the uncertainty in a better way. Using Type-2 Fuzzy thresholding techniques, different regions and abnormal lesions can be separated. Image processing can be done by FMM using triangular norms. Using T2FS, collection of undesirable scraps can be made while noise exist. In image processing, image enrichment, clustering, thresholding, edge detection and morphological image processing are easy to be done using T2FS. Application of single image analysis is always not reliable and therefore image processing based on T2 Fuzzy system has been considered [23, 24]. Borderline between two compatible regions is called an edge.

Using unit of the regional array, sense of the trial edge will be done at different points. Real world issues are levelheaded of various structures at various scales and an ideal image cannot be expected. The technique of selecting and detecting acute disruption in an image is called edge detection. DICOM is worn to store, transfer and pass on the medical images (MIs). Most of the MIs are saved in DICOM pattern where one can store data of an image and header as well and per file there is one slice in general. Singe color images are called gray scale which accommodate the knowledge of only gray level but not about color. Every pixel has some number of bits that determines available number of various gray levels [25-30].

The paper is organized in the following manner. In section II, literature review has been done related to the present work. In section III, basic definitions required for developing the concept have been described. In section IV, operational laws have been proposed for TIT2FN. In section V, aggregation properties have been proved using weighted arithmetic and geometric operators. In section VI, the theory of image processing and the role of T2FS and Yager norms is presented. In section VII, applied Type-2 fuzzy logic in edge detection for DICOM image in two dimensional through MATLAB. In section VIII, conclusion and future work is given.

II. Review of Literature

The authors of, [1] described Aggregation operators elaborately with their advanced direction and applications. [2] explained about gathering of the information and its related aggregation operators. [3] proposed Frank Aggregation Operators (AOs) and its mathematical properties for TIT2FSs and applied in a decision making problem. [4] studied t norms of Yager and Hamacher and also metric space on fuzzy logic. [5] utilized AOs in the process of decision making under the environment of probabilistic fuzzy.

[17] proposed fuzzy image processing (FIP) using Dubois and Prade triangular norm. [22] proposed a methodology for an image condensation and rehabilitation on a Lossy image using fuzzy relational equations. [23] proposed a technique for image analysis with the application of morphological operators with the support of uninorms.

[24] described and explained very clearly about the role of theoretical fuzzy logic strategies in medical image processing. [25] reviewed the applications of type-2 fuzzy systems in the field of image processing. [26] presented a comprehensive depiction of imitation of an image with the help of fuzzy logic. [27] established an algorithm for edge detection under fuzzy environment where instability of a digital image for every pixel has been calculated.

[28] proposed a methodology for fusion of image under intuitionistic fuzzy setting. [29] introduced a new technique for edge detection with the support of representation of fuzzy image and pixels. [30] examined and done a comparative analysis of various techniques of edge detection. From this review it is found that there is no work has been done for edge detection on DICOM image using Type-2 fuzzy logic. This is the motivation of the present work.

III. Basic Definitions

The following basic concepts are given for the better understanding of the paper.

A. Aggregation Operator [3]

Let $(M_\alpha)_{\alpha\in[0,1]}$ be a group of aggregation operators (AOs) which is non-decreasing. If A is an AO then

$$M_A : \bigcup_{n\in N} [0,1]^n \to [0,1].$$

B. Triangular Interval Type-2 Fuzzy Set (TIT2FS) [3]

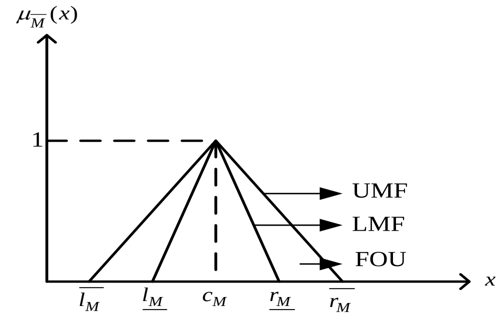


Fig. 1. TIT2FS.

The membership function (MFs) are developed using triangular fuzzy number in IT2FS called TIT2FS. In IT2FS, upper and lower MFs represented by a triangular fuzzy number $\overline{M} = \left\langle [\underline{l_M}, \overline{l_M}], c_M, [\underline{r_M}, \overline{r_M}] \right\rangle$ called TIT2FS and are defined by

$$LMF_{\overline{M}}(x) = \begin{cases} \dfrac{x-\overline{l_M}}{c_M - \overline{l_M}} & , \quad \overline{l_M} \le x < c_M \\ 1 & , \quad x = c_M \\ \dfrac{x-\underline{r_M}}{c_M - \underline{r_M}} & , \quad c_M \le x < \underline{r_M} \\ 0 & , \quad otherwise \end{cases} \tag{1}$$

$$UMF_{\overline{M}}(x) = \begin{cases} \dfrac{x-\underline{l_M}}{c_M - \underline{l_M}} & , \quad \underline{l_M} \le x < c_M \\ 1 & , \quad x = c_M \\ \dfrac{x-\overline{r_M}}{c_M - \overline{r_M}} & , \quad c_M \le x < \overline{r_M} \\ 0 & , \quad otherwise \end{cases} \tag{2}$$

Where $\underline{l_M}, \overline{l_M}, c_M, \underline{r_M}, \overline{r_M}$ are the measuring points on TIT2FS satisfying $0 \le \underline{l_M} \le \overline{l_M} \le c_M \le \underline{r_M} \le \overline{r_M} \le 1$. If we consider X as a set of real numbers, a TIT2FS in X is called TIT2FN. The FOU is the area between lower and upper membership functions in figure 1. If $\underline{l_M} = \overline{l_M}, \underline{r_M} = \overline{r_M}$, then $UMF_{\overline{M}}(x)$ = $LMF_{\overline{M}}(x)$ for all the values of x in X, then the TIT2FS will become Type-1 case. Here FOU is the footprint of Uncertainty.

C. Ranking formula for TIT2FN [3]

Let $\overline{M} = \langle [A,B], C, [D,E] \rangle$ where $A = \underline{l_M}, B = \overline{l_M}, C = c_M, D = \underline{r_M}, E = \overline{r_M}$ be the TIT2FN. The ranking value is defined by

$$Rank(\overline{M}) = \left(\frac{A+E}{2} + 1 \right) \times \frac{A+B+D+E+4C}{8} \tag{3}$$

D. Yager Triangular Norms [7]

$\underset{Y}{\otimes}$ is Yager product (T Norm) and $\underset{Y}{\oplus}$ is a Yager sum (T conorm) and are defined as follows.

$$r \underset{Y}{\otimes} s = \max\left(1-[(1-r)^{\eta}+(1-s)^{\eta}]^{\frac{1}{\eta}},0\right), \eta>0, \ for\ all\ r,s \in [0,1]^2 \tag{4}$$

$$r \underset{Y}{\oplus} s = \min\left(\left(r^{\eta}+s^{\eta}\right)^{\frac{1}{\eta}},1\right), \eta>0, \ for\ all\ r,s \in [0,1]^2 \tag{5}$$

E. Triangular Interval Type-2 Fuzzy Yager Weighted Arithmetic (TIT2FYWA) Operator [3]

Consider a set of TIT2FNs and the operator $TIT2FYWA_{\varepsilon}: \Omega^n \to \Omega$ is defined by $TIT2FYWA_{\varepsilon}\left\langle \overline{M_1}, \overline{M_2},...,\overline{M_n}\right\rangle \ = \varepsilon_1 \underset{Y}{\bullet} \overline{M_1} \underset{Y}{\oplus} \varepsilon_2 \underset{Y}{\bullet} \overline{M_2} \underset{Y}{\oplus} ... \underset{Y}{\oplus} \varepsilon_n \underset{Y}{\bullet} \overline{M_n}$

and its weight vector is $\varepsilon = (\varepsilon_1, \varepsilon_2,...,\varepsilon_n)^T$ and the sum of the weight vectors is equal to 1, when $\varepsilon = \left(\frac{1}{n}, \frac{1}{n},...,\frac{1}{n}\right)^T$, triangular interval type-2 fuzzy Yager weighted arithmetic operator will become triangular interval type-2 fuzzy Yager arithmetic averaging operator of dimension n and is defined by

$$TIT2FYAA_{\varepsilon}(\overline{M_1}, \overline{M_2},...,\overline{M_n}) = \frac{1}{n} \underset{Y}{\bullet} \left(\overline{M_1} \underset{Y}{\oplus} \overline{M_2},..., \underset{Y}{\oplus} \overline{M_n}\right). \tag{6}$$

F. Triangular Interval Type-2 Fuzzy Yager Weighted Geometric (TIT2FYWG) Operator [3]

Let $\overline{M} = \left([\underline{l_{M_p}}, \overline{l_{M_p}}], c_{M_p}, [\underline{r_{M_p}}, \overline{r_{M_p}}]\right), \quad p=1,2,...,n$ be a set of TIT2FNs. Triangular *Interval Type-2 fuzzy Yager Weighted Geometric Mean Operator (TIT2FYWA),* TIT2FYWG:$\varepsilon^{n} \to \varepsilon$ is $\text{TIT2FYWG}_{\varepsilon}\left(\overline{M_1}, \overline{M_2},...,\overline{M_n}\right) = \overline{M_1}^{\bullet_Y \varepsilon_1} \underset{Y}{\otimes} \overline{M_2}^{\bullet_Y \varepsilon_2} \underset{Y}{\otimes} ... \underset{Y}{\otimes} \overline{M_n}^{\bullet_Y \varepsilon_n}$ .Here also sum of all weight vectors is equal to 1, when $\varepsilon = \left(\frac{1}{n}, \frac{1}{n},...,\frac{1}{n}\right)^T$, triangular interval type-2 fuzzy Yager weighted arithmetic operator will become triangular interval type-2 fuzzy Yager geometric averaging operator of dimension n and is defined by

$$TIT2FYGA_{\varepsilon}(\overline{M_1}, \overline{M_2},...,\overline{M_2}) = \ \frac{1}{n}\left(\overline{M_1} \underset{Y}{\otimes} \overline{M_2} \underset{Y}{\otimes},..., \underset{Y}{\otimes} \overline{M_n}\right)^{\hat{Y} \frac{1}{n}} . \tag{7}$$

IV. Proposed Operational Laws

Let $\overline{M}, \overline{M_1}, \overline{M_2}$ be three TIT2FNs and $\eta > 0$, then we define their operational laws as follows.

A. Addition

Consider, ${}_1 = \underset{p=1}{\overset{2}{sum}}\left(\underline{l_{M_p}}\right), B_1 = \underset{p=1}{\overset{2}{sum}}\left(\overline{l_{M_p}}\right), C_1 = \underset{p=1}{\overset{2}{sum}}\left(c_{M_p}\right),$

$D_1 = \underset{p=1}{\overset{2}{sum}}\left(\underline{r_{M_p}}\right), E_1 = \underset{p=1}{\overset{2}{sum}}\left(\overline{r_{M_p}}\right)$

$$\overline{M_1} \underset{Y}{\oplus} \overline{M_2} = \left(\left[\min\left(A_1^{\frac{1}{\eta}},1\right), \min\left(B_1^{\frac{1}{\eta}},1\right)\right], \ \min\left(C_1^{\frac{1}{\eta}},1\right), \left[\min\left(D_1^{\frac{1}{\eta}},1\right), \min\left(E_1^{\frac{1}{\eta}},1\right)\right]\right) \tag{8}$$

B. Multiplication

Consider,

$A_2 = \underset{p=1}{\overset{2}{sum}}\left(1-\underline{l_{M_p}}\right)^{\eta}, B_2 = \underset{p=1}{\overset{2}{sum}}\left(1-\overline{l_{M_p}}\right)^{\eta},$

$C_2 = \underset{p=1}{\overset{2}{sum}}\left(1-c_{M_p}\right)^{\eta},$

$D_2 = \underset{p=1}{\overset{2}{sum}}\left(1-\underline{r_{M_p}}\right)^{\eta}, E_2 = \underset{p=1}{\overset{2}{sum}}\left(1-\overline{r_{M_p}}\right)^{\eta}.$

$$\overline{M_1} \underset{Y}{\otimes} \overline{M_2} = \left\{\left[\max\left(1-A_2^{\frac{1}{\eta}},0\right), \max\left(1-B_2^{\frac{1}{\eta}},0\right)\right], \max\left(1-C_2^{\frac{1}{\eta}},0\right), \left[\max\left(1-D_2^{\frac{1}{\eta}},0\right), \ \max\left(1-E_2^{\frac{1}{\eta}},0\right)\right]\right\} \tag{9}$$

C. Multiplication by an ordinary number

Consider, $A = \underline{l_M}, B = \overline{l_M}, C = c_M, D = \underline{r_M}, E = \overline{r_M}$

$$k \underset{Y}{\oplus} \overline{M} = \left\{\left[\min\left\langle A^{\frac{k}{\eta}},1\right\rangle, \min\left\langle B^{\frac{k}{\eta}},1\right\rangle\right], \min\left\langle C^{\frac{k}{\eta}},1\right\rangle, \left[\min\left\langle D^{\frac{k}{\eta}},1\right\rangle, \min\left\langle E^{\frac{k}{\eta}},1\right\rangle\right]\right\} \tag{10}$$

D. Power

Consider

$A_3 = 1-\underline{l_M}, B_3 = 1-\overline{l_M}, C_3 = 1-c_M, \ D_3 = 1-\underline{r_M}, E_3 = 1-\overline{r_M}$

$$\overline{M}^{\hat{Y} k} = \left\{\left[\max\left(1-\left[A_3^{\eta}\right]^{k/\eta},0\right), \max\left(1-\left[B_3^{\eta}\right]^{k/\eta},0\right)\right], \right.$$

$$\max\left(1-\left[C_3^{\eta}\right]^{k/\eta},0\right),$$

$$\left[\max\left(1-\left[D_3{}^{\eta}\right]^{k/\eta},0\right),\max\left(1-\left[E_3{}^{\eta}\right]^{k/\eta},0\right)\right]\Bigg\} \quad (11)$$

V. Proposed Theorems

Here the mathematical properties of aggregation properties for TIT2FN using TIT2FYWG and TIT2FYWA operators are proved and they are playing an important role in image processing.

Consider a collection of TIT2FNs $\overline{M}=\left([\underline{l_{M_p}},\overline{l_{M_p}}],c_{M_p},[\underline{r_{M_p}},\overline{r_{M_p}}]\right),\quad p=1,2,...,n$

Where $0\le \underline{l_M}\le \overline{l_M}\le c_M\le \underline{r_M}\le \overline{r_M}\le 1$

A. Theorem

The aggregation value of these fuzzy numbers using TIT2FYWG operator is again a TIT2FN and

$$TIT2FYWG_{\varepsilon}\left\langle \overline{M_1},\overline{M_2},...,\overline{M_n}\right\rangle=\left\{\left[\max\left(1-\left[A_n{}^{\eta}\right]^{\varepsilon_p/\eta},0\right),\max\left(1-\left[B_n{}^{\eta}\right]^{\varepsilon_p/\eta},0\right)\right],\max\left(1-\left[C_n{}^{\eta}\right]^{\varepsilon_p/\eta},0\right),\left[\max\left(1-\left[D_n{}^{\eta}\right]^{\varepsilon_p/\eta},0\right),\max\left(1-\left[E_n{}^{\eta}\right]^{\varepsilon_p/\eta},0\right)\right]\right\} \quad (12)$$

Where the weight vector is $\varepsilon=\left(\varepsilon_1,\varepsilon_2,...,\varepsilon_n\right)^T,\varepsilon_n\ge 0$,

the sum of the weight vectors is equal to 1.

Proof:

Here use mathematical induction method.

Case (i): For $n=2$.

Consider, $P_1=\left(1-\underline{l_{M_1}}\right)^{\eta},Q_1=\left(1-\overline{l_{M_1}}\right)^{\eta},R_1=\left(1-c_{M_1}\right)^{\eta}$

$S_1=\left(1-\underline{r_{M_1}}\right)^{\eta},T_1=\left(1-\overline{r_{M_1}}\right)^{\eta}$

Using Yager power operation

$$\overline{M_1}^{\bullet_Y k}=\left\{\left[\max\left(1-P_1^{\frac{k}{\eta}},0\right),\max\left(1-Q_1^{\frac{k}{\eta}},0\right)\right],\max\left(1-R_1^{\frac{k}{\eta}},0\right),\left[\max\left(1-S_1^{\frac{k}{\eta}},0\right),\max\left(1-T_1^{\frac{k}{\eta}},0\right)\right]\right\}.$$

Consider, $P_2=\left(1-\underline{l_{M_1}}\right)^{\eta},Q_2=\left(1-\overline{l_{M_1}}\right)^{\eta},R_2=\left(1-c_{M_1}\right)^{\eta}$,

$S_2=\left(1-\underline{r_{M_1}}\right)^{\eta},T_2=\left(1-\overline{r_{M_1}}\right)^{\eta}$

$$\overline{M_2}^{\bullet_Y k}=\left\{\left[\max\left(1-P_2^{\frac{k}{\eta}},0\right),\max\left(1-Q_2^{\frac{k}{\eta}},0\right)\right],\max\left(1-R_2^{\frac{k}{\eta}},0\right),\left[\max\left(1-S_2^{\frac{k}{\eta}},0\right),\max\left(1-T_2^{\frac{k}{\eta}},0\right)\right]\right\}.$$

$$TIT2FYWG_{\varepsilon}\left(\overline{M_1},\overline{M_2}\right)=\overline{M_1}^{\bullet_Y \varepsilon_1}\underset{Y}{\otimes}\overline{M_2}^{\bullet_Y \varepsilon_2}$$

$$=\left\{\left[\max\left(1-\left[\sum_{p=1}^{2}\left(1-\max\left(1-\left(1-A_2\right)\frac{\varepsilon_p}{\eta},0\right)\right)\right],0\right),\max\left(1-\left[\sum_{p=1}^{2}\left(1-\max\left(1-\left(1-B_2\right)\frac{\varepsilon_p}{\eta},0\right)\right)\right],0\right)\right],\max\left(1-\left[\sum_{p=1}^{2}\left(1-\max\left(1-\left(1-C_2\right)\frac{\varepsilon_p}{\eta},0\right)\right)\right],0\right),\left[\max\left(1-\left[\sum_{p=1}^{2}\left(1-\max\left(1-\left(1-D_2\right)\frac{\varepsilon_p}{\eta},0\right)\right)\right],0\right),\max\left(1-\left[\sum_{p=1}^{2}\left(1-\max\left(1-\left(1-E_2\right)\frac{\varepsilon_p}{\eta},0\right)\right)\right],0\right)\right]\right\}.$$

$A_4=\left(1-\underline{l_{M_p}}\right)^{\eta},B_4=\left(1-\overline{l_{M_p}}\right)^{\eta},C_4=\left(1-c_{M_p}\right)^{\eta}$,

$D_4=\left(1-\underline{r_{M_p}}\right)^{\eta},E_4=\left(1-\overline{r_{M_p}}\right)^{\eta}$

$$=\left\{\left[\max\left(1-\left(\sum_{p=1}^{2}\left(1-\max\left(1-A_4{}^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right),0\right),\max\left(1-\left(\sum_{p=1}^{2}\left(1-\max\left(1-B_4{}^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right),0\right)\right],\right.$$

$$\max\left(1-\left(\underset{p=1}{\overset{2}{sum}}\left(1-\max\left(1-C_4^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right),0\right)\Bigg),$$

$$\Bigg[\max\left(1-\left(\underset{p=1}{\overset{2}{sum}}\left(1-\max\left(1-D_4^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right),0\right)\Bigg),$$

$$\max\left(1-\left(\underset{p=1}{\overset{2}{sum}}\left(1-\max\left(1-E_4^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right),0\right)\Bigg]\Bigg\}.$$

$$=\Bigg\{\Bigg[\max\left[1-A_2^{\frac{\varepsilon_p}{\eta}},0\right],\max\left[1-B_2^{\frac{\varepsilon_p}{\eta}},0\right]\Bigg],\max\left[1-C_2^{\frac{\varepsilon_p}{\eta}},0\right],$$

$$\Bigg[\max\left[1-D_2^{\frac{\varepsilon_p}{\eta}},0\right],\max\left[1-E_2^{\frac{\varepsilon_p}{\eta}},0\right]\Bigg]\Bigg\}$$

For $n=k,$

$$A_k=\underset{p=1}{\overset{k}{sum}}\left(1-\underline{l_{M_p}}\right)^{\eta},B_k=\underset{p=1}{\overset{k}{sum}}\left(1-\overline{l_{M_p}}\right)^{\eta},C_k=\underset{p=1}{\overset{k}{sum}}\left(1-c_{M_p}\right)^{\eta},$$
$$D_k=\underset{p=1}{\overset{k}{sum}}\left(1-\underline{r_{M_p}}\right)^{\eta},E_k=\underset{p=1}{\overset{k}{sum}}\left(1-\overline{r_{M_p}}\right)^{\eta}.$$

$$TIT2FYWG_{\varepsilon}\left\langle\overline{M_1},\overline{M_2},...,\overline{M_k}\right\rangle$$

$$=\Bigg\{\Bigg[\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-\left[A_k\right]^{\frac{\omega_p}{\eta}},0\right)\right)\right],0\right),$$

$$\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-B_k^{\frac{\omega_p}{\eta}},0\right)\right)\right],0\right)\Bigg],$$

$$\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-C_k^{\frac{\omega_p}{\eta}},0\right)\right)\right],0\right),$$

$$\Bigg[\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-D_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),$$

$$\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-E_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right)\Bigg]\Bigg\}.$$

$$=\Bigg\{\Bigg[\max\left[1-A_k^{\frac{\varepsilon_p}{\eta}},0\right],\max\left[1-B_k^{\frac{\varepsilon_p}{\eta}},0\right]\Bigg],$$

$$\max\left[1-C_k^{\frac{\varepsilon_p}{\eta}},0\right],\Bigg[\max\left[1-D_k^{\frac{\varepsilon_p}{\eta}},0\right],\max\left[1-E_k^{\frac{\varepsilon_p}{\eta}},0\right]\Bigg]\Bigg\}.$$

For $n=k+1,$

$$TIT2FYWG_{\varepsilon}\left\langle\overline{M_1},\overline{M_2},...,\overline{M_k}\right\rangle\underset{Y}{\otimes}TIT2FYWG_{\varepsilon}\left\langle\overline{M_{k+1}}\right\rangle$$

$$=\Bigg\{\Bigg[\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-A_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),$$

$$\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-B_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right)\Bigg],$$

$$\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-C_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),$$

$$\Bigg[\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-D_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),$$

$$\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-E_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right)\Bigg]\Bigg\}.$$

$$\underset{Y}{\otimes}\Bigg\{\Bigg[\max\left(1-P_{k+1}^{\frac{\omega_{k+1}}{\eta}},0\right),\max\left(1-Q_{k+1}^{\frac{\varepsilon_{k+1}}{\eta}},0\right)\Bigg],$$

$$\max\left(1-R_{k+1}^{\frac{\varepsilon_{k+1}}{\eta}},0\right),$$

$$\Bigg[\max\left(1-S_{k+1}^{\frac{\varepsilon_{k+1}}{\eta}},0\right),\max\left(1-T_{k+1}^{\frac{\varepsilon_{k+1}}{\eta}},0\right)\Bigg]\Bigg\}.$$

$$=\Bigg\{\Bigg[\max\left(1-\left[\underset{p=1}{\overset{k+1}{sum}}\left(1-\max\left(1-A_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),$$

$$\max\left(1-\left[\underset{p=1}{\overset{k+1}{sum}}\left(1-\max\left(1-B_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right)\Bigg],$$

$$\max\left(1-\left[\underset{p=1}{\overset{k+1}{sum}}\left(1-\max\left(1-C_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),$$

$$\Bigg[\max\left(1-\left[\underset{p=1}{\overset{k+1}{sum}}\left(1-\max\left(1-D_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),$$

$$\max\left(1-\left[\underset{p=1}{\overset{k+1}{sum}}\left(1-\max\left(1-E_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right)\Bigg]\Bigg\}.$$

$$=\left\{\left[\max\left[1-A_{k+1}^{\frac{\varepsilon_p}{\eta}},0\right],\max\left[1-B_{k+1}^{\frac{\varepsilon_p}{\eta}},0\right]\right],\right.$$

$$\left.\max\left[1-C_{k+1}^{\frac{\varepsilon_p}{\eta}},0\right],\left[\max\left[1-D_{k+1}^{\frac{\omega_p}{\eta}},0\right],\max\left[1-E_{k+1}^{\frac{\varepsilon_p}{\eta}},0\right]\right]\right\}.$$

Hence the result holds for all the values of n.

B. Theorem(Idempotency)

If $\overline{M_p}=\overline{M}$ for all the values of p then $TIT2FYWG_\varepsilon\left\langle\overline{M_1},\overline{M_2},...,\overline{M_n}\right\rangle=\overline{M}$. (13)

Proof:

By theorem A,

$$TIT2FYWG_\varepsilon\left\langle\overline{M_1},\overline{M_2},...,\overline{M_n}\right\rangle$$

$$=\left\{\left[\max\left(1-A_n^{\frac{\varepsilon_p}{\eta}},0\right),\max\left(1-B_n^{\frac{\omega_p}{\eta}},0\right)\right],\right.$$

$$\left.\max\left(1-C_n^{\frac{\varepsilon_p}{\eta}},0\right),\left[\max\left(1-D_n^{\frac{\varepsilon_p}{\eta}},0\right),\max\left(1-E_n^{\frac{\varepsilon_p}{\eta}},0\right)\right]\right\}.$$

$$TIT2FYWG_\varepsilon\left\langle\overline{M_1},\overline{M_2},...,\overline{M_n}\right\rangle$$

$$=\left\{\left[\max\left(1-A_4^{\underset{p=1}{\overset{k+1}{sum}}\left(\frac{\varepsilon_p}{\eta}\right)},0\right),\max\left(1-B_4^{\underset{p=1}{\overset{k+1}{sum}}\left(\frac{\omega_p}{\eta}\right)},0\right)\right],\right.$$

$$\left.\max\left(1-C_4^{\underset{p=1}{\overset{k+1}{sum}}\left(\frac{\varepsilon_p}{\eta}\right)},0\right),\left[\max\left(1-D_4^{\underset{p=1}{\overset{k+1}{sum}}\left(\frac{\varepsilon_p}{\eta}\right)},0\right),\max\left(1-E_4^{\underset{p=1}{\overset{k+1}{sum}}\left(\frac{\varepsilon_p}{\eta}\right)},0\right)\right]\right\}$$

$$=\left\{\left[\max\left(1-A_5^{\frac{1}{\eta}},0\right),\max\left(1-B_5^{\frac{1}{\eta}},0\right)\right],\right.$$

$$\left.\max\left(1-C_5^{\frac{1}{\eta}},0\right),\left[\max\left(1-D_5^{\frac{1}{\eta}},0\right),\max\left(1-E_5^{\frac{1}{\eta}},0\right)\right]\right\}.$$

$$=\left\{\left[A_3,B_3,C_3,D_3,E_3\right]\right\}=\left\{\left[A,B\right],C,\left[D,E\right]\right\}=\overline{M}$$

C. Theorem(Boundary)

Let

$$\overline{M}^{+}=\left\{\left[\max_{p=1}^{n}\left(\underline{l_{M_p}}\right),\max_{p=1}^{n}\left(\overline{l_{M_p}}\right)\right],\max_{p=1}^{n}c_{M_p},\left[\max_{p=1}^{n}\left(\underline{r_{M_p}}\right),\max_{p=1}^{n}\left(\overline{r_{M_p}}\right)\right]\right\}$$

$$\overline{M}^{-}=\left\{\left[\min_{p=1}^{n}\left(\underline{l_{M_p}}\right),\min_{p=1}^{n}\left(\overline{l_{M_p}}\right)\right],\min_{p=1}^{n}c_{M_p},\left[\min_{p=1}^{n}\left(\underline{r_{M_p}}\right),\min_{j=p}^{n}\left(\overline{r_{M_p}}\right)\right]\right\}$$

Then $\overline{M}^{-}\le TIT2FYWG_\varepsilon\left\langle\overline{M_1},\overline{M_2},...,\overline{M_n}\right\rangle\le\overline{M}^{+}$ (14)

Proof:

Since,

$$\min_{p=1}^{n}\left(\underline{l_{M_p}}\right)\le\underline{l_{M_p}}\le\max_{p=1}^{n}\left(\underline{l_{M_p}}\right),\ \min_{p=1}^{n}\left(\overline{l_{M_p}}\right)\le\overline{l_{M_p}}\le\max_{p=1}^{n}\left(\overline{l_{M_p}}\right),$$

$$\min_{p=1}^{n}\left(c_{M_p}\right)\le c_{M_p}\le\max_{p=1}^{n}\left(c_{M_p}\right),$$

$$\min_{p=1}^{n}\left(\underline{r_{M_p}}\right)\le\underline{r_{M_p}}\le\max_{p=1}^{n}\left(\underline{r_{M_p}}\right),\ \min_{p=1}^{n}\left(\overline{r_{M_p}}\right)\le\overline{r_{M_p}}\le\max_{j=1}^{n}\left(\overline{r_{M_p}}\right).$$

we have,

$$1-\max_{p=1}^{n}\left(\underline{l_{M_p}}\right)\le\underline{l_{M_p}}\le1-\min_{p=1}^{n}\left(\underline{l_{M_p}}\right)$$

$$\Rightarrow\min\left(\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(\underline{l_{M_p}}\right)\right)^{\eta}\right]^{\frac{1}{\eta}},1\right)\le\min\left(\left[\underset{p=1}{\overset{n}{sum}}\left(1-\left(\underline{l_{M_p}}\right)\right)^{\eta}\right]^{\frac{1}{\eta}},1\right)$$

$$\le\min\left(\left[\underset{p=1}{\overset{n}{sum}}\left(1-\min\left(\underline{l_{M_p}}\right)\right)^{\eta}\right]^{\frac{1}{\eta}},1\right).$$

$$\Rightarrow \min\left(\left[\left(1-\max\left(\underline{l_{Mp}}\right)\right)^{\eta}\right]^{\underset{p=1}{\overset{n}{prod}}\left(\frac{1}{\eta}\right)},1\right) \le \min\left(\left[\left(1-\underline{l_{M_p}}\right)^{\gamma}\right]^{\underset{p=1}{\overset{n}{prod}}\left(\frac{1}{\eta}\right)},1\right)$$

$$\le \min\left(\left[\left(1-\min\left(\underline{l_{M_p}}\right)\right)^{\eta}\right]^{\underset{p=1}{\overset{n}{prod}}\left(\frac{1}{\eta}\right)},1\right)$$

$$\Rightarrow \min\left(\left[\left(1-\max\left(\underline{l_{M_p}}\right)\right)\right],1\right) \le \min\left(\left[\left(1-\underline{l_{M_p}}\right)\right],1\right)$$

$$\le \min\left(\left[\left(1-\min\left(\underline{l_{M_p}}\right)\right)\right],1\right)$$

$$\Rightarrow \min\left(\min\left(\underline{l_{M_p}}\right),1\right) \le \min\left(1-\underset{p=1}{\overset{n}{sum}}\left(\underline{l_{M_p}}\right),1\right) \le \min\left(\max\left(\underline{l_{M_p}}\right),1\right)$$

$$\Rightarrow \min\left(\underline{l_{M_p}}\right) \le \min\left(1-\underset{p=1}{\overset{n}{sum}}\left(\underline{l_{M_p}}\right),1\right) \le \max\left(\underline{l_{M_p}}\right)$$

Similarly we have,

$$\min\left(\overline{l_{M_p}}\right) \le \min\left(1-\underset{p=1}{\overset{n}{sum}}\left(\overline{l_{M_p}}\right),1\right) \le \max\left(\overline{l_{M_p}}\right),$$

$$\min\left(c_{M_p}\right) \le \min\left(1-\underset{p=1}{\overset{n}{sum}}\left(c_{M_p}\right),1\right) \le \max\left(c_{M_p}\right)$$

$$\min\left(\underline{r_{M_p}}\right) \le \min\left(1-\underset{p=1}{\overset{n}{sum}}\left(\underline{r_{M_p}}\right),1\right) \le \max\left(\underline{r_{M_p}}\right),$$

$$\min\left(\overline{r_{M_p}}\right) \le \min\left(1-\underset{p=1}{\overset{n}{sum}}\left(\overline{l_{M_p}}\right),1\right) \le \max\left(\overline{r_{M_p}}\right)$$

By using the ranking value formula for TIT2FN and using the arithmetic average ranking value,

$$R\left(\overline{M}\right)=\left(\frac{\underline{l_{M_p}}+\overline{r_{M_p}}}{2}+1\right)\times\frac{\underline{l_{M_p}}+\overline{l_{M_p}}+\underline{r_{M_p}}+\overline{r_{M_p}}+4c_M}{8}$$

$$\le\left(\frac{\max\left(\underline{l_{M_p}}\right)+\max\left(\overline{r_{M_p}}\right)}{2}+1\right)\times\left(\underset{p=1}{\overset{n}{\max}}\left(\underline{l_{M_p}}\right)+\underset{p=1}{\overset{n}{\max}}\left(\overline{l_{M_p}}\right)+\underset{p=1}{\overset{n}{\max}}\left(\underline{r_{M_p}}\right)\right.$$

$$\left.+\underset{p=1}{\overset{n}{\max}}\left(\overline{r_{M_p}}\right)+4\underset{p=1}{\overset{n}{\max}}\left(c_{M_p}\right)\right)\times 8^{-1}=R\left(\overline{M}^{+}\right)$$

Hence the result.

D. Theorem

If $t>0$ for all the values of p then

$$TIT2FYWG_{\varepsilon}\left(\overline{M_1}^{\bullet_Y t},\overline{M_2}^{\bullet_Y t},...,\overline{M_n}^{\bullet_Y t}\right)$$

$$=TIT2FYWG_{\varepsilon}\left(\overline{M_1},\overline{M_2},...,\overline{M_n}\right)^{\bullet_Y t} \quad (15)$$

Proof:

$$\overline{M}^{\bullet_Y t}=\left\{\left[\max\left(1-A_5^{\frac{t}{\eta}},0\right),\max\left(1-B_5^{\frac{t}{\eta}},0\right)\right],\max\left(1-C_5^{\frac{t}{\eta}},0\right),\right.$$

$$\left.\left[\max\left(1-D_5^{\frac{t}{\eta}},0\right),\max\left(1-E_5^{\frac{t}{\eta}},0\right)\right]\right\}.$$

$$=\left\{\left[\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-A_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),\right.\right.$$

$$\left.\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-B_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right)\right],$$

$$\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-C_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),$$

$$\left[\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-D_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),\right.$$

$$\left.\left.\max\left(1-\left[\underset{p=1}{\overset{k}{sum}}\left(1-\max\left(1-E_k^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right)\right]\right\}.$$

$$TIT2FYWG_{\varepsilon}\left\langle\overline{M_1},\overline{M_2},...,\overline{M_n}\right\rangle$$

$$=\overline{M_1}^{\bullet_Y \varepsilon_1}\underset{Y}{\otimes}\overline{M_2}^{\bullet_Y \varepsilon_2}\underset{Y}{\otimes}...\underset{Y}{\otimes}\overline{M_n}^{\bullet_Y \varepsilon_n}$$

$$=\left\{\left[\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left(1-A_2\right)^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),\right.\right.$$

$$\left.\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left(1-B_2\right)^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right)\right]$$

$$\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left(1-C_2\right)^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),$$

$$\left[\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left(1-D_2\right)^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right),\right.$$

$$\left.\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left(1-E_2\right)^{\frac{\varepsilon_p}{\eta}},0\right)\right)\right],0\right)\right]\Bigg\}.$$

$$=\left\{\left[\max\left(1-A_n^{\frac{t\varepsilon_p}{\eta}},0\right),\max\left(1-B_n^{\frac{t\varepsilon_p}{\eta}},0\right)\right],\right.$$

$$\left.\max\left(1-C_n^{\frac{t\varepsilon_p}{\eta}},0\right),\left[\max\left(1-D_n^{\frac{t\varepsilon_p}{\eta}},0\right),\max\left(1-E_n^{\frac{t\varepsilon_p}{\eta}},0\right)\right]\right\}. \tag{16}$$

Also since,

$$TIT2FYWG_{\varepsilon}\left(\overline{M_1},\overline{M_2},...,\overline{M_n}\right)^{\bullet_Y t}$$

$$=\left\{\left[\max\left(1-A_n^{\frac{t\varepsilon_p}{\eta}},0\right),\max\left(1-B_n^{\frac{t\varepsilon_p}{\eta}},0\right)\right],\right.$$

$$\left.\max\left(1-C_n^{\frac{t\varepsilon_p}{\eta}},0\right),\left[\max\left(1-D_n^{\frac{t\varepsilon_p}{\eta}},0\right),\max\left(1-E_n^{\frac{t\varepsilon_p}{\eta}},0\right)\right]\right\}^{\bullet_Y t}.$$

$$=\left\{\left[\max\left(\left(1-A_n^{\frac{t}{\eta}}\right)^{\varepsilon_p},0\right),\max\left(\left(1-B_n^{\frac{t}{\eta}}\right)^{\varepsilon_p},0\right)\right],\max\left(\left(1-C_n^{\frac{t}{\eta}}\right)^{\varepsilon_p},0\right),\right.$$

$$\left.\left[\max\left(\left(1-D_n^{\frac{t}{\eta}}\right)^{\varepsilon_p},0\right),\max\left(\left(1-E_n^{\frac{t}{\eta}}\right)^{\varepsilon_p},0\right)\right]\right\}.$$

$$=\left\{\left[\max\left(\left(1-A_n^{\frac{t\varepsilon_p}{\eta}}\right),0\right),\max\left(\left(1-B_n^{\frac{t\varepsilon_p}{\eta}}\right),0\right)\right],\right.$$

$$\left.\max\left(\left(1-C_n^{\frac{t\varepsilon_p}{\eta}}\right),0\right),\left[\max\left(\left(1-D_n^{\frac{t\varepsilon_p}{\eta}}\right),0\right),\max\left(\left(1-E_n^{\frac{t\varepsilon_p}{\eta}}\right),0\right)\right]\right\} \tag{17}$$

Since (16) = (17), hence the result.

E. Theorem(Stability)

If $t>0$, $\overline{M_{n+1}}=\left(\left[\underline{l}_{M_{n+1}},\overline{l}_{M_{n+1}}\right],c_{M_{n+1}},\left[\underline{r}_{M_{n+1}},\overline{r}_{M_{n+1}}\right]\right)$ then

$$TIT2FYWG_{\varepsilon}\left(\overline{M_1}^{\bullet_Y t}\underset{Y}{\otimes}\overline{M_{n+1}},\overline{M_2}^{\bullet_Y t}\underset{Y}{\otimes}\overline{M_{n+1}},...,\overline{M_n}^{\bullet_Y t}\underset{Y}{\otimes}\overline{M_{n+1}}\right)$$

$$=TIT2FYWG_{\varepsilon}\left(\overline{M_1},\overline{M_2},...,\overline{M_n}\right)^{\bullet_Y t}\underset{Y}{\otimes}\overline{M_{n+1}} \tag{18}$$

Proof:

$$TIT2FYWG_{\varepsilon}\left(\overline{M_1}^{\bullet_Y t}\underset{Y}{\otimes}\overline{M_{n+1}},\overline{M_2}^{\bullet_Y t}\underset{Y}{\otimes}\overline{M_{n+1}},...,\overline{M_n}^{\bullet_Y t}\underset{Y}{\otimes}\overline{M_{n+1}}\right)$$

$$=\left\{\left[\max\left(1-A_n^{\frac{t\varepsilon_p}{\eta}},0\right)\underset{Y}{\otimes}\overline{M_{n+1}},\max\left(1-B_n^{\frac{t\varepsilon_p}{\eta}},0\right)\underset{Y}{\otimes}\overline{M_{n+1}}\right],\right.$$

$$\max\left(1-C_n^{\frac{t\varepsilon_p}{\eta}},0\right)\underset{Y}{\otimes}\overline{M_{n+1}},$$

$$\left.\left[\max\left(1-D_n^{\frac{t\varepsilon_p}{\eta}},0\right)\underset{Y}{\otimes}\overline{M_{n+1}},\max\left(1-E_n^{\frac{t\varepsilon_p}{\eta}},0\right)\underset{Y}{\otimes}\overline{M_{n+1}}\right]\right\}$$

$$=\left\{\left[\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left[\underset{q=\{p,n+1\}}{sum}\left(1-\underline{l}_{M_q}\right)^{\eta}\right]^{\frac{t}{\eta}},0\right)^{\varepsilon_p}\right)\right],0\right),\right.\right.$$

$$\left.\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left[\underset{q=\{p,n+1\}}{sum}\left(1-\overline{l}_{M_q}\right)^{\eta}\right]^{\frac{t}{\eta}},0\right)^{\varepsilon_p}\right)\right],0\right)\right],$$

$$\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left[\underset{q=\{p,n+1\}}{sum}\left(1-c_{M_q}\right)^{\eta}\right]^{\frac{t}{\eta}},0\right)^{\varepsilon_p}\right)\right],0\right),$$

$$\left[\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left[\underset{q=\{p,n+1\}}{sum}\left(1-\underline{r}_{M_q}\right)^{\eta}\right]^{\frac{t}{\eta}},0\right)^{\varepsilon_p}\right)\right],0\right),\right.$$

$$\left.\left.\max\left(1-\left[\underset{p=1}{\overset{n}{sum}}\left(1-\max\left(1-\left[\underset{q=\{p,n+1\}}{sum}\left(1-\overline{r}_{M_q}\right)^{\eta}\right]^{\frac{t}{\eta}},0\right)^{\varepsilon_p}\right)\right],0\right)\right]\right\}.$$

$$=\left\{\left[\max\left(\left(1-A_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-\underline{l}_{M+1}\right)^{\eta}\right]^{\frac{1}{\eta}}\right)^{\underset{p=1}{\overset{n}{sum}}\varepsilon_p},0\right),\right.\right.$$

$$\max\left(\left(1-B_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-\overline{l_{M+1}}\right)^{\eta}\right]^{\frac{1}{\eta}}\right)^{\overset{n}{\underset{p=1}{sum}}\varepsilon_p},0\right)\Bigg]\Bigg],$$

$$\max\left(\left(1-C_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-c_{M+1}\right)^{\eta}\right]^{\frac{1}{\eta}}\right)^{\overset{n}{\underset{p=1}{sum}}\varepsilon_p},0\right),$$

$$\Bigg[\max\left(\left(1-D_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-\underline{r_{M+1}}\right)^{\eta}\right]^{\frac{1}{\eta}}\right)^{\overset{n}{\underset{p=1}{sum}}\varepsilon_p},0\right),$$

$$\max\left(\left(1-E_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-\overline{r_{M+1}}\right)^{\eta}\right]^{\frac{1}{\eta}}\right)^{\overset{n}{\underset{p=1}{sum}}\varepsilon_p},0\right)\Bigg]\Bigg]\Bigg\}. \qquad (19)$$

Based on the theorem A and the operational law,

$$\text{TIT2FYWG}_{\varepsilon}\left\langle\overline{M_1},\overline{M_2},\ldots,\overline{M_n}\right\rangle^{\bullet t}_{Y}\underset{Y}{\otimes}\overline{M_{n+1}}$$

$$\left\{\left[\max\left(1-A_n^{\frac{t\varepsilon_p}{\eta}},0\right),\max\left(1-B_n^{\frac{t\varepsilon_p}{\eta}},0\right)\right],\right.$$

$$\left.\max\left(1-C_n^{\frac{t\varepsilon_p}{\eta}},0\right),\left[\max\left(1-D_n^{\frac{t\varepsilon_p}{\eta}},0\right),\ \max\left(1-E_n^{\frac{t\varepsilon_p}{\eta}},0\right)\right]\right\}.$$

$$\underset{Y}{\otimes}\left\langle\left[\underline{l_{M_{n+1}}},\overline{l_{M_{n+1}}}\right],c_{M_{n+1}},\left[\underline{r_{M_{n+1}}},\overline{r_{M_{n+1}}}\right]\right\rangle$$

$$=\left\{\left[\max\left(\left(1-A_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-\underline{l_{M+1}}\right)^{\eta}\right]^{\frac{1}{\eta}}\right),0\right),\right.\right.$$

$$\left.\max\left(\left(1-B_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-\overline{l_{M+1}}\right)^{\eta}\right]^{\frac{1}{\eta}}\right),0\right)\right],$$

$$\max\left(\left(1-C_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-c_{M+1}\right)^{\eta}\right]^{\frac{1}{\eta}}\right),0\right),$$

$$\left[\max\left(\left(1-D_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-\underline{r_{M+1}}\right)^{\eta}\right]^{\frac{1}{\eta}}\right),0\right),\right.$$

$$\left.\left.\max\left(\left(1-E_n^{\frac{\varepsilon_p}{\eta}}\right)+\left(\left[\left(1-\overline{r_{M+1}}\right)^{\eta}\right]^{\frac{1}{\eta}}\right),0\right)\right]\right\} \qquad (20)$$

Here, (19) = (20).

Hence the result.

F. *Theorem(Image Contrast)*

For given arguments $\overline{M}_p, p=1,2,\ldots,n$ and the parameter $\eta\in(1,+\infty)$ then TIT2FYWG operator is monotonically non-decreasing (MND) with respect to the parameter.

Proof:

To prove the operator is MND with respect to the parameter, we have to prove the same for every reference point function is MND w.r.t the parameter.

Since $0\le\underline{l_M}\le\overline{l_M}\le c_M\le\underline{r_M}\le\overline{r_M}\le 1$, $\max\left(1-A_5^{\frac{k}{\eta}},0\right)>0.$

And it is true for all the reference points.

Hence the result.

Note: The above theorems also can be proved by using TIT2FYWA operator.

VI. Theory of Image Processing and Role of Yager Norms

The advantage of considering Yager triangular norms is having maximum and minimum operators which will be much useful in Image Processing while filtering.

A. *Associativity of Yager T Norms [8]*

Each fuzzy norm should be satisfied the associativity property to compute the norm for more than two values using continual manner as follows.

Consider the associativity property for Yager T Norm (YTN)

$$YTN\left[TN\left(\overline{M}_1,\overline{M}_2\right),\overline{M}_3\right]$$

$$=\max\left\{1-\left[\left[1-T\left(\overline{M}_1,\overline{M}_2\right)\right]^{\eta},\left(1-\overline{M}_3\right)^{\eta}\right]^{1/\eta},0\right\}$$

$$=\max\left\{1-\left\{\left[1-\left[\max\left\{1-\left[\left[\left(1-\overline{M}_1\right)^{\eta}+\left(1-\overline{M}_2\right)^{\eta}\right]^{1/\eta}\right]^{\eta},0\right\}\right]\right]+\left(1-\overline{M}_3\right)^{\eta}\right\}^{1/\eta},0\right\}$$

$$=\max\left\{\left[\left\{1-1+\left\{\max\left[1-\left(\left(1-\overline{M}_1\right)^{\eta}+\left(1-\overline{M}_2\right)^{\eta}\right)^{1/\eta}\right]^{\eta},0\right\}+\left(1-\overline{M}_3\right)^{\eta}\right\}^{1/\eta}\right],0\right\}$$

$$= \max\left\{\left[\left\{\max\left\{\left[1-\left(\left(1-\overline{M}_1\right)^\eta+\left(1-\overline{M}_2\right)^\eta\right)^{1/\eta}\right]^\eta,0\right\}+\left(1-\overline{M}_3\right)^\eta\right\}^{1/\eta}\right],0\right\}$$

$$= \max\left\{\left(\left[1-\left(\left(1-\overline{M}_1\right)^\eta+\left(1-\overline{M}_2\right)^\eta\right)^{1/\eta}\right]^\eta+\left(1-\overline{M}_3\right)^\eta\right)^{1/\eta},0\right\}$$

$$= \max\left\{\left(\left[1-\left(1-\overline{M}_1\right)^\eta+\left(\left(1-\overline{M}_2\right)^\eta+\left(1-\overline{M}_3\right)^\gamma\right)^{1/\eta}\right]^\eta\right)^{1/\eta},0\right\}$$

$$= YTN\left[\overline{M}_1, TN\left(\overline{M}_2,\overline{M}_3\right)\right]. \quad (21)$$

Similarly for Yager T conorm (YTCN).

Here we can consider $\overline{M}_1$ and $\overline{M}_2$ as the Interval Type-2 Triangular Fuzzy Number. If $\max\left(\overline{M}_1,\overline{M}_2\right)=1$ then (YTN) will become $\min\left(\overline{M}_1,\overline{M}_2\right)$. If $\min\left(\overline{M}_1,\overline{M}_2\right)=0$ then (YTCN) will become $\max\left(\overline{M}_1,\overline{M}_2\right)$. In the same manner we can have for YTCN. Here the definition of YTN accords the inference around the effectiveness of the norm and γ, its complimentary parameter. Generally, it allows tuning between the norms.

If η approaches 0, then YTN will be $\min\left(\overline{M}_1,\overline{M}_2\right)$ only when $\max\left(\overline{M}_1,\overline{M}_2\right)=1$ i.e., their drastic product. If η approaches 1, then YTN becomes $\max\left[(\overline{M}_1+\overline{M}_2-1),0\right]$.

B. Role of Associativity of Yager T Norm in Image Processing

Using Fuzzy Set approach, we can generalize a binary morphology into MFM. Morphological operations are the basic tools to modify the image I_1 over the structural aspect of I_2. To study the structure of I_1, size and shape of the I_2 are chosen accordingly.

C. Morphological Operations [24]

(i). Erosion (Maximum)

(ii). Dilation (Minimum)

(iii). Opening and Closing (Idempotency)

D. Role of T-Norms in Image Processing [17]

For constructing FM, we use Conjunctions and Implications. Among these two, we used conjunctions (t-norms) here and from the below, the representation of mathematical properties in image processing has been explained.

1) Commutativity:

The result of IDS application on two successive points P and Q is the same as applying on them in inverse order, since the value of flapped points is the sum of values of all data diluted on that point and therefore the operator is commutative.

2) Monotonicity:

If the brightness of P is less than or equal to Q then all the data points in brightness of P is less than or equal to brightness with respect to the corresponding data points of brightness of Q.

Therefore for any point n, the brightness appeared from P is $n+aP$, where a is proportional to inverse of distance. Similarly, the brightness appears from Q is $n+aQ$ Since $a>0$, the brightness of n appeared from P is less than or equal to that of from Q.

3) Associativity:

Assume that P, Q and R are the sources of light going to affect to the point n by IDS.

For every source, IDS increases the brightness with respect to the distance regardless of other sources.

On the point n, the order of applying IDS does not affect the distance.

Sum of effects of P, Q and R is the value of n. Therefore, the operator is associative.

4) Idempotency:

This property and its generalization is used for the morphological operation opening and closing.

5) Neutrality of 0:

Consider a pyramid of height 0, sum of this with others does not influence them. Therefore, 0 is the neutral element.

E. Morphological Gradient (MG) [24]

It is useful to detect an edge and act as a first approximation to a morphological segmentation. MG is the discrepancy between

a) dilation and erosion

b) dilation and the original image

c) original image and its erosion

F. T-Norm and Image Compression(IC) [17]

IC is based on Fuzzy Relational (FR) Equations and it is a grayscale image of size $C\times D$ as a FR $\Re\in F(A,B)$ where, $A=\{a_1,a_2,...,a_C\}$, $B=\{b_1,b_2,...,b_D\}$ enclosed the depth range of each pixel into $[0,1]$.

$CS=\left\{CS^{(R)},CS^{(G)},CS^{(B)}\right\}\subset F(A,B)$ represent the color image on RGB Color Space (CS). Here $CS^{(R)}$, $CS^{(G)}$ and $CS^{(B)}$ are the Red, Green and Blue color spaces.

For clarity, gray scale image (GSI) will be considered. The GSI $\Re\in F(A,B)$ is restrict into $\wp\in F(I\times J)$, through a max t-norm FR equations with composition $\wp = \max_{b \in B}\left\{V_j(b)TN \max_{a \in A}\left[U_i(a)TN\square\,(a,b)\right]\right\}$, where TN is a continuous t-norm, $U_i \in U = \{U_1, U_2, ..., U_I\} \subset \mathrm{F}(A)$ and $V_i \in V = \{V_1, V_2, ..., V_J\} \subset \mathrm{F}(B)$ are the coders.

The shape of the FSs of coders are the triangular line, preferable for IC. We can adjust the compression rate of IC by the sum of FSs consist in U and V ad is defined by $\zeta = \frac{IJ}{AB}$. Here IJ and AB the compressed and original image coefficients respectively. By adjusting the parameter η, YTN will all the continuous T-norm where as Zadeh's and major t norms cannot do the same. Though Frank t-norm can do the same, due to the computational complexity, we prefer Yager's t norm for image processing.

G. Role of T2FS [24]

Here the components of an image processing and the role of T2FS is correlated.

1) Image Contrast Enrichment:

The most common image enrichment method is histogram equalization. Since an image has an imprecise pixel grey values, it may not produce acceptable results in IP. To handle the ambiguity of the gray values, Fuzzy methods have been suggested by many researchers.

By adjusting the membership values, the contrast of the image is increased by contrast intensification operator and it transforms the higher MVs to much higher and lower MVs to much lower in a nonlinear aspect. Since this aspect considers whole image, global histogram fails to produce satisfactory results.

Though the fuzzy methods deals ambiguity well and produced proper enrichment, it fails in some case and hence T2FS has been considered for this purpose since it deals more uncertainties.

2) Image Segmentation:

Region boundaries of an image may not have a fine growth, therefore fuzzy decision is used to check whether the pixel exists to a region and T2FS may be applied to get better threshold images.

3) Clustering:

The images have different regions with different pitch, clustering collects the similar pixels in a group with membership value 1 and collects different pixels in different group with membership value 0. But in fuzzy clustering the pixel associate to different number of groups and hence the MVs are not 0.

4) Edge Detection:

Since most of the images have poor brightness, the proper decision cannot be taken in checking the existence of an edge in an image. Edges may be enriched before carrying out the edge detection. In taking off the edge due to ambiguity, fuzzy method may be useful and may not find better edges. At this junction T2FS is useful as it handles more uncertainties.

5) Morphology:

Which is a non-linear image processing technique and is used to shape the image features. Here also T2FS plays an important role to get better results.

VII. Application of Image Proessing

Fig. 2. Shows that the Architecture has been proposed for edge detection on DICOM image using triangular norms

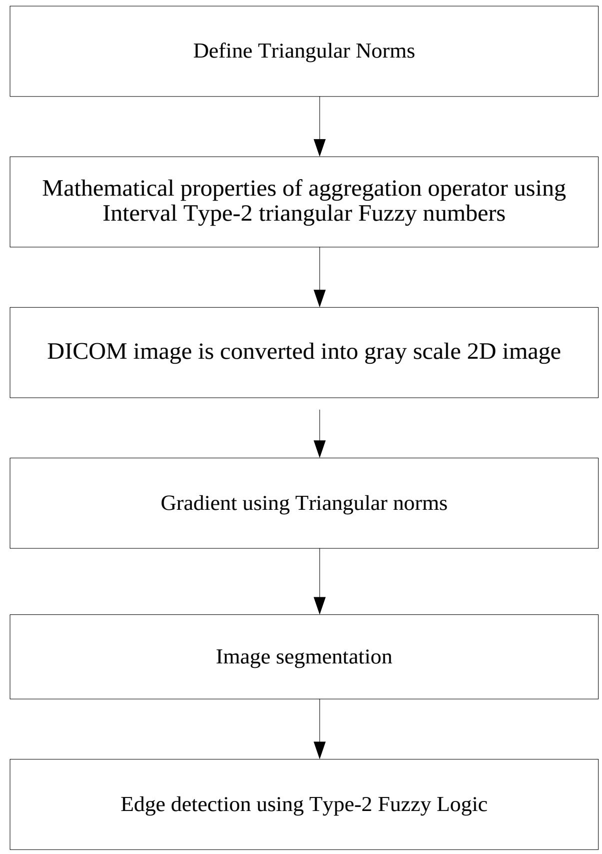


Fig. 2. Architecture for Edge Detection.

Using MATLAB 2015a, triangular norms has been applied in medical image processing from a patient DICOM image. In this case 3D image is converted to 2D image.

In Fig. 3, the image is collected from our experimental data set from a patient DICOM image in the Fig. 7. From this Fig. 7. the clear image Fig. 8. Has been considered for the experiment.

Size of the image = 512 x 517.

Mean of the image = 28.83.

Standard deviation = 60.79

Mean absolute deviation = 40.03.

To identify the gradient of the image by dilation-erosion, triangular norms are used.

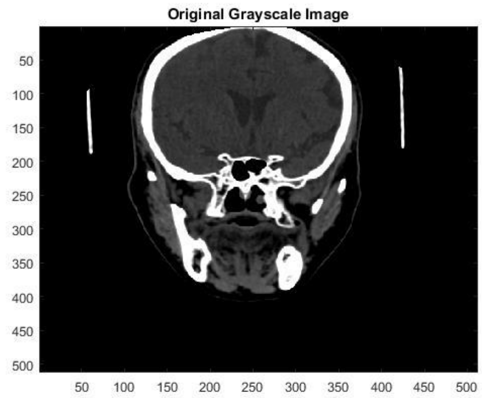


Fig. 3. Original gray scale DICOM image.

Structuring elements are used in gradient value.

Image Erosion is

0.9961	1.0000	0
1.0000	0.9961	0
1.0000	1.0000	1.0000

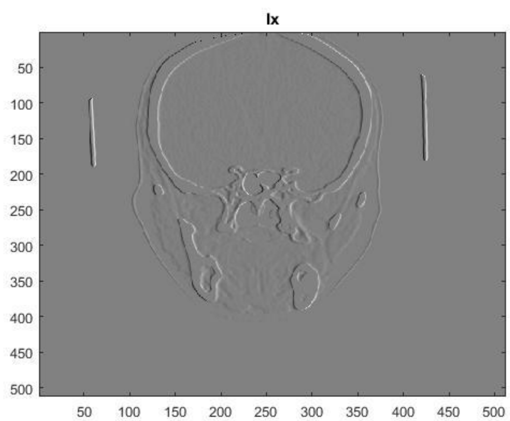


Fig. 4. Gradient through x axis.

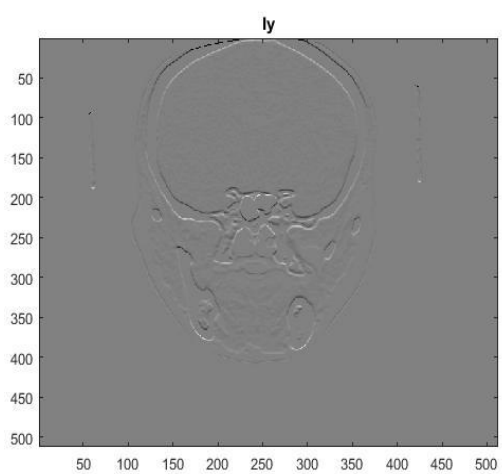


Fig. 5. Gradient through y axis.

Gradient through y axis.The below figures are the output of the image processing application in edge detection through triangular norms by MATLAB 2015 a.

Fig. 4. is the gradient through x axis and Fig. 5. is the gradient through the y axis.

The figures reveals that the image gradient to identify the region uniformly.

Fig. 6. is the output of the edge detection through T2 fuzzy by our experimental output using MATLAB 2015a

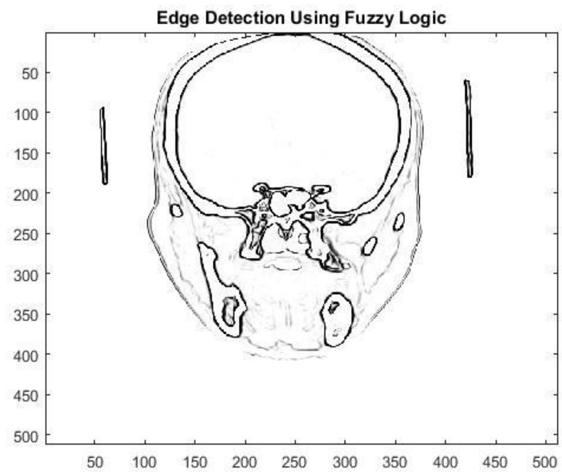


Fig. 6. Edge detection image.

Fig. 6. shows that the edges of the object through FIS and equating the pixel on both direction. If the edge is block then pixel is not 0.

Edge detection plays a vital role in image identification. It is observed that, fuzzy logic edge detector helps in reducing the memory for saving medical images.

VIII. Conclusion

In this paper, operational laws of addition, multiplication, power and multiplication by an arbitrary number using Yager triangular norms for TIT2FN are derived. Also some properties of aggregation operation using Fuzzy Yager Weighted Geometric operators have been proved. Since Yager aggregation operator contains minimum and maximum operator, it will be act as a morphological filters in medical image processing. Detailed representation of the mathematical properties in image processing is presented. Also, the gradient of the DICOM image of MRI scan of a patient using Triangular norms is found and done edge detection using MATLAB in T2 fuzzy logic. The future work is planned to apply T2 Fuzzy logic in edge extraction on medical image in 3D models.

Data Availability statement

The DICOM data used to support the findings of this study are available from the corresponding author upon request.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary Materials

The data set in Fig. 7. is the montage of the images in a single file and is from a patient MRI. This MRI which is in the 3D form is converted to 2D form (DICOM) using MATLAB2015a. The 3D format consists of 25 DICOM file formats; the montage of the images is obtained as a single frame. Out of these 25 DICOM images a clear full image is chosen as in Fig. 8. Using Dilation corrosion method, the gradient is identified. The edge detection is performed through triangular norms using MATLAB 2015a.

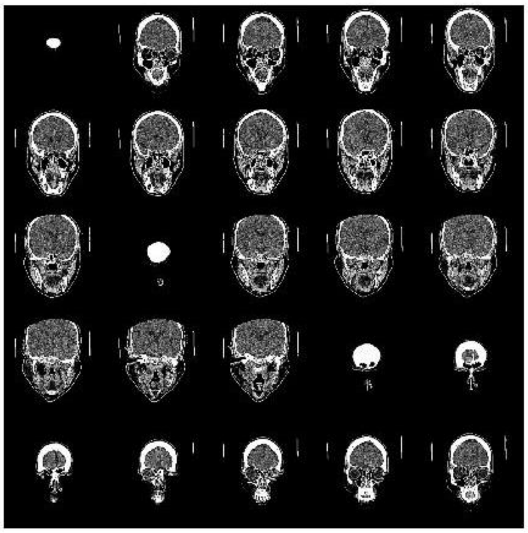


Fig. 7. Montage of the images.

Fig. 8. Clear image from montage.

REFERENCES

[1] T. Calvo, G. Mayor and R. Messiar, "Aggregation Operators: New Trends and Application," 2012, Springer Publications.

[2] D. Vivona and M. Divari, "On Information Reception and Some of its Aggregation Operators," 8th International Summer School on Aggregation Operators, Kotowice, Poland, University of Silesia, 2015, pp.233-237.

[3] J. Qin and X. Liu, "Frank Aggregation operators for Triangular Interval Type-2 Fuzzy Set and its Application in Multiple Attribute Group Decision Making," Journal of Applied Mathematics, Vol. 1, 2014, pp.1-24.

[4] F. Castro-Company and P. Tirado, "On Yager and Hamacher t-norms and Fuzzy Metric spaces," International Journal of Intelligent Systems, Vol.29, No.12, 2014, pp.1173-1180.

[5] E. Czogala and H. Zimmermann, "The Aggregation Operators for Decision Making in Probabilistic Fuzzy Environment," Fuzzy Sets and Systems, Vol.9, No.1, 1984, pp.183-196.

[6] T. Calvo and R. Mesiar, "Weighted Means Based on Triangular Conorms," International Journal of Uncertain Fuzziness Knowledge Based Systems, Vol.29, No.12, 2001, pp.1173-1180.

[7] M. M. Gupta and J. Qi, "Theory of t-norms and Fuzzy Inference Methods," Fuzzy Sets and Systems, Vol. 40, No. 3, 1991, pp.431-450.

[8] D. Dubois and H. Prade, "On the Use of Aggregation operations in Information Fusion Processes," Fuzzy Sets and Systems, Vol. 142, No.3, 2004, pp.143-161.

[9] M. Detynieki, "Mathematical Aggregation Operators and their Application to Video Querying," Ph.D. Thesis in Artificial Intelligence Specialty of Paris, 2000, pp.1-104.

[10] R. Mesiar, A. Kolesarova, T. Calvo and M. Komornikova, "A review of aggregation functions," In: Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, H.Bustince et al. (Eds.), Springer, Berlin, 2008, pp.121-144.

[11] D. Dubois and H. Prade, "A review of fuzzy set aggregation Connectives, Information Sciences," Vol. 36, No.1-2, 1985, pp. 85-121.

[12] R. R. Yager, "Aggregation operators and fuzzy systems modeling," Fuzzy Sets and Systems, Vol. 67, No.2, 1994, pp.129-145.

[13] L. A. Zadeh, "The concept of a linguistic variable and its application to Approximate Reasoning-I," Information Sciences, Vol.8, No.3, 1975, pp.199-249.

[14] C. H. Chiu and W. J. Wang, "A simple computation of MIN and MAX operations for fuzzy numbers," Fuzzy Sets and Systems, Vol.126, No. 2, 2002, pp.273-276.

[15] Z. Gera and J. Dombi, "Exact calculations of extended logical operations on fuzzy truth values," Fuzzy Sets and Systems, Vol.159, No.11, 2008, pp.1309-1326.

[16] S. Zhou, F. Chiclana, R. John and J. Garibaldi, "Type-1 OWA operators for aggregating uncertain information with uncertain weights induced by type-2 linguistic Quantities," Fuzzy Sets and Systems, Vol.159, No.24, 2008, pp.3281-3296.

[17] K Franke, M. Koppen and B.Nickolay, "Fuzzy Image Processing by using Dubois and Prade Fuzzy Norms," Pattern Recognition, 2000, pp.1-4.

[18] A. A. Kiaei, S. B. Shouraki, S.H. Khasteh, M. Khademi and A. R. G. Samani, "New S-norm and T-norm Operators for Active Learning Method," Fuzzy Optimization and Decision Making, 2010, pp.1-11.

[19] J. Dombi, "A general class of fuzzy operators, the DeMorgan class of fuzzy operators and fuzziness measures induced by fuzzy operators," Fuzzy Sets and Systems, Vol. 8, No.2, 1982, pp.149-163.

[20] Y. Yao and J. Wang, "Interval based Uncertain Reasoning using Fuzzy and Rough Sets," Advances in Machine Intelligence & Soft-Computing, 1996, pp.1-20.

[21] K. Franke and M. Koppen, "A computer-based system to support forensic studies on handwritten documents," International Journal on Document Analysis and Recognition, 2000, pp.1-13.

[22] K. Hirota, H. Nobuhara, K. Kawamoto and S. I. Yoshida, "On a Lossy Image Compression/ Reconstruction Method Based on Fuzzy Relational Equations," Iranian Journal of Fuzzy Systems, Vol.1, No.1, 2004, pp.33-42.

[23] M. G. Hidalgo, A. M. Torres, D. R. Aguilera, J. T. Sastre, "Image Analysis Applications of Morphological Operators on Uninorms," Soft computing techniques for uncertainty management in image processing, 2009, pp. 630-635. ISBN: 978-989- 95079-6-8.

[24] T. Chaira, "Medical Image Processing: Advanced Fuzzy Set Theoretic Techniques," CRC Publications, 2015, ISBN: 9781498700450-CAT#K24522.

[25] O. Castillo, M. A. Sanchez, C. I. Gonzalez and G. E. Martinez, "Review of Recent Type-2 Fuzzy Image Processing Applications," Information, Vol. 8, No. 97, 2017, pp. 1-18.

[26] I. Sheikh and K. A. Khan, "Simulation of Image Edge Detection using Fuzzy Logic in MATLAB," International Journal of Computer & Mathematical Sciences, Vol. 6, No. 5, 2017, pp. 19-22.

[27] Kenjharayoobchandio and Yasarayaz, "Fuzzy Logic Based Digital Image Edge Detection," International Journal of Electrical, Electronics and Data Communication, Vol. 6, No. 2, 2018, pp.18-22.

[28] A. K. Dutta, "Intuitionistic Fuzzy Logic Implementation in Image Fusion Technique," Asian Journal of Research in Computer Science, Vol. 1. No. 1, 2018, pp. 1-7.

[29] D. G. Privezentsev, A. L. Zhiznyakov, A. V. Astafiev and E. V. Pugin, "Using fuzzy fractal features of digital images for the material surface analysis," Journal of Physics, Vol. 944, 2018, pp. 1-5.

[30] R. Maini and H. Aggarwal, "Study and Comparison of Various Image Edge Detection Techniques," International Journal of Image Processing, Vol.3, No.1, 2009, pp.1-12.

Content Analysis of Privacy Management Features in Geosocial Networking Application

Syarulnaziah Anawar, Yeoh Wai Hong, Erman Hamid, Zakiah Ayop
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

***Abstract*—Geosocial networking application allows user to share information and communicate with other people within a virtual neighborhood or community. Although most geosocial networking application include privacy management features, one the challenge is to improve privacy management features design. To overcome this challenge, the adaptation of privacy-related theories offers a concrete way to comprehend and analyze how the privacy management features are used as tangible research results that facilitate user and system developer in understanding privacy management. This paper attempt to propose a standardized privacy management features in geosocial networking application from market perspectives that could be utilized by researchers and application developers to demonstrate or measure privacy management features. The objective of this paper is two-fold: First, to map the theoretical constructs guided by Communication Privacy Management (CPM) theory into privacy management features in geosocial networking application. Second, to evaluate the reliability of the proposed features using content analysis. Content analysis is performed on 1326 geosocial networking apps in the market (Google Play store and App Store) to determine the reliability of the proposed privacy management features through inter-coder reliability analysis. The primary findings of the content analysis show that many of the privacy management features with low reliability are from Boundary Turbulence construct. Furthermore, only 6 out of 13 proposed features are deemed reliable, namely, specific grouping, visibility setting, privacy policy, violation, imprecision and inaccuracy. The proposed privacy management features may aid researchers and system developers to focus on the best privacy management features for improving geosocial networking application design.**

Keywords*—Privacy management; communication privacy management theory; social network; geosocial network; content analysis

I. Introduction

Geosocial networking application has gain increased popularity over the years that allow information sharing and communication between peoples. The difference between geosocial network and traditional social network is its location-aware capability. Geosocial networking combines real-time location-reporting capabilities with traditional social network functionality [1]. Geosocial networking applications provide its service by utilizing a number of techniques, such as geocoding, geotagging, and geolocation [2]. These capabilities enhance the functionality of social network. For example, user can update their current location status by using the “Check-in” feature.

However, disclosing user and location information can lead to several privacy risks. According to [1], some of the privacy concerns in geosocial networking are location, absence, and co-location privacy. Location privacy concerns on how detailed the location information that a user wishes to disclose, absence privacy concerns on the absence of privacy violation when the location information reveals user’s absence on specific location, while co-location privacy concerns on the “tagging” feature in geosocial network. User’s effort is very important in managing privacy sensitivity values in order to achieve higher privacy level and result [3]. Some of the user privacy aspects must be compromised because location information is essential in geosocial networking, and the application should provide automatic support in determining real-time location of the users.

The aim of this study is to propose privacy management features for geosocial networking application. Recognizing substantial challenge in incorporating theoretical privacy constructs into geosocial networking application design, this study attempts to propose a standardized privacy management features that could be utilized by researchers and application developers to demonstrate or measure privacy management features. By doing so, researcher will have better understanding on how most market players manage user privacy in geosocial network and then adapt the reliable privacy features into geosocial networking applications.

This paper presents the analysis of privacy management features and describes the content analysis performed on geosocial networking apps in the market. This paper is set to address research question: “What are the suitable privacy management features to represent the theory-privacy constructs in a geosocial networking application?”. The rest of the paper is organized as follows: The first section explains how the privacy constructs are conceptualized as IT artifacts by mapping the constructs derived from Communication privacy management (CPM) theory into privacy management features. Next, this paper presents the data collection method in order to draw suitable sample of geosocial networking application to be evaluated in the content analysis. Finally, the proposed privacy management features are evaluated using inter-coder reliability analysis, highlighting reliable features that contributed positively to future privacy management design.

II. Related Work

Privacy management has been studied widely in the research area of social networking. Although most social networking applications include privacy management features, the advantage of privacy management is not well understood by social networking users due to poor design of privacy management in the social networking applications [4]. Therefore, one of the most significant research in this area is to improve privacy management features design. Including users into application design is challenging because they should be well-informed about the geosocial networking applications in order to be part of the decision making process to achieve the anticipated application's goal [5].

To overcome this challenge, the adaptation of privacy-related theories offers a concrete way to comprehend and analyze how the privacy management features are used as tangible research results that facilitate user and system developer in understanding privacy management. Some studies have investigated the role of theory driven privacy constructs in privacy management. One of the earlier effort can be seen in [6], where they offer analysis of privacy management practice and the usability aspect of the privacy management. Their work adopted Adaptive Structuration Theory (AST) to establish reliable predictors of online privacy management in social setting. These predictors are particularly useful in measuring user appropriation; which refers to the process of technology adoption and adaptation by user.

Cho [7] applied Communication privacy management (CPM) theory to examine the influence of cultural differences among Facebook users from the US, Singapore, and South Korea in privacy management strategies. On the other hand, Wilkinson [8] has developed a User-Tailored Privacy by Design framework drawn from Privacy by Design philosophy, that combines multiple privacy management features into a single intelligent user interface. In the context of geosocial networking, [9] provide a comparative privacy analysis of several existing geosocial network, and provide discussions on privacy and security recommendation to enhance the protection of privacy in geosocial networking. However, they do not offer substantial recommendation on privacy management features design.

Some studies have investigated the role of theoretical foundation in improving privacy management design from user perspectives. [10] has established content analysis that emphasize on potential damage to users through information security and privacy infringements in mobile health apps. However, there is lack of studies that analyze privacy management features particularly in geosocial networking from market perspectives. Designing the privacy management features from market perspectives may provide knowledge on how the privacy features is taking part in actual practice in geosocial networking application. Consequently, the privacy management features may then be evaluated as part of the tangible results of the theoretical constructs understudy.

III. Conceptualizing Privacy Constructs as IT Artifacts

The challenge in investigating theory-driven privacy management in geosocial networking application is how to conceptualize the privacy management theory into an IT artifact. This study perceives the importance of an IT artifact to provide a tangible research results in order to reach practitioners and stakeholders in the social networking. In the context of this study, the IT artifact is presented as a set of features in social networking application that can be used as building blocks to enforce privacy management among the users. The adaptation of the theoretical foundation offers a concrete way to comprehend and analyze how privacy are used as tangible research results that facilitate users' behavior. The privacy management features are derived by mapping the theoretical conception in the theoretical constructs through various tools in an existing social networking application.

A. *Privacy Management Constructs*

This study integrates constructs from prominent theory in privacy management, Communication privacy management (CPM) theory as the underlying theoretical foundation for the content analysis conducted in this study. CPM [11][12] is primarily focused on how the decisions of revealing or disclosing private information are made by people. In CPM, privacy is considered to be a process of opening and closing a boundary to others [11]. Margulis [13] regarded CPM as "the most valuable privacy theory for understanding interpersonal computer-mediated communication". As geosocial networking is an integration of location-aware services with online social networks. The services pose substantial privacy threats: user location information may be tracked and leaked to third parties [14]. Therefore, obfuscation is also included as one of the privacy management constructs that could be studied to determine its influence on privacy against neighbor in this study.

The following discussion will examine how these constructs could be applied to social networking, especially to privacy management behavior as guided by CPM theory [11]:

1) Privacy rule characteristics: Refers to how people obtain rules of privacy and understand the properties of those rules to decide how information will be shared. When rules are applied, people may create imaginary or metaphorical boundaries, around their information .

2) Boundary coordination: Refers to when there are multiple parties creating boundary, the information is considered co-owned. All the co-owners should have same understanding on how the privacy should be managed. Coordination of boundaries of privacy and disclosure by the coowner of information based on boundary permeability, linkage, and ownership.

3) Boundary Turbulance: Refers to turbulence among co-owners when rules are not mutually understood and when the management of private information comes into conflict with the expectations of each owner.

4) Obfuscation: Refers to how private information is presented in a falsified manner through data masking. To achieve this, data are deliberately scrambled to inhibit unauthorized access to sensitive materials.

B. *Mapping Privacy Constructs into Privacy Management Features*

This study first mapped the previously defined privacy constructs into privacy management features. All privacy management features represent each construct and correspond to mobile geosocial networking setting in order to maintain coherency of this study. Table 1 illustrates how four constructs in the study: Privacy Rule Characteristics, Boundary Coordination, Boundary Turbulence, and Obfuscation could be applied to geosocial networking setting. A list of privacy management features that represent the constructs in geo-social networking application is identified in Table 1. The privacy management features are mapped by reviewing relevant literature on application of CPM theory in mobile application development and conducting discussion with two mobile apps developers.

TABLE I. MAPPING OF PRIVACY CONSTRUCT INTO PRIVACY MANAGEMENT FEATURES

Construct	Feature	Description
Privacy Rule Characteristics	Service access	Feature that allows user to choose whether to allow services, e.g. location service, “find my friend” service to access device data.
	Specific grouping	Feature that allows user to create/join user group for specific communication, i.e. interest based, relationship based.
Boundary coordination	Visibility setting	Feature that allows user to choose who can see their private information.
	Activity Log	Feature that allows user to review who can see their activity.
	Tagging	Feature that allows user to manage the information they are tagged in.
Boundary turbulence	Classification	Feature that allows user to classify another user’s role from their perspective.
	Privacy policy	Feature that declares the terms and regulations on information privacy that user needs to obey.
	Education	Feature that educates user the proper ways to control their privacy.
	Notification	Feature that notifies users to control their privacy.
	Violation	Feature that allows user to take action against privacy-violated content/user.
Obfuscation	Imprecision	Feature that allows user to lower the detail of private information disclosed.
	Inaccuracy	Feature that allows user to provide false information to protect their privacy.
	Vagueness	Feature that allows user to include linguistic terms i.e. “near”, “around” in providing their information.

IV. PROPOSED PRIVACY MANAGEMENT FEATURES

In this section, the proposed privacy management features are explained in details. These features will be used further in the content analysis on existing geosocial networking applications.

A. *Service Access*

Service access is a feature that allows the application to access the services of mobile device such as GPS location data, contact information etc. to retrieve data and use it on application. In social network, users can share self-generated content, as well as data and information that is automatically obtained from embedded sensors in mobile device [15]. Most mobile social networking applications show a popup dialogue requesting permission to access location information from the user. Some applications only provide a notification message to notify on automatic acquiring of user location information without allowing users’ control over that process [16]. These accesses sometimes are done without user’s conscience. Without proper management, the location tracking service might violate user’s privacy. Therefore, a privacy feature that allows user managing the service access is needed.

B. *Specific Grouping*

Specific grouping feature allows user to create a group within a social network based on interest, hobby etc. People prefer to use closed-type social networking services which allow certain participants, who are only invited to the group, to communicate together on a basis of small groups, such as family, friends, alumni, and school clubs. This feature motivates user to utilize social network more as it promotes better quality of interaction and privacy.

C. *Visibility Setting*

The concept of visibility is widely discussed in regards to social networking privacy [17]. Visibility refers to level of easiness for other peoples to view user profile, information, or posts. Generally, the higher visibility of user information to other peoples, the lesser users’ control over their privacy. As most social networking platforms offer association and connectivity, users have the ability to view other users’ profiles directly or through a common connection. Therefore, a geosocial networking application should offer visibility setting feature as a privacy-preserving mechanism that support users’ controlling decisions regarding who can view their information such as location information or being ‘nearby’ as named in many application [16].

Fig. 1. Proposed Visibility Setting based on Onion Metaphor.

Onion metaphor from Social Penetration Theory [18] will be adapted in this feature. Figure 1 illustrates the visibility setting illustration which adapts the idea of onion metaphor. The outer layer shows lower privacy level thus leads to disclosure on more common audience. While the inner layer shows higher privacy level and the disclosure is more on specific target.

D. Activity Log

Activity log is a feature that allows user to review their activities in social network. Many geosocial networking users are not aware that a deleted post is not permanently removed from the geosocial networking platform and still can be accessible due to the ease at which information can be saved, shared, and reposted [17]. Some geosocial networking platforms do not offer users' control over their activity stream, making them unaware of all the events that are added to their activity stream, nor who has access to their activity stream [17]. Generally, activity stream can be referred as the timeline or feed that display all of a particular user's activities such as posts, share, and likes. In addition, access control option over sensitive information like location history and health data is an essential requirement in geosocial networking application. For instance, it is important for a user to have access control option, that able to control friend's view of the availability duration of users' locations history when they visited several locations during a certain period of time [16].

E. Tagging

In order to facilitate personal information sharing users' interactions, most geosocial networking application provide four basic functionalities including publishing, recommending, tagging, and pushing [19]. Tagging functionality enables a user to make a reference to their friends' usernames when the user publishes social activities in their account, thus motivating users' interactions. However, this feature introduces the ownership of such information belongs to multiple users. The co-owners will have the responsibility and right in managing the privacy of that information. Therefore, a privacy feature is needed to manage the privacy of tagged information. It is complicated as it involves many users. Turbulence can be happened easily if the privacy is not well coordinated.

F. Classification

Geosocial networking users can be classified into several profile based on their privacy preferences [20]. As a result, users may not demonstrate same behaviours when regulating their privacy management strategies. Consider Anne, a user who is concerned about her privacy, but at the same time likes to make her whereabouts known to her friends. She may wish that at certain circumstances her location information will be viewed only by a selected close friends. For instance, when she is at home with her mother on weekends morning. Therefore, user classification is needed to help user in disclosing the information to correct audience. The difference between this feature and specific grouping is that how user classify others will not affect other people's experience, while specific grouping is based on a collective understanding among multiple users.

G. Privacy Policy

An online privacy policy is a statement that informs users how a service provider handles (e.g., collect, use, access, control) users' personal information [21]. Upon joining or signing up for a geosocial networking application, a user must agree to the privacy policy provided by the application provider. Previous study [22] has emphasized the important role of privacy policy in users' perceived privacy. When a privacy turbulence happens, an inexperienced social network user can refer to privacy policy to understand the choice he has to overcome the turbulence.

H. Education

Many social network users report difficulties in managing their privacy settings [23]. Therefore, privacy experts suggest that users must be given exhaustive control over their privacy to help them regulate their privacy boundaries. In the context of geosocial networking, privacy education is often manifested in the form of notifying users of information sharing practices. The notification is done through textual notices embedded in privacy authorization dialogues [24], and visual icons [25]. In addition, tips about a privacy feature may serve as a helpful reminder to the user prior knowledge [26].

I. Notification

Notification is a feature that provides a popup message to remind users on their incompleteness of privacy configuration or potential privacy risk. The lack of knowledge and awareness of various security tools and option available in smartphones is one of the main contributing factor of data breaches that implicated users' privacy. One of the notification approaches to support privacy decisions is privacy nudging. Nudges refers to soft paternalistic intervention that influence user behavioral and decision making, while allowing user privacy decision to be revised if their expectation is not met [27]. As nudging acknowledges that subtle differences in application design can possibly affect users, nudges are usually in a form of persuasive cue and were embedded in a notification through feedback, defaults, norms, and saliency of features [28]. These notifications help users to better understand their privacy right and configuration, and increase their awareness on privacy threats.

J. Violation

Geosocial networking application must implement certain counter measures when user or data privacy is violated. To suit the context of this study, the scope of interpersonal privacy protection behaviours is broadened to the management of relational boundaries (e.g. friending, and unfriending), territorial boundaries (e.g. untagging posts or photos or deleting unwanted content posted by others), network boundaries (e.g. hiding one's friend list from others), and interactional boundaries (e.g. blocking other users or hiding one's online status to avoid unwanted chats) [29].

K. Inaccuracy, Imprecision, and Vagueness

[30] suggested three types of imperfection that can be used in location information, namely, inaccuracy, imprecision and vagueness. In the context of location privacy, inaccuracy refers to providing different location information instead of the real location. Imprecision refers to providing location information

in the form of region to represent the real location, and vagueness refers to using linguistic terms like “far from” or “near” in the conveyed location.

V. Data Collection Methods

To select suitable application for this study, all related existing application in the mainstream digital distribution platforms which are App Store by Apple Inc. and Google Play Store are reviewed. In App Store, Social Networking category is chosen. Total number of existing applications is 246. For Google Play Store, two categories were emphasized: Social and Communication. For each category, in the Top Free section, it lists 540 related apps which is free and have the most download. Most app developers distribute their apps in both platform. To overcome the duplication problem, the duplicated apps in Play Store will be excluded.

A. Inclusion and Criteria

To obtain the most suitable sample, all related apps are filtered by applying inclusion and exclusion criteria. By doing so, it is expected to improve the study results. General exclusion and inclusion criteria were established to limit the scope of apps being evaluated.

Table II shows the inclusion criteria that will be used in this study. First two criteria have been used in identifying the app pool. For price, the app should be free to use as there is no necessity to use paid apps for academic study. Download count indirectly shows how large the social network userbase is. Although it might not be an accurate measure as it does not reflect the number of active users, it implies the number of users who tested the app. Due to this reason, these apps have research value. This study set the minimum threshold for download count as 100,000. To ensure the app is functioning in Malaysia, the location in App Store and Google Play Store was set as Malaysia. Although these apps are under Social and Communication category, some of them do not actually serve for social purposes. Therefore, this study include only social network/social discovery app.

TABLE II. Inclusion Criteria

Criteria	**Condition**
Type of OS supported	iOS/Android
Category	Social/Communication
Price	Free
Download Count	More than 100,000
Availability in Malaysia	Yes
Type of service	Social network/ Social discovery
Language	English

Among the apps searched with keywords, apps will not be reviewed if it meets one or more of the exclusion criteria. Firstly, the apps for dating and purposes will be excluded because the proposed privacy features are for geosocial network and it is decided to be inappropriate to test these features in these apps. They do not serve as a fully functional social network. The apps which is served for adult only or LGBT community are excluded as these apps contains sexually explicit content and unhealthy communication. Also, there are many apps which is only served as an additional service to a social network. For instances, Facebook Lite, Facebook Mentions, Facebook Groups are the add-on apps for Facebook. Therefore, these type of apps are excluded as well.

B. Data Sampling and Collection

Inter-coder reliability analysis is used by employing independent coders to evaluate the proposed privacy management features and get the same conclusion [31]. The reason to measure reliability is to demonstrate the trustworthiness (truthful) of the proposed features. Coders received the two-phase training and guidance from the researcher and they will evaluate the same case (apps) to maintain consistency in evaluation.

Three coders are involved in the content analysis. The coders evaluate the privacy features in the sample and code the finding in either “1” for existing feature or “0” for non-existing feature. Before content analysis starts, the coders are required to attend a training to get familiar with the coding purpose and procedure.. Two applications which are Google+ and BeeTalk have been selected as training material. The code represents the opinion of coder, in whether the feature is existing on the application. For Google+, there are differing opinions on Classification feature among coders. Meanwhile for BeeTalk, there is a perfect agreement among all features. Coders are advised to review their answer with each other to identify the cause of disagreement. If it is due to carelessness, coders can re-code their result to reach agreement.

ReCal3 (“Reliability Calculator for 3 or more coders”) [32] is an online utility that is used to calculate the inter-coder reliability coefficients for nominal data coded by the coders. The coding result will be entered into an Excel and uploaded to ReCal3. It will then calculate the reliability of the feature. Reliability means how far the agreement among coders. Higher reliability means the coders have more similar expectation and idea on that feature. This study accepts the privacy feature based on the acceptance level of reliability coefficient. The rejected features will be discarded from the study.

Based on the comparison among reliability coefficients, Krippendorff’s Alpha is the best statistic to use in this study as it has tougher standard on determining the reliability of variable. It is important to ensure the privacy features are reliable in order to conclude that such privacy features are agreed by multiple persons on their characteristics. The difference of Krippendorff’s Alpha with other statistic techniques is that it includes observed and expected disagreement. Consequently, it provides more accurate approximation of reliability. It also has three benefits which makes it a better statistic. First, it can be used for any number of coders. It also can be used for any sample sizes and different type of variables. The “bootstrapping” system allows alpha to replace missing value with existing values samples form. Alpha value of 0.667 is the minimum acceptable limit.

C. Data Screening

Figure 2 shows the process of screening the most suitable apps for this study. During the initial screening phase, Google Play Store has significantly more apps than App Store. This is due to two categories (Social and Communication) were selected into the screening process. Meanwhile App Store only has categorized all the related apps into Social Networking, therefore the number of apps appeared to be lesser.

Firstly, the applications are evaluated based on inclusion criteria. As the result shown, 230 out of 246 apps in App Store have been selected for inclusion. Meanwhile Google Play Store has only 261 out of 1080 apps selected. This is because Google Play Store has many apps that do not related to social network or social discovery, especially in Communication category which only has 30 apps included for next phase. Then the included applications are filtered based on exclusion criteria as shown in Table III.

From Table III, it can be seen that most of the applications are either add-on based on social network or social network manager. These add-ons are used to enhance the functionalities of existing social network. For example, Facebook Groups allows user to manage their group in Facebook better than using the Facebook app itself. Social network manager is used to manage different social network in one platform. User can receive information and notification from different social network in one app. It provides convenience to the users who have multiple social network accounts. Finally, duplicate apps in Google Play and Apps Store that are similarly named from the same developer were removed from the dataset, leaving 65 apps for content analysis. The content analysis of the apps were examined in the next section.

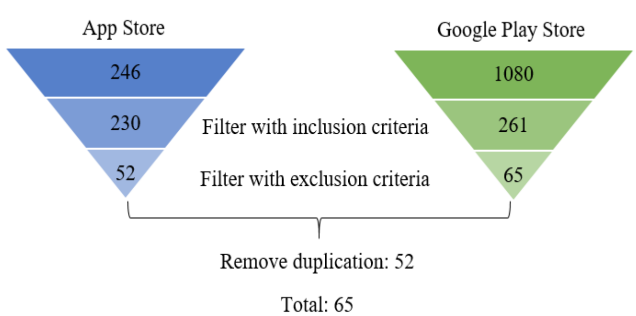


Fig. 2. Data Screening.

TABLE III. INCLUSION CRITERIA

Criteria	App Store	Play Store
Purpose	28	35
Audience Target	12	16
Nature of Service	138	155
Total	178	196

VI. RESULTS AND DISCUSSIONS

A. Inter-Coder Reliability Analysis

Table IV shows the result of the inter-coder reliability analysis. The result shows that six out of 13 proposed features are agreed by the coders to be reliable. The accepted features are specific grouping, visibility setting, privacy policy, violation, imprecision and inaccuracy.

TABLE IV. RESULTS OF INTER-CODER RELIABILITY ANALYSIS

Feature	Krippendorff's Alpha	Result
Specific Grouping	0.747028862	Accepted
Visibility setting	0.673782157	Accepted
Violation	0.709551657	Accepted
Imprecision	0.673782157	Accepted
Inaccuracy	0.754934211	Accepted
Privacy policy	0.757657902	Accepted
Education	0.533431571	Rejected
Notification	0.530925926	Rejected
Vagueness	0.260912698	Rejected
Service access	0.180339632	Rejected
Activity log	0.48760181	Rejected
Tagging	0.569097294	Rejected
Classification	0.527827293	Rejected

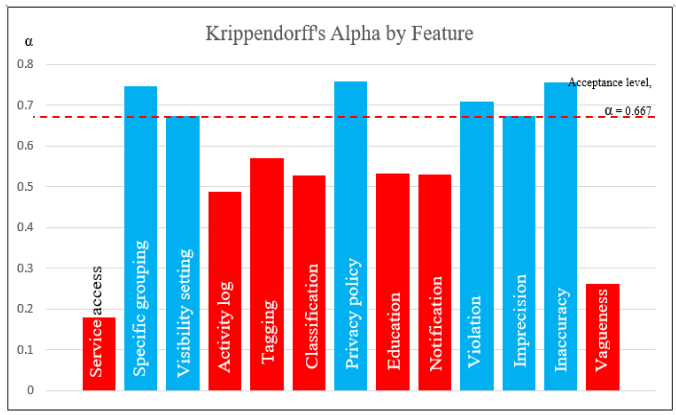


Fig. 3. Proposed Privacy Features based on Krippendorff's Alpha.

Figure 3 shows the bar chart of content analysis result. Based on the Figure 3, it can be seen that that many of the privacy management features with low reliability are from Boundary Turbulence construct as prevalence in most cases is low.

B. Discussions

Based on the inter-coder reliability analysis, the result shows that six out of 13 proposed features are agreed by the coders to be reliable. The accepted features are *specific grouping*, *visibility setting*, *privacy policy*, *violation*, *imprecision* and *inaccuracy*. From the findings, it can be seen that privacy policy feature is the most agreeable privacy feature as it has the highest reliability values (α=0.757657902). Privacy policy can be found in almost every application as it is required by law to protect user. Due to that, it is undeniable that privacy policy is a must-have feature in geosocial networking application.

Specific grouping feature is a privacy feature that allows user to define the user group to disclose information with. Based on the result, it shows a high reliability as well (α=0.747028862). This study views it as a reliable feature that cannot be neglected when developing a social network, especially for neighbourhood. User should be allowed to define smaller group within neighbourhood based on e.g. interest, hobby. By doing so, it protects the privacy of user as some information like body health conditions are sensitive to disclose, even to whole neighbourhood.

However, service access feature has a very low result value (α=0.180339632). After discussing with coders, the problem is identified. It is hard to define whether the feature is existing in an application. Some applications will prompt a popup message asking for permission to access services such as messaging and location [15]. The aforementioned privacy feature is not included in this study as the setting is done on the operating system level. This study consider such privacy feature exists only if the application provides options for user to decide whether to allow application to access service. This study put an emphasis in privacy features implementation on application level to provide a guideline for potential practitioner. However, coders are unsure on such feature existence as they have different opinion on the definition. Therefore, the agreement is very low.

Visibility setting feature allows user to define their information is visible to whom. This privacy feature is very common among social networking application as it is widely discussed in regards to social networking privacy [16]. Every private information in social network should be allowed to define its visibility by its owner. However, many social networks did not provide this feature on every information provided by user. This can be seen as the limitation of the development of social network. Nonetheless, this privacy feature is agreed by the coders as a reliable feature due to the results (α=0.673782157), which is slightly above the acceptance value. Figure 4 shows the example on how the visibility setting should be implemented in creating an event in neighbourhood geosocial network.

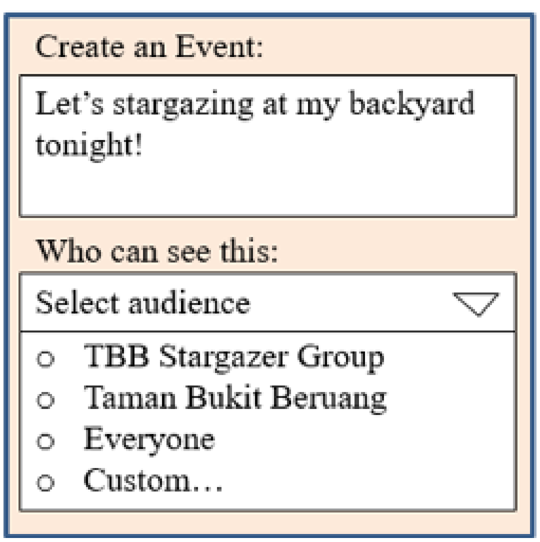


Fig. 4. Example of Visibility Setting Feature.

Activity log has low reliability values (α=0.48760181). The feature is rejected mainly due to its lack of existence in many applications. The purpose of this feature is for users to review their activity on social network to check whether any of them violates their privacy e.g. disclosed information to wrong audience. This feature provides a safety measure to revert their mistakes in order to protect their privacy. However, many applications do not implement this feature. This may imply the value of this feature is not worthy to develop. Therefore, this study does not make this an exception from rejection.

The purpose of tagging is to manage the tagged information to achieve privacy control. Therefore, a privacy mechanism is needed to manage this tagged information. However, this feature's alpha did not meet the acceptance level. Based on coder's opinion, there is a number of applications does not provide tagging function or could not identify the feature as some applications only allow including other users in messaging but not information sharing. Due to this reason, disagreements happened and the reliability value became low (α=0.569097294).

Violation is a feature to manage the privacy violated user or content. This feature acts as a counter measure for user to react on the violation. When a turbulence in a relationship happens, users can choose to revise the relationship. User can either choose to negotiate with that user, remove that relationship, or block that user without notifying. There are many ways to handle privacy violation. If that user is causing irritation or harm to the public, reporting to authority can be done. It can be seen that this feature is very important in managing privacy. It shows high reliability values (α=0.709551657) as well.

Classification is a feature to classify the user into type such as friend, family, close friend, etc. It works similar to specific grouping except it emphasizes on defining user type and works as a personal list of multiple types of relationship. Based on the coding result, the existence of this feature is low. Also, based on the feedback of coders, they had a tough time on identifying the feature as they confused it with specific grouping. These factors contribute to low reliability values (α=0.527827293).

Education consists of any material that provides knowledge to user on privacy management. This includes tutorial on privacy configuration, FAQ (Frequently Asked Question), privacy guideline, etc. This feature is analyzed to be not reliable mainly due to differing opinions of coders. One of the coders thought that some education feature was not enough to be identified to be education, while the others thought many of the privacy features are education-tailored. Scenario like this caused disagreements and led to low reliability value (α=0.533431571). Also some studies found that educational tips could be considered annoying by users [33]. Therefore, this feature will be excluded in this study.

Notification is a feature that provides a popup message to remind users on their incompleteness of privacy configuration or potential privacy risk. It shows low reliability values (α=0.530925926). Based on coders' feedback, it is said that identifying the occurrence of this feature is difficult as these reminders usually appear after the user using the application for some time. Therefore, the coder who received notification would identify this feature as exist, and vice versa. In addition, some studies [33] have found that using notification such as nudges would likely cause annoyance among users, which may hinder effective deployment of privacy nudges.

Imprecision allows user to lower the details of disclosed information. Fr instance, instead of full house address, user can choose to disclose the approximate area of living only. This feature provides option to cater different user privacy requirement. By doing so, user can still be expected to provide truthful information willingly as they can choose what they wish to disclose. However, this feature's reliability values (α=0.673782157) are just barely passed the acceptance level. This is due to some disagreement among coders that the idea of imprecision might exist in an application, but not for all the information. imprecision feature might occur in personal information but not location information. Therefore, in the same application, the coders might have different opinion on the existence of imprecision. However, this study does not reject this feature even though it might not be actually reliable. Figure 5 shows an example of imprecision technique application that is viable in geosocial network. To protect the user's location privacy, user can choose to lower the details about his location information.

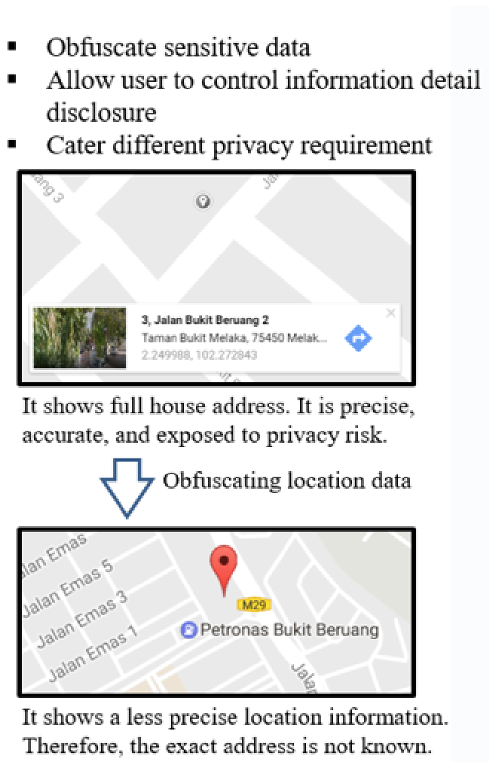


Fig. 5. Example of Imprecision Feature.

For inaccuracy feature, the reliability values (α=0.754934211) are the second highest among all the features. This feature is defined as "feature that allows user to provide false information to protect their privacy". The coders consider inaccuracy feature exist if the application does not validate the information entered by user and enforce to obtain the information from sensor (e.g. location data by using GPS location). Therefore, the coding process of this feature was lenient if based on the definition. Nonetheless, this feature is still playing a significant role in managing the privacy. By providing inaccurate information to protect their privacy, user can prevent unwanted attention from other users. For example, a user wishes to conceal the fact that he is hospitalized as to prevent their neighbours from knowing and causing potential harassment. He updates his location information by manually input some tourism spots so that his neighbours may assume he is on vacation. In a more technical way, inaccuracy can be done in location privacy by showing a location shifted away from the actual coordination. This masking technique provides convenience to the user in protecting their privacy.

Vagueness is a feature that allows user to include linguistic terms i.e. "near", "around" in providing information. This feature has a very low reliability value (α=0.260912698) probably due to two reasons: 1. Its inexistence in most applications. 2. Coders' difficulty in identifying the feature. Most applications do not provide an option for user to enter these approximation terms. When user provides imprecise or inaccurate information, the information must have a value. This is the nature of how computer applications and database work. By providing a vague information, it leads to difficult data handling and processing, where cannot be handled by most applications. They can only provide such vague information as a suggestion based on user's input. Therefore, this feature is rejected.

VII. Conclusion and Future Work

In summary, this paper presented privacy management features that represents theoretical constructs derived from Communication Privacy Management theory. To provide empirical evaluation of the proposed features, a content analysis was performed on 1326 geosocial networking apps from the market. After data screening, 65 apps were analyzed using inter-coder reliability analysis. The primary findings of the content analysis showed that many of the privacy management features with low reliability are from Boundary Turbulence construct as prevalence in most cases is low. The findings show that that 6 out of 13 proposed features are deemed reliable. The reliable privacy management features are specific grouping, visibility setting, privacy policy, violation, imprecision and inaccuracy. The proposed privacy management features may aid researchers and system developers to focus on the best privacy management features for improving geosocial networking application design. Consequently, it will provide an insight on how most market players implement privacy management in geosocial networking and then adapt the reliable privacy features into geosocial networking applications.

In general, this study found 7 out of 13 privacy features with low inter-coder reliability. Based on the Krippendoff result. The low value of alpha may not necessarily reflect low level of agreement, but due to the prevalence of privacy features in all apps is very low. Therefore, the reliability of this study can be improved by corroborating other prominent privacy theories in the research of managing online privacy as present study only includes four privacy management processes in the model. In addition, further extension of this work may include the use of mix-methodology such as case study, or phenomenology to further analyze and gain more insight on privacy management in geosocial network application. Case study can be conducted by using an existing popular geosocial networking application as a case, then collecting the data about user experience. This is required to investigate and accommodate reliability issues in the present study and highlight specific user behavior with privacy features, that may be used to confirm and increase the reliability the proposed privacy management features.

Acknowledgment

This paper is funded by Global Commission on the Stability of Cyberspace (GCSC) Grant (GLUAR/HGCC/2018/FTMK-CACT/A00015). A high appreciation to Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for facilitating the work done in this paper.

References

[1] C. R. Vicente, D. Freni, C. Bettini, and C. S. Jensen, “Location-related privacy in geo-social networks,” IEEE Internet Computing, vol. 15, no. 3, pp.20-27, 2011.

[2] S.L. Low, S. Anawar, Z. Ayop, M. R. Baharon, E. Hamid, “Self-organized Population Segmentation for Geosocial Network Neighborhood”, International Journal of Advanced Computer Science and Applications, Vol. 9, No. 9, pp. 230-235, 2018.

[3] G. Lugano, and P. Saariluoma, “To Share or not to share: Supporting the user decision in Mobile Social Software applications,” International Conference on User Modeling, Springer, Berlin, Heidelberg, pp. 440-444, July 2007.

[4] C. A. Dwyer, and S. R. Hiltz, “Designing privacy into online communities,” Proceedings of Internet Research 9.0, Copenhagen, Denmark, October 2008.

[5] S. Anawar, G. P. Ananta, Z. Z. Abidin, Z. Ayop, and S. Yahya, “Conceptualizing autonomous engagement in participatory sensing design: A deployment for weight-loss self monitoring campaign,” In 2013 IEEE Conference on e-Learning, e-Management and e-Services (IC3e), pp. 111-116, Dec 2013.

[6] C. Dwyer, S. R. Hiltz, M. S. Poole, J. Gussner, F. Hennig, S. Osswald, and B. Warth, “Developing reliable measures of privacy management within social networking sites,” In 2010 43rd Hawaii International Conference on System Sciences, IEEE, pp. 1-10, Jan 2010.

[7] H. Cho, B. Knijnenburg, A. Kobsa, and Y. Li, “Collective Privacy Management in Social Media: A Cross-Cultural Validation,” ACM Transactions on Computer-Human Interaction (TOCHI), vol. 25(3), pp 17, 2018.

[8] D. Wilkinson, S. Sivakumar, D. Cherry, B. P. Knijnenburg, E. M. Raybourn, P. Wisniewski, and H. Sloan, “Work in Progress: User-Tailored Privacy by Design”, 2017.

[9] S. Gambs, O. Heen, and C. Potin, “A comparative privacy analysis of geosocial networks,” In Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, pp. 33-40, Nov 2011.

[10] T. Dehling, F. Gao, S. Schneider, and A. Sunyaev, “Exploring the far side of mobile health: information security and privacy of mobile health apps on iOS and Android,” JMIR mHealth and uHealth, vol. 3(1), 2015.

[11] S. Petronio, “Communication boundary management: A theoretical model of managing disclosure of private information between marital couples,” Communication Theory, vol. 1(4), pp.311-335, 1991.

[12] S. Petronio, “Boundaries of privacy: Dialetics of disclosure,” Albany, NY: State University of New York Press, 2002.

[13] S. T. Margulis, “Three theories of privacy: An overview,” In Privacy online, Springer, Berlin, Heidelberg, pp. 9-17, 2011.

[14] B. Carbunar, R. Sion, R. Potharaju, and M. Ehsan, “Private badges for geosocial networks,” IEEE Transactions on Mobile Computing, vol. 13(10), pp.2382-2396, 2014.

[15] A. S. Teles, F.J.D.S. e Silva, and M. Endler, “Situation-based privacy autonomous management for mobile social networks,” Computer Communications, No. 107, pp.75-92, 2017.

[16] R. Ajami, N. Al Qirim, and N. Ramadan, “Privacy issues in mobile social networks,” Procedia Computer Science, No. 10, pp.672-679, 2012.

[17] J. Fox, and J.J. Moreland, “The dark side of social networking sites: An exploration of the relational and psychological stressors associated with Facebook use and affordances,” Computers in Human Behavior, No. 45, pp.168-176, 2015.

[18] I. Altman, D. Taylor, “Social penetration: The development of interpersonal relationships,” Holt, Rinehart and Winston, NewYork, 1973.

[19] Y. Li, Y. Li, Q. Yan, and R.H. Deng, “Privacy leakage analysis in online social networks,” Computers & Security, No. 49, pp.239-254, 2015.

[20] P.J. Wisniewski, B.P. Knijnenburg, and H.R. Lipford, “Making privacy personal: Profiling social network users to inform privacy education and nudging,” International Journal of Human-Computer Studies, No. 98, pp.95-108, 2017.

[21] N.F. Awad, and M.S. Krishnan, “The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization,” MIS quarterly, pp.13-28, 2006.

[22] Y. Chang, S.F.Wong, C.F. Libaque-Saenz, and H. Lee, “The role of privacy policy on consumers’ perceived privacy,” Government Information Quarterly, 2018.

[23] H.R. Lipford, A. Besmer, and J. Watson, “Understanding Privacy Settings in Facebook with an Audience View,” UPSEC, 8, pp.1-8, 2008.

[24] B.P. Knijnenburg, and A. Kobsa, “Making decisions about privacy: information disclosure in context-aware recommender systems,” ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 3(3), p.20, 2013.

[25] J.Y. Tsai, S. Egelman, L. Cranor, and A. Acquisti, “The effect of online privacy information on purchasing behavior: An experimental study,” Information Systems Research, vol. 22(2), pp.254-268, 2011.

[26] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L.F. Cranor, and Y. Agarwal, “Your location has been shared 5,398 times!: A field study on mobile app privacy nudging,” In Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 787-796, ACM , April 2015.

[27] S. Kankane, C. DiRusso, and C. Buckley, “Can We Nudge Users Toward Better Password Management?: An Initial Study,” In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, p. LBW593, ACM, April 2018.

[28] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L.F. Cranor, S. Komanduri, P.G. Leon, N. Sadeh, F. Schaub, M. Sleeper, and Y. Wang, “Nudges for privacy and security: understanding and assisting users’ choices online,” ACM Computing Surveys (CSUR), vol. 50(3), p.44, 2017.

[29] P.J. Wisniewski, A.K.M. Najmul Islam, H.R. Lipford, and D.C. Wilso, “Framing and Measuring Multi-dimensional Interpersonal Privacy Preferences of Social Networking Site Users,” Communications of the Association for information systems, vol. 38(1), 2016.

[30] Y. Liu, Z. Yang, X. Wang, and L. Jian, “Location, localization, and localizability,” Journal of Computer Science and Technology, vol. 25(2), pp.274-297, 2010.

[31] H.E. Tinsley, and D.J. Weiss, “Interrater reliability and agreement,” In Handbook of applied multivariate statistics and mathematical modeling, pp. 95-124, 2000.

[32] D.G. Freelon, “ReCal: Intercoder reliability calculation as a web service,” International Journal of Internet Science, vol. 5(1), pp.20-33, 2010.

[33] L. Jedrzejczyk, B.A. Price, A.K. Bandara, and B. Nuseibeh, “On the impact of real-time feedback on users' behaviour in mobile location-sharing applications,” In ACM Proceedings of the Sixth Symposium on Usable Privacy and Security, p. 14, July 2010.

Differentiation of Brain Waves from the Movement of the Upper and Lower Extremities of the Human Body

Brian Meneses-Claudio
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades
Lima, Perú

Witman Alvarado-Diaz
Student Member, IEEE
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades, Lima, Perú

Avid Roman-Gonzalez
Senior Member, IEEE
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades, Lima, Perú

***Abstract*—Currently, the study of brain waves has shown a type of alternative communication, in addition to the different applications that can be made with the brain waves obtained from each individual. The OpenBCI is an open source platform for electroencephalography (EEG), in addition to a device called Cyton Board capable of collecting brain waves and these can be sent to a computer to be processed. In this research work, a computer-machine interface is presented that may be able to collect the brain waves of individuals and process them, this is to indicate the differences between the thinking of the imaginary movement of the left and right arm, also of the leg left and right brain signals. Then, to use these brain wave differences in applications focused on people with physical disabilities.**

Keywords—OpenBCI; cyton board; extremity movement; brain wave differentiation

I. Introduction

Currently, there are different types of communication, either between two people or more people, in order to transmit information or data and the reception. New types of communication are also created due to the current need of society such as brainwave communication for people who have different physical and / or medical problems.

Neurons communicate by means of electrical impulses can be measured by equipment capable of receiving brain waves like shows in these articles [1] [2], because they have electrodes that are located on the scalp to obtain brain signals from the individual, being an alternative communication system. Each action and / or thinks of the individual will be captured and represented in brain waves, knowing the meaning of each action.

The cerebral electrical impulses are always in constant action because we never stop thinking in other words, the brain is always working continuously, then when we imagine doing actions without moving, there is also an alteration in the brain waves.

In this research work, we will present the implementation of a brain-machine interface, for the classification of electroencephalographic signals, in this case the difference of signals between the movement of the left and right arm, in addition to the difference between the left and right leg.

The purpose of this project is to obtain the differences between the signals in order implement devices capable of acting and the reading of the electroencephalographic signals, while the individual would only have to imagine the movement of some of the extremities.

II. Methodology

To obtain signals from the brain, there are multiple devices, some of which have been described in [1], [2] and [3]; for the present work we will use the system provided by OpenBCI, we will use the Cyton Board in addition to the Daisy module, which together provide us with 16 channels, for obtaining data.

It is described that the communication of the Cyton Board with the PC is done through *Bluetooth*, so the OpenBCI Dongle is used, which provides us with the necessary communication to obtain data.

Fig. 1. Cyton Board and Daisy Module.

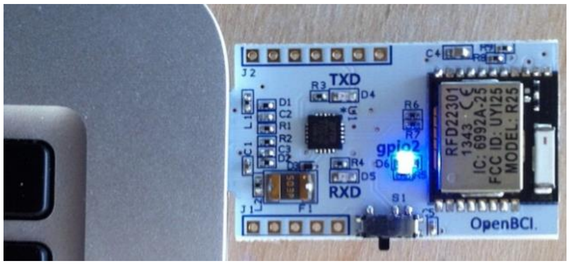


Fig. 2. OpenBCI Dongle.

For electrode placement, the international 10/20 system described in [3] is used for our case showed in figure 3, we will see the difference of the signals of the imaginary movement of the arms and legs, the electrodes will be placed in positions C3, C4, P3 and P4 [6] in addition to the reference connections in the ear lobes corresponding to connections A1 and A2.

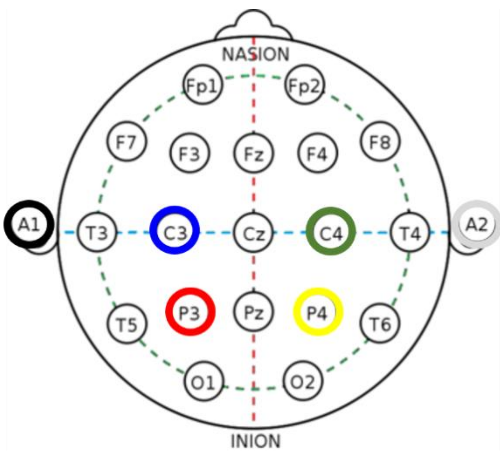


Fig. 3. Electrode Placement.

In order to improve and facilitate the placement of the electrodes in the indicated positions, the Ultracortex Mark IV system showed in figure 4 [4] was acquired, which is an open-source helmet, printable in 3D.

For the reception of data, a graphical interface was developed, which we can control the whole process of obtaining data, described in [1], it will also be possible to control the previous processing of the data.

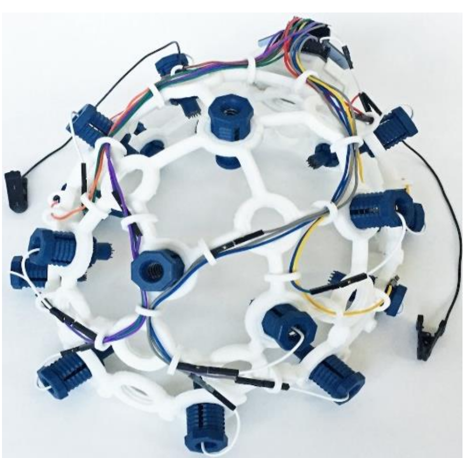


Fig. 4. Ultracortex Mark IV EEG Headset.

TABLE 1. TYPES OF BRAINWAVES

BRAIN WAVES SIGNALS	
DELTA SIGNAL	0 – 3,99 Hz
THETA SIGNAL	4 – 7,99 Hz
ALFA SIGNAL	8 – 13,99 Hz
BETA SIGNAL	14 – 29,99 Hz
GAMMA SIGNAL	30 – 70 Hz

There are different types of brainwaves, each brainwave has ranges of frequency. In the table I, it will show the types of brainwaves and the frequency of each one.

Our study is based on the Brainwave type Alfa because it shows us the differences between the channel analyzed. In the results, we show the brainwave signals of the 4-channel analyzed also, to explain about the differences and indicate the imaginary movement of the upper and lower extremities.

III. RESULTS

A. Graphic Interface

The graphic interface has 2 main buttons: the connect button is responsible for starting Python and the necessary libraries; the start button is responsible for starting the data capture with the specified parameters; in addition, the reader has three secondary buttons that are the Python button, responsible for starting only Python; the BCILAB button responsible for starting only the libraries, in addition to the analyze button, which will result in 4 graphs in which we can buy the signals obtained.

The graphic interface has an area where the user must enter basic information such as: name, repetitions, time and rest; repetitions are the number of times the system will capture data for each patient, as well as the number of different files that it will generate; in the request of time the time in which the person, to whom the data is being recorded, must remain, must be concentrated in the indications given; at rest you must enter the rest time you will have, the person to whom, are taking the data, by default it will be 3 seconds between one repetition and another.

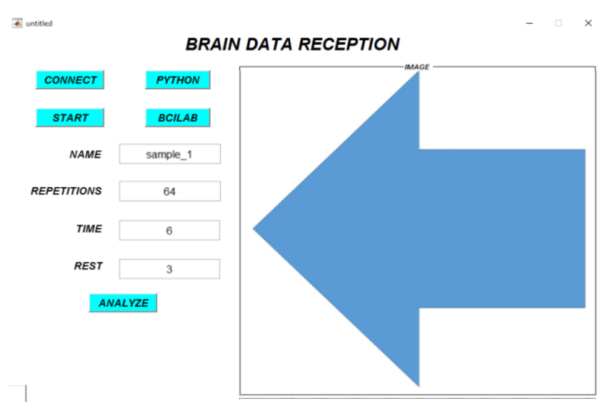


Fig. 5. Graphic Interface.

An area was placed where the respective indications will be shown for the people to whom the data is being taken, this area shows random images, which correspond to the four most known directions (right, left, forward and backward).

Finally, the user interface is as the following Figure 5, in which we can see all the elements described above.

This stage of the project is the most important because it is where all the necessary processing for the operation of the system will be carried out. Before processing, it is necessary to capture data, for which it takes 9 seconds, 6 data and 3 rest, repeating 64 times the process, this capture was made to 10 different people, being 6 women and 4 men, as Test phase, they are people with no physical disabilities, as a first study, we want to know about what they are imaging and then it will be use for people with physical disabilities.

In order to better observe the signals, 2 filters were applied, a filter with a pass band 5 to 15 Hz and a band reject filter to eliminate the 60 Hz noise [7], was implemented with the following code:

```
% FILTER REJECT-BAND
Fs=250;% sampling frequency 250
flownot=55;% lower cutting frequency
fhighnot=65;% higher cutting frequency
Ornot=1;% filter order
[bnot,anot]=butter(Ornot,[flownot,fhighnot]/(F
s/2),'stop');

% PASS-BAND FILTER
Fs=250;% sampling frequency
flow=5;% lower cutting frequency
fhigh=15;% higher cutting frequency
Or=6;% filter order
[b,a]=butter(Or,[flow,fhigh]/(Fs/2),'bandpass'
);
```

The Fast Fourier Transform FFT was found and all the signals corresponding to a single direction were averaged resulting in 4 comparative signals for each participant, in which it is shown that it is possible to differentiate the signals for the orders that were shown in the graphical interface.

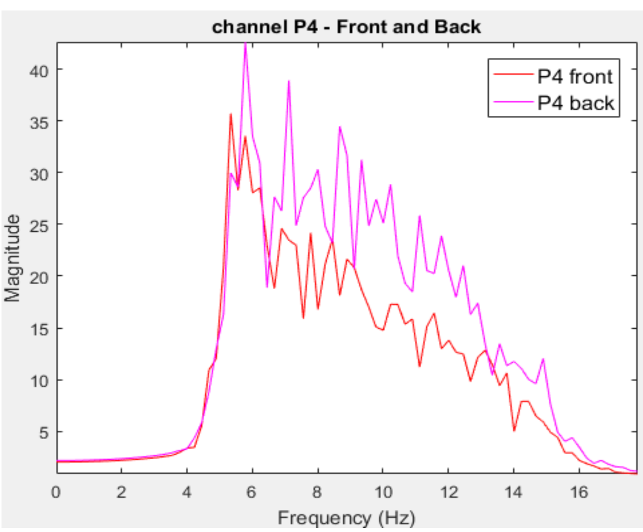


Fig. 6. Comparison in the Position P4.

In the figure 6 and 7, it will show the brainwaves in the channel P3 and P4, these signals are the imaginary movement of the left and right legs.

In the following image, we can observe the clear differentiation that exists between the two signals corresponding to two different directions, for the same electrode, which was placed in position P4 of the international system 10/20 [5]; it can mainly be observed that the most notable differentiation is between approximately 8 and 13 Hz, and these values correspond to the frequency range of the alpha waves.

As in the previous image, the comparison is also made in position P3, in which we also see that there is a clear differentiation of the signals. It should be noted that the signals that are within the range 8 - 13 Hz are the most relevant for the study [6].

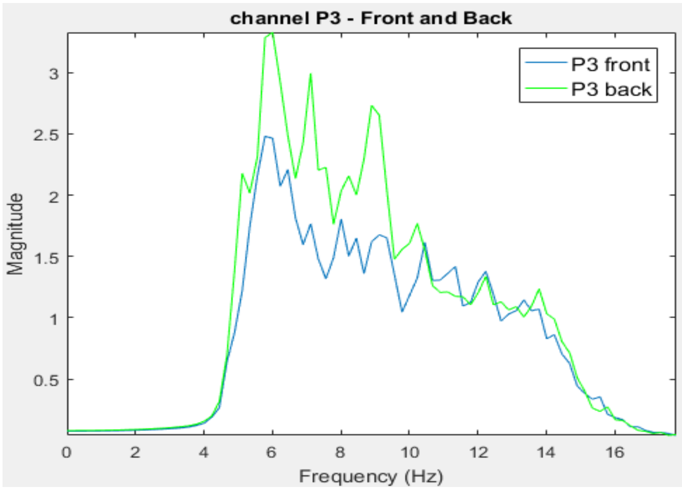


Fig. 7. Comparison in the Position P3.

In the figure 8 and 9, it will show the brainwaves in channel C3 and C4, these signals are the imaginary movement of the left and right arms.

Additionally, we analyzed the signals coming from the imaginary movement of the arms, one of which we can see the next image, which corresponds to the position C4.

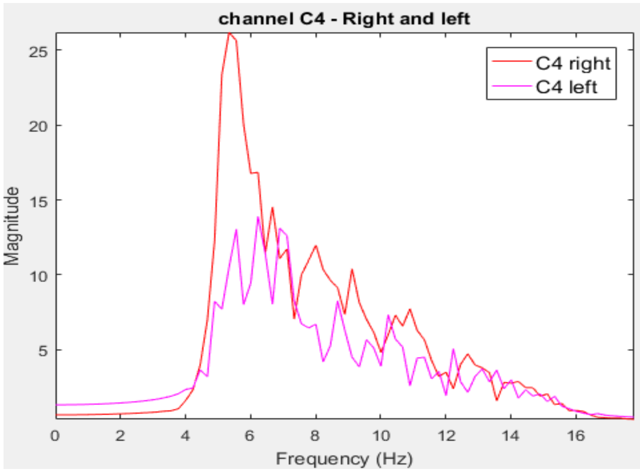


Fig. 8. Comparison in the Position C4.

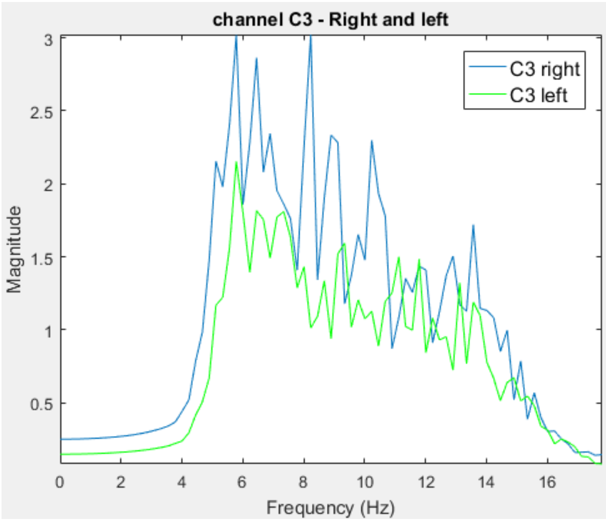


Fig. 9. Comparison in the Position C3.

Finally, we compare the signals in position C3, where we also see that it is possible to achieve a differentiation of the signals.

With the analysis of the signals, we can realize that it is possible to achieve the differentiation of the same with 4 different instructions, which were given as follows:

Arrow up = imaginary movement of the Right leg.

Arrow Backward = imaginary movement of the Left leg.

Arrow to the right = imaginary movement of the Right arm.

Arrow to the left = imaginary movement of the Left arm.

This shows that in the future, we will be able to apply differentiation algorithms of the signals, in order to control real-world objects.

Mainly the help is sought for people with physical disabilities that require electronic systems capable of actions based on the person's thinking [8] [9].

IV. Discussion

The present work confirms the processing of brain waves for the identification of the differentiation of thoughts of the movement of the upper and lower extremities, besides identifying the difference also indicate that, obtaining these results, they can be applied to electronic equipment to act at the moment of identifying the differentiation.

Currently, there are several studies that study with brain waves for processing and then their study, in this case shows the difference that exists in the imagination of the movement of the lower upper extremities.

The study of communication by brain waves is an innovative technology because it can improve the quality of life of people who have physical and / or motor disabilities.

As a future work, we are going to use these data in electronic devices for people who have physical disabilities because they have just their brain as a method of communication, so using these data, we can build a wheelchair moving by their brainwaves, it will work just imaging the movement of the upper and lower extremities.

V. Conclusions

It is concluded that, knowing the differentiation of the cerebral waves of the movement of the upper and lower extremities, electronic devices can be constructed that can work only with the action of the thought of the movements.

It is concluded that the application of an electro conductive gel was used in order to facilitate obtaining data from the electrodes, as well as to more easily identify the differentiation of brain waves.

It is concluded that the individual must be on alert because, during the analysis, the study can be somewhat ambiguous, being a bit boring for the individual. For this reason, alertness is required, in addition, to do well the tasks indicated.

The differences from the brainwaves showed in this paper are remarkable, also these data are digital so a microcontroller could use them to move electronic devices and help people with motor disabilities.

As a future work, theses data are going to be useful to make complete system for follow the instructions showed by the brain imagining movement for people with motor disabilities. Also, a wheelchair is going to be implemented and using the brainwaves, people with motor disabilities can move by themselves just imaging movement for their arms and legs.

References

[1] W. Alvarado-Díaz, B. Meneses-Claudio, and A. Roman-Gonzalez, “Implementation of a brain-machine interface for controlling a wheelchair,” in 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, CHILECON 2017 - Proceedings, 2017.

[2] B. Meneses-Claudio, W. Alvarado-Diaz, A. Roman-Gonzalez, and E. Z. Villaorduña, “Implementation of a Wireless system for the processing and comparison of cerebral waves of patients with Amyotrophic Lateral Sclerosis through matlab identifying their basic needs,” in 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, CHILECON 2017 - Proceedings, 2017.

[3] A. Roman-Gonzalez, “Communication Technologies Based on Brain Activity,” pp. 91–96, 2012.

[4] OpenBCI, “Ultracortex Mark IV | Documentation,” 2016. [Online]. Available: http://docs.openbci.com/Headware/01-Ultracortex-Mark-IV. [Accessed: 31-Jul-2018].

[5] Alcoba, S., Minguillón, J., Morillas, C., Carrillo, R., Pelayo, F., & López-Gordo, M. Á. "Plataforma inalámbrica configurable para BCI". 7o Simp. CEA Bioingeniería, Málaga, 27-32, 2015.

[6] Rodríguez, V., Jahel, G., Cárdenas, U., & Fernando, D.. Implementación de un Sistema BCI para el Análisis del Comportamiento de Bioseñales Neurológicas" (Bachelor's thesis, Universidad del Azuay), 2017.

[7] Correa Oliva, C. J. "Adquisición y procesamiento multisensorial en FPGA para EEG", 2017.

[8] Iglesias López, P. "Brain-Computer Interfaces: Desarrollo de un sistema de identificación de estados mentales alfa", 2016.

[9] Jiménez Guevara, M. P. "Sistema electrónico de iluminación (ON-OFF) mediante el control de señales cerebrales basado en tecnología EEG" (Bachelor's thesis), 2017.

Conceptual Model for Measuring Transparency of Inter-Organizational Information Systems in Supply Chain: Case Study of Cosmetic Industry

Maryam Toofani[1], Alireza Hassanzadeh[2*]
Department of information Technology Management,
Faculty of Management, Tarbiat Modares University,
Tehran, Iran.

Ali Rajabzadeh Ghatari[3]
Industrial Management Department, Faculty of
Management, Tarbiat Modares University,
Tehran, Iran.

***Abstract*—The role of information systems has changed effectively in organizational performance, and today, information systems are creating value for organizations. This study aims to provide a conceptual model for measuring the transparency of inter-organizational information systems in the supply chain. The statistical population of this research includes all managers and staff of cosmetic cosmetics companies in Tehran while companies are engaged in different sectors and their number is 500. About 218 people were questioned through the calculation with the Cochran formula. A conceptual model for measuring the transparency of inter-organizational information systems in the supply chain was developed based on a review of the theoretical concepts. Researcher made questionnaire have been used to measure variables of the research model. The validity of the research tool was confirmed by experts and the reliability of the tool was reported 0.85 by Cronbach's alpha. According to T-statistics, transparency of resources, inter-organizational trust, and environmental assurance are positive and significant in measuring transparency of inter-organizational information systems at a level of 0.01 and they are above average.**

Keywords*—Transparency of information systems; supply chain; cosmetic industry, measuring transparency; inter-organizational

I. Introduction

Research on transparency of inter-organizational information systems in supply chain is not only scientifically important, but also vital to Iran's industry in current circumstances. The problems with information systems of supply chain logistics of the Iranian cosmetic industry are not only due to its technical nature, but also due to the lack of transparency and participatory communication between various inter-institutional parts of the information systems of this chain; these problems become more complicated by continuous changes in technology. Currently, significant role of information and communication technology (ICT) in coordinating intra-organizational activities is to a degree that it is regarded as the core of supply chain management. Transparency of information systems provides new business opportunities in line with economic and executive power of organizations.

The global competitive market and the changing needs of customers have led organizations to broadly focus on improving supply chain performance [1]. A complex performance management system involves many management processes such as identifying measurement scales, setting targets, planning, communicating, monitoring, reporting, and feedback. These processes are embedded in most of the information system outputs. These system outputs perform measurement and monitoring operations on key performance indicators (KPIs) that are decisive for optimizing the supply chain performance. Performance measurement is vital for companies to improve the efficiency of the supply chain and performance [2]. The supply chain involves a network of contributors and different operating channels from within and outside the organization that affect the utility of supply chain gains [3]. Nowadays, the efficient supply chain is a potential way to maintain competitive advantage and improve organizational performance. Thus, there is no competition among organizations, but it is among supply chains and it is shown that more efficient supply chain performance improves the efficiency of the organization [4]. Moreover, the role of information systems has changed effectively in organizational performance, and today, information systems are creating value for organizations [5]. Supply chain management studies emphasize the importance of integrating in-house information and its role in the supply chain [6]. Information technology in the supply chain (SC) has changed the nature of business competition and cooperation from one company to another and from the supply chain to another supply chain [7]. In the competition of supply chains, transparency through enterprise information systems (IOS) is a key determinant of supply chain competition [8]. Definition of visibility and transparency of the supply chain means that availability of suppliers or the sharing of information in a supply chain is met as they are considered as key concepts and applications of their operations and, in their view, they take advantage of mutual benefits. Recognizing the central role of the information system in supply chain transparency and visibility, researchers in both Supply Chain Management (SCM) and Information Systems (IS) have investigated the sharing of information between enterprises through organizational information systems [9]. According to Barratt and Oke [10], the result of the transparency of sharing information can clearly lead to improved operational performance of a supply chain. Wei and Wang [11] also state that supply chain transparency has a direct impact on the strategic performance of the supply chain.

Today, the use of cosmetics is an integral part of life and as Iran is the second largest cosmetics consumer in the Middle East and the seventh importer of cosmetics in the world [12]. The cosmetics and beauty industry is one of the world's largest and most prosperous industries. According to a study about the size of the cosmetics market in Iran by the central bank and other official Iranian sources, there is no information on the share of cosmetics in the Iranian households' basket. Declaring that 30% cosmetics in the Middle East are consumed in Iran, TMBA Statistical Institute announced that Iran is the seventh country in the consumption of cosmetics industry. According to Baharat Research Center in London, the size of the cosmetics market in Iran in 2007 was estimated at $ 2 billion. According to a research conducted by the International Research Institute, now this figure in Iran is now estimated at $ 5 billion (Journal of Top Economics, 2015). In explaining the supply chain issue, it should be acknowledged that information in a supply chain is a key tool in the present day in decision making for the survival and development of enterprises. The link between all operations and processes in the supply chain is performed through information; expansion of these communications will enable firms in a supply chain to make the right decisions in order to develop themselves and maximize the profitability of the supply chain [13]. With these lines and models proposed for of information systems in the supply chain, the dimensions, and variables of transparent communications space, the existence of specific accesses, and the use of IT specific standards have been less explored. Moreover, this research focuses on transparency of inter-organizational information systems that add the innovation of this paper. Accordingly, this research has been conducted to investigate the dimensions of transparency of inter-organizational information systems in the supply chain in the cosmetic industry. Hence, the research seeks to identify dimensions of transparency of inter-organizational information systems in the supply chain of the cosmetic industry. In order to identify the dimensions of transparency of inter-organizational information systems, the variables of transparency of resources, inter-organizational trust and environmental assurance are considered as dimensions of transparency of inter-organizational information systems in the supply chain.

II. Theoretical Foundations

A. *Transparency of Information Systems:*

Information systems play the role of integration and coordination between different supply chain segments and the efficiency of this system has a direct impact on the performance of supply chain performance. Definition of visibility and transparency of the supply chain means that availability of suppliers or the sharing of information in a supply chain is met as they are considered as key concepts and applications of their operations and, in their view, they take advantage of mutual benefits. Recognizing the central role of the information system in supply chain transparency and visibility, researchers in both Supply Chain Management (SCM) and Information Systems (IS) have investigated the sharing of information between enterprises through organizational information [9].

B. *Organizational Trust:*

Organizational trust sources, for example, familiarity through reciprocal interaction, are calculated based on the interests and norms that lead to predictability and reliability. Organizational trust involves a specific relationship, because it represents the level of trust in the partner organization that has been created by the members of an organization [14].

C. *Lack of Environmental Assurance:*

Lack of environmental assurance refers to the lack of information on environmental factors that influence decision-making [15]. Researchers emphasize the importance of sharing information in support of the supply chain [16]. Sharing better information will lead to better performance [17]. This initial uncertainty of the environment faced by supply chain participants is affected by demand fluctuations and industry speed [17].

D. *Environmental Assurance:*

It is assurance about the privacy and compliance with information security in the system [16].

E. *Supply Chain:*

Supply chain involves a network of contributors and different operating channels from within and outside the organization that affects the utility of supply chain gains [3]. In addition to goods and services, this chain also includes a dynamic and continuous flow of information between different steps [1].

III. Literature Review

In a study entitled "Investigating the relationship between supply chain complexity and financial performance by reviewing the moderating role of the complexity of manufacturing companies in Khuzestan Province," Bagheri's findings [18] indicate that reducing and managing complexity in the supply chain leads to better performance, lower costs, and more integrated supply chain. Therefore, controlling and managing the level of complexity can be considered as a strategic issue for companies. This highlights the need to address complexity as a management issue. Hofmann et al. [19] investigated the relationship between supply chain management and operating costs. The research aimed to summarize and analyze what is known regarding activity-based costing (ABC) applications in the context of supply chain management (SCM). The review illustrated four main areas for further research: determination of the role of management accounting in SCM (including supply chain finance), integration of time-driven ABC with radio frequency identification (RFID) technology and automatic data collection, analysis of inter-organizational management tools in supply chains in multiple negotiation rounds, and standardization of cost accounting data in supply chains. Lee et al. [20] conducted a research on transparency of organizational information systems and supply chain performance with the aim of examining the impact of transparency of organizational information systems and supply chain performance. The results show that transparency of organizational information systems has a positive impact on supply chain performance; it was measured this study as operational performance. According to the transparency of organizational information systems, factors such as property attributes, organizational trust, complementary resources, and shared governance structures are important while lack of environmental assurance and interdependence do not have a significant effect on supply chain performance. Ignoring transparency of resources, environmental assurance, and IT standards can be considered as the weakness of this research. Ayağ et al. [21] have identified the support requirements and supply chain management strategies in the dairy industry. They used the Fuzzy Quality Function Deployment method to understand the requirements of customers and their entry into the design of the chain. Information on the quality of goods and services is one of the most important sources of uncertainty in supply chain modeling, because they are the perception and conceptual perceptions of the outside world. This paper addresses these uncertainties about the quality and customer satisfaction.

IV. Research Methodology

This research can be considered as an applied research since the present study attempts to present a conceptual model for measuring the transparency of inter-organizational information systems in the supply chain in cosmetic industry. The statistical population of the present study consists of two varieties of management experts in the supply chain with research background and industry experts who are experts in the field of cosmetic industry, which are active in their respective companies. For sampling in this study, a random sampling method is used and the number of samples is calculated using the Cochran formula. With the reference to authoritative databases and cosmetic industry companies, the number of members of the statistical community consists of academic experts and industry professionals was estimated 500 industry professionals; in this regard, 218 academic experts and industry professionals are selected as the statistical sample of the research. Descriptive and inferential statistics have been used to analyze the data in this study. In the descriptive part, mean, and standard deviation will be used; in inferential part, Kolmogorov-Smirnov tests, Lawshe content validity and one-sample t-test were used. Data were analyzed using SPSS-23 software.

In this research, an alpha value of 0.7 and higher is considered suitable for instrument reliability That reports in the Table 1. Therefore, measurements reliability is performed using Cronbach's alpha and SPSS 23 software.

TABLE I. Reliability Of The Research Tool

	Subscale	Questions	Cochran alpha
Transparency of information systems	transparency of resources	16-18	0.89
	inter-organizational trust	11-15	0.80
	environmental assurance	6-10	0.82

A. Main hypothesis:

A conceptual model for measuring transparency of inter-organizational information systems in the supply chain includes three dimensions of transparency of resources, inter-organizational trust, and environmental assurance.

B. Secondary Hypotheses:

1) Transparency of resources is a dimension of measuring transparency of inter-organizational information systems in the supply chain.

2) Inter-organizational trust is a dimension of measuring transparency of inter-organizational information systems in the supply chain.

3) Environmental assurance in environmental systems is a dimension of measuring transparency of inter-organizational information systems in the supply chain.

V. Research Findings

A. Descriptive Statistics

Table 2 reports indexes of descriptive statistics for the sample included the mean and standard deviation, and variance for the variables studied in this research.

TABLE II. Descriptive Indexes of Research Variables

Variables	Number of samples	Mean	Standard deviation	Variance
transparency of resources	218	3.34	2.96	8.65
inter-organizational trust	218	3.24	1.85	6.54
environmental assurance	218	3.74	1.36	9.69

B. Evaluation of Research Data Normality

Kolmogorov-Smirnov test was used to test the data. Given that parametric methods are used in societies with normal functions and nonparametric methods are used in non-normalized societies, normal or abnormal distribution of research data should be examined at first. Kolmogorov-Smirnov (KS) test is employed in this regard. If the research data is normal, Pearson correlation coefficient will be used and otherwise t a nonparametric method, will be used to test the research hypotheses.

TABLE III. Kolmogorov Smirnov Test To Examine Research Variables Normality (N=218)

Variables	Status	Z	Sig. (2-tailed)
transparency of resources	Independent	0.956	0.502
inter-organizational trust	Independent	0.875	0.458
environmental assurance	Independent	0.698	0.387

According to Table 3, Kolmogorov-Smirnov test shows that significance level for all variables not significant and data distribution is normal. Then, T-test and ANOVA variance are used to examine the research hypotheses. Questions from 16 to 18 of the questionnaire were used to measure the transparency variable. Frequency, mean, and standard deviation of responses are given in Table 4:

TABLE IV. FREQUENCY OF RESPONDENTS' ANSWER OF TRANSPARENCY OF RESOURCES

Items	Frequency	Mean	Standard deviation
Item 16: Use of information systems of mutual benefit	218	3.34	1.36
Item 17: Relationships Between Manufacturers of Cosmetic Industry	218	3.24	1.14
Item 18: Access to information	218	3.74	1.09

As represented in Table 4, three items have been employed to measure transparency of resources; they are responded by 218 research subjects. As shown, means response to Item 16 (3.34), item 17 (3.45) and item 18 (3.58) are obtained and all means are higher than average.

Questions from 11 to 15 of the questionnaire were used to measure the variable of inter-organizational trust. Frequency, mean, and standard deviation of responses are given in Table 5.

TABLE V. FREQUENCY OF RESPONDENTS' ANSWER TO INTER-ORGANIZATIONAL TRUST

Items	Standard deviation	Mean	Frequency
Item 11: mutual respect between managers and employees	218	3.65	1.05
Item 12: An Intimate and Intimate Relationship Between Employees and Managers	218	3.40	1.28
Item 13: Observing human aspects in a competitive environment	218	3.28	1.15
tem 14: helping each other in cases of crisis	218	3.91	1.12
Item 15: The relationship between products based on honesty and truth	218	3.63	1.10

As represented in Table 5, five items have been employed to measure inter-organizational trust; they are responded by 218 research subjects. As shown, means response to Item 11 (3.65), item 12 (3.40), item 13 (3.28), item 14 (3.91), and item 15 (3.63) are obtained and all means are higher than average.

Questions from 6 to 10 of the questionnaire were used to measure the variable of environmental assurance. Frequency, mean, and standard deviation of responses are given in Table 6.

TABLE VI. TABLE 4. FREQUENCY OF RESPONDENTS' ANSWER TO ENVIRONMENTAL ASSURANCE

Items	Standard deviation	Mean	Frequency
Item 6: Human Resources Privacy Policy	218	2.97	1.07
Item 7: Information Exchange Security	218	3.71	1.05
Item 8: industry market stability	218	3.37	1.10
Item 9: Share information between production	218	3.57	1.21
Item 10: Rules and regulations in terms of effectiveness	218	3.41	2.2

As represented in Table 6, five items have been employed to measure of environmental assurance; they are responded by 218 research subjects. As shown, means response to Item 6 (2.97), item 7 (3.71), item 8 (3.37), item 9 (3.57), and item 10 (3.41) are obtained and all means except item 5 are higher than average.

C. First Hypothesis Testing and Analysis:

Transparency of resources is a dimension of measuring transparency of inter-organizational information systems in the supply chain.

The first hypothesis suggests that the transparency is considered as an independent variable. Lawshe test is used to examine content validity of research items. Table 7 reports the value of CVE index resulted from experts' opinion; all indexes are higher than 0.50 indicating the high validity of items representing transparency of resources.

TABLE VII. CONTENT VALIDITY OF STATEMENTS IN THE QUESTIONNAIRE OF TRANSPARENCY OF RESOURCES IN SUPPLY CHAIN

Questions	Variable	Accept	Reject	CVR
16	transparency of resources	11	1	0.83
17		12	-	1
18		12	-	1

Moreover, One-sample T-test has been used to test first hypothesis. The results are shown in Table 8.

TABLE VIII. DESCRIPTIVE INDICATORS OF TRANSPARENCY OF RESOURCES IN THE SUPPLY CHAIN

Mean	Standard deviation	Mean standard error
3.34	3.11	11.70

TABLE IX. ONE-SAMPLE T-TEST FOR EVALUATING TRANSPARENCY OF RESOURCES IN THE SUPPLY CHAIN

T-statistics	Freedom degree	Significance level	Mean difference
5.57	217	0.01	0.30

According to Table 9, T-statistics of transparency of resources in supply chain is positive and significant at the level of 0.01; it is higher than average. In other words, transparency of resources is a dimension of measuring transparency of inter-organizational information systems in the supply chain.

D. Second Hypothesis Testing and analysis:

Inter-organizational trust is a dimension of measuring transparency of inter-organizational information systems in the supply chain.

The second hypothesis suggests that inter-organizational trust is considered as an independent variable. Lawshe test is used to examine content validity of research items. The following table reports the value of CVE index resulted from experts' opinion; all indexes are higher than 0.50 indicating the high validity of items representing inter-organizational trust.

TABLE X. TABLE 9. CONTENT VALIDITY OF STATEMENTS IN THE QUESTIONNAIRE OF INTER-ORGANIZATIONAL TRUST IN SUPPLY CHAIN

Questions	Variable	Accept	Reject	CVR
11	Inter-organizational trust	12	-	1
12		11	1	0.83
13		12	-	1
14		11	1	0.83
15		12	-	1

TABLE XI. DESCRIPTIVE INDICATORS OF INTER-ORGANIZATIONAL TRUST IN THE SUPPLY CHAIN

Mean	Standard deviation	Mean standard error
3.24	3.54	10.21

Moreover, one-sample T-test has been used to test second hypothesis. The results are shown in Table 11.

TABLE XII. ONE-SAMPLE T-TEST FOR EVALUATING INTER-ORGANIZATIONAL TRUST IN THE SUPPLY CHAIN

T-statistics	Freedom degree	Significance level	Mean difference
9.25	217	0.01	0.45

According to Table 12, T-statistics of inter-organizational trust (9.25) in supply chain is positive and significant at the level of 0.01; it is higher than average. In other words, inter-organizational trust is a dimension of measuring transparency of inter-organizational information systems in the supply chain.

E. Third Hypothesis Testing and analysis:

Environmental assurance in environmental systems is a dimension of measuring transparency of inter-organizational information systems in the supply chain.

The second hypothesis suggests that environmental assurance is considered as an independent variable. Lawshe test is used to examine content validity of research items. Table 13 reports the value of CVE index resulted from experts' opinion; all indexes are higher than 0.50 indicating the high validity of items representing environmental assurance.

TABLE XIII. CONTENT VALIDITY OF STATEMENTS IN THE QUESTIONNAIRE OF ENVIRONMENTAL ASSURANCE IN SUPPLY CHAIN

Questions	Variable	Accept	Reject	CVR
6	Environmental assurance	12	-	1
7		11	1	0.83
8		10	2	0.67
9		12	-	1
10		12	-	1

Moreover, one-sample T-test has been used to test second hypothesis. The results are shown in Table 14.

TABLE XIV. DESCRIPTIVE INDICATORS OF ENVIRONMENTAL ASSURANCE IN THE SUPPLY CHAIN

Mean	Standard deviation	Mean standard error
8.34	2.96	8.34

TABLE XV. ONE-SAMPLE T-TEST FOR EVALUATING ENVIRONMENTAL ASSURANCE IN THE SUPPLY CHAIN

T-statistics	Freedom degree	Significance level	Mean difference
7.89	217	0.01	0.40

According to Table 15, T-statistics of environmental assurance (7.9) in supply chain is positive and significant at the level of 0.01; it is higher than average. In other words, environmental assurance is a dimension of measuring transparency of inter-organizational information systems in the supply chain.

VI. CONCLUSION

Question 1: Is transparency of resources regarded as a dimension of measuring transparency of inter-organizational information systems in the supply chain?

Findings of the research on the first question and the role of transparency of resources in measuring transparency of inter-organizational information systems indicates with regard to T-statistics (5.57) that the role of transparency of resources is positive and significant in measuring the transparency of inter-organizational information systems at the level of 0.01; it is higher than average. In other words, transparency of resources is a dimension of measuring transparency of inter-organizational information systems in the supply chain. It can be concluded according to the experts' opinion that transparency of resources in the cosmetics industry is in a favorable position. The research findings about role of transparency of resources in measuring transparency of inter-organizational information systems are in line with the results of Nilipour Tabatabaei et al. [22], Jafari et al. [23], Mirfakhredini et al. [24], Hosseini et al. [25], Maboudi et al. [26], Shibaji Panda [27], Lee et al. [20], and Sufian et al. [28]. These researchers point to transparency of resources as a component of measuring transparency of inter-organizational information systems. In the interpretation of this finding, one can say that transparency of resources, which refers to the amount of resource assets along with the assets of partners, is in a favorable position in the cosmetic industry. Accordingly, organization will have more information from its partners and competitive advantages in the value of the supply chain will be greater as much as company resources are more.

Question 2: Is inter-organizational trust regarded as a dimension of measuring transparency of inter-organizational information systems in the supply chain?

Findings of the research on the second question and the role of inter-organizational trust in measuring transparency of inter-organizational information systems indicates with regard to T-statistics (9.25) that the role of inter-organizational trust is positive and significant in measuring the transparency of inter-organizational information systems at the level of 0.01; it is higher than average. In other words, inter-organizational trust is a dimension of measuring transparency of inter-organizational information systems in the supply chain. It can be concluded according to the experts' opinion that inter-organizational trust in the cosmetics industry is in a favorable position. The research findings about role of inter-organizational trust in measuring transparency of inter-organizational information systems are in line with the results of Mirfakhredini et al. [24], Hosseini et al. [25], Maboudi et al. [26], Shibaji Panda [27], and Lee et al. [20]. These researchers point to inter-organizational trust as a component of measuring transparency of inter-organizational information systems. In the interpretation of this finding, one can say that inter-organizational trust is in a favorable position in the cosmetic industry. Accordingly, organization will have more information from its partners and competitive advantages in the value of the supply chain will be greater as much as company resources are more. Inter-organizational trust is a key aspect of communication in the cosmetic industry, which is a kind of capital. Repeat interaction, computation based on benefits, and norms that determine predictability and reliability are counted as source of organizational trust.

Question 3: Is environmental assurance regarded as a dimension of measuring transparency of inter-organizational information systems in the supply chain?

Findings of the research on the third question and the role of environmental assurance in measuring transparency of inter-organizational information systems indicates with regard to T-statistics (7.89) that the role of environmental assurance is positive and significant in measuring the transparency of inter-organizational information systems at the level of 0.01; it is higher than average. In other words, environmental assurance is a dimension of measuring transparency of inter-organizational information systems in the supply chain. It can be concluded according to the experts' opinion that environmental assurance in the cosmetics industry is in a favorable position. The research findings about role of environmental assurance in measuring transparency of inter-organizational information systems are in line with the results of Nilipour Tabatabaei et al. [22], Jafari et al [23], Mirfakhredini et al. [24] Hosseini et al. [25], Maboudi et al. [26], Shibaji Panda [27], Lee et al. [20], and Sufian et al. [28]. These researchers point to environmental assurance as a component of measuring transparency of inter-organizational information systems. In the interpretation of this finding, one can say that Environmental assurance in the cosmetic industry leads to better information sharing and helps to realize organizational performance.

This initial uncertainty of the environment faced by supply chain participants is affected by demand fluctuations and industry speed and they can improve performance by overcoming it.

A major limitation of this study is that it was conducted in a single organization and it cannot be generalized to all organizations. In addition, the number of variables studied was limited to 6 dimensions and a quantitative method was used.

References

[1] A. Amid, Ghodsypour ,S.H., and Obrien.Christopher "A Weighted additive Fuzzy Multiobjective Model for the Supplierselection problem under price Breaks in a Supply chain", International Journal of production Economics, vol.104, no. 2, pp.382-393, 2006.

[2] C. Shphard, H. Gunter, "Measuring supply chain performance: current research and future directions", Internatonal Journal of Produtivity and performance Management, vol. 55, no. 3/4, pp. 242- 258, 2006.

[3] T.Y. Eng, "Mobile supply chain Management: challenges for Implementation", Technovation,vol.26, pp.682-686, 2006.

[4] L. Suhong, B. Ragu-Nathanb, T.S. Rangu- Nathanb, and S. Subba Raob, "Impact of Supply chain Management Practices on Competitive Advantage anad Organizational Performance", Omega the Internatinal Journal of management Science , vol.34, pp.107-124, 2004.

[5] L. Choon, and B. OH, "Evaluating Lnformation Strategic Planning and Evaluating information Sterategic planning and Evaluation System and its Application", Journal of system Integration, vol.10, pp.207-222, 2001.

[6] W. Christina, L. Kee-hung, T.C.E. Chengb, Y.H. Venus Lunb, "The role of IT-enabled collaborative decision making in inter-organizational information integration to improve customer service performance", International Journal of Production Economics,vol.159, pp.56–65, 2015.

[7] A. Rai, R. Patnayakuni, & N. Seth, "Firm performance impacts of digitallyenabled supply chain integration capabilities", MIS Quarterly, vol. 30, pp.225–246, 2006.

[8] K. K. Kim, S. Y. Ryoo, & M. D. Jung, "Inter-organizational information sys-tems visibility in buyer–supplier relationships: The case of telecommunicationequipment component manufacturing industry" Omega, vol. 39, pp. 667–676, 2011.

[9] K. Zhu, "Information transparency of business-to-business electronic mar-kets: A game-theoretic analysis", Management Science, vol. 50, pp. 670–685, 2004.

[10] M. Barratt, & A. Oke, "Antecedents of supply chain visibility in retail supplychains: A resource-based theory perspective", Journal of Operations Management, vol. 25, pp.1217–1233, 2007.

[11] H.-L. Wei, & E. T. G. Wang, "The strategic value of supply chain visibility:Increasing the ability to reconfigure", European Journal of Information Systems, vol. 19, pp.238–249, 2010.

[12] M. Kiani, F. Moguei, "Socioeconomic factors affecting the use of cosmetics by girls under age 20 Yazdanshahr, Najafabad", Journal of Skin and Beauty, vol. 4 no. 1, pp. 1-9, 2013.

[13] R Samiezadeh, & R Hosseini "Supply Chain Management (IT) and Information Technology (SCM)", Second National Conference on Information Technology, 2005.

[14] A. Zaheer, B. McEvily, & V. Perrone, "Does trust matter? Exploring the effectsof interorganizational and interpersonal trust on performance", Organization Sci-ence, vol. 9, pp. 141–159, 1998.

[15] K. K. Kim, S.H. Park, S. Y. Ryoo, & S. K. Park, "Inter-organizational cooperationin buyer–supplier relationships: Both perspectives", Journal of Business Research, vol. 63, pp. 863–869, 2010.

[16] F.T.S. Chan, A.Y.L. Chong, L. Zhou,"Anempirical investigation of factors affecting ecollaboration diffusion in SMEs". International Journal of Production Economics, vol. 138, no. 2, pp.329–344, 2012.

[17] C.W.Y. Wong, C.Y. Wong, and S. Boon-itt, "Green service practices: performance implications and the role of environmental management systems", Service Science, 2013.

[18] M. Bagheri, "Investigating the relationship between supply chain complexity and financial performance by reviewing the moderating role of complexity leverage of manufacturing companies in Khuzestan Province", Shahid Chamran University of Ahvaz, Faculty of Economics and Social Sciences, 2015

[19] E. Hofmann, E. Hofmann, J. Bosshard, & J. Bosshard, "Supply chain management and activity-based costing: Current status and directions for the future", International Journal of Physical Distribution & Logistics Management, vol. 47, no. 8, pp. 712-735, 2017.

[20] J. Lee, H. A. Kao, & S. Yang, "Service innovation and smart analytics for industry 4.0 and big data environment", Procedia Cirp, vol. 16, pp. 3-8, 2014.

[21] Z. Ayağ, F. Samanlioglu, & G. Büyüközkan, "A fuzzy QFD approach to determine supply chain management strategies in the dairy industry", Journal of Intelligent Manufacturing, vol. 24, no. 6, pp.1111-1122, 2013.

[22] S. Nilipour Tabatabaei, B. Khayambashi, B. Karbasian, & M. Shariati, "Optimization of Information Technology Application in Supply Chain Management and Marketing of Air Products by AHP Method", New Marketing Research, vol. 2, no. 2, pp.163-143, 2012.

[23] A. Jafari, M. Forghani, & A. Forghani "Supply Chain Performance Evaluation by Fuzzy Analytical Hierarchy Process (A Case Study of a Food Industry)", Commercial Letter, vol. 15, no. 61, pp.21- 49, 2011.

[24] S. Mirfakhredini, H. Andlib Ardekani, D. Rezaei, M. Rezaei Asl, "Using Multi-Attribute Decision Making Techniques to Assess Supply Chain Risk Factors (Case Study: Information Technology for Small and Medium Enterprises)", Industrial Management Studies, vol. 8, no. 21, pp.107-130, 2011.

[25] S. Hosseini, M. Muhammadi, & M. Pishvaei, "Supply chain strategy and production system selection", Strategic Management Studies, vol. 1, no.2, pp.89-112, 2010.

[26] M. Maboudi, H. Javanshir, A. Rashidi, & P. Valipour, "Investigating the Effect of Supply Chain Management on Customer Satisfaction in the Textile Industry", Textile Technology (Textile Science and Technology), vol. 5, no.1, pp.13 – 23, 2009.

[27] P. Shibaji, "Coordination of a socially responsible supply chain using revenue sharing contract", Transportation Research, vol. 67, pp.2–104, 2014.

[28] Q. Sufian, T. Monideepa, "Supply chain information system sstrategy :Impacts on supply chain performance and firm performance", International Journal of Production Economics, vol. 147, pp.40–350, 2014.

Biological Feedback Controller Design for Handwriting Model

Mariem BADRI[1] Afef ABDELKRIM[3]
Laboratory of Automation Research (LA. RA), (ENIT)
National School of Engineering of Carthage (ENICarthage)
Tunis, Tunisia

Ines CHIHI[2]
Laboratory of Automation Research (LA. RA)
National School of Engineering of Tunis
Tunis, Tunisia

***Abstract*—This paper deals with a feedback controller of PD (proportional, derivative) type applied to the process of handwriting. The considered model for this study describes the behavior of the system "hand and pen" to forearm muscles forces, applied for the production of handwriting. The applied approach considers memory recall of error signal between model outputs and experimental data to reach a desired trajectory position, a rapid dynamic and stable model response. The control technique is applied in order to expand the handwriting model response to a larger database of graphic traces. The obtained results illustrated the reliability of closed loop control to command the handwriting system, and to ensure its robustness against unknown inputs such as muscles forces that could vary from an individual to another and increase model complexity.**

***Keywords*—Handwriting system; biological system; feedback control; PD controller; muscles forces signals**

I. Introduction

Hand loss occurs due to different causes and has been increased in many countries. It has without exception, profound economic psychological and social impacts. A hand serves many roles. It is acting as a multifunctional tool for interacting with the world. Thus, prosthesis hand is vital for everyone that is in need of it. However, most of the proposed prosthesis offer limited and simple functions as opening, closing, grasp objects by pinching type, etc. They are unable to perform movements requiring some precision such as handwriting with conserving the individual characteristics of the writer (roundness or sharpness, inclination, regular or irregular spacing between letters, etc.). These characteristics are maintained even when we write with body segments as incongruous as the mouth and the foot.

In [1-3], the increase in functionality of hand prosthesis is mainly based on the progression of the control strategies. The most used control approaches are based on the amplifier electrical activity of the muscles, named ElectroMyoGraphy signal (EMG) which allows encoding directly the orders generated by the brain or muscles forces. These signals are strongly related to the muscles forces as mentioned in [4,5] which are used in several hand's characterisation approaches. Indeed, [6] used these forces to develop a highly realistic human hand and forearm model. In 2015, [7] proposed a stable force-myographic control for prosthetic hand using incremental learning. Robust hand force estimation is also studied recently in [8].

Other researches such as [9-12] show that the complex movement of hand writing can be modelled on the plane (x,y), from two most active forearm muscles, named "Abductor Pollicis Longus" and "Extensor Carpi Ulnaris". They assure vertical and horizontal displacement, respectively. The control theories are based on two electromyography signals EMG1 and EMG2 of these forearm muscles.

In order to refine the proposed handwriting models in the literature, we propose to ameliorate the mathematical model proposed in [12]. The biological process is compared, in this work, to a dynamic system model described by a set of two nonlinear differential equations. It is controlled by the muscles forces and incorporates viscosity coefficient and a frictional force applied during writing act between the pen tip and the writing plant.

The main contribution of this paper is to model handwriting motion by a feedback controller wildly used in literature [13-15]. Quite recently, a considered attention has been paid to field of biomedical control system and a growing interest gave emerged for applying feedback controller in this area. In [16], researchers focused on the simulation of reflexive muscle. Literature in [17] was devoted to the development of automotive safety system, active muscles response were implemented using feedback control of a nonlinear model of human arm.

In this paper, the considered controller is of PD type (i.e. Proportional and derivative). Indeed, the purpose is to reduce modelling error and improve system stability to the studied uncertain inputs, due to their nature, and increase the model dynamic. Actually, in the considered model we allow the integration of the human expertise decision in the studied system during the human handwriting motion, different elements react and intervene in the same time to produce homogeneous writing, readable and without deviation, such as the system of perception, the brain, the muscles, etc. In this context, [18,19] presented a generalized handwriting model and proved that the pen-tip position is detected by the eyes. Information are transmit to the brain for analyse and comparison with the desired position. Then, electrical activities are sent to muscles to execute the desired writing movement.

The proposed handwriting model can be applied in several areas, such as, trajectory generator for hand-robot control, neuro-prosthetic devices; rehabilitation practices and even to in diagnose of neurological disease like Parkinson and Alzheimer.

The paper is organized as follows; Section II is devoted to the experimental approach proposed for characterizing the handwriting system. Section III reports the studied model and control of handwriting system. Section IV discusses the background of our study by outlining a feedback control technique of PD type applied to control function of handwriting model, and illustrated by simulation results and analysis. Finally section V concludes with a summary.

II. Experimental Approach

In order to characterize handwriting movements, an experimental approach was proposed in [9], in Hiroshima City University, to record at the same instant cursive Arabic letters or geometric forms and two forearm muscles EMG signals. The measuring data were synchronized by sending a step signal from the parallel interface port on the computer to the data recorder. This experimentation had required the following equipment:

- Digital table of the brand "WACOM, KT-0405-RN".
- Pre-amplifiers "TEAC, AR-C2EMG1".
- Data recorder "TEAC, DR-C2".
- Bipolar surface electrodes (MEDICOTEST, Blue Sensor N-00-S).
- Computer.

Fig. 1 indicates the positioning of electrodes on the writer's arm. Electrodes, indicated by "ch1", is relative to the first muscle and those relative to the second muscle are indicated by "ch2 ".

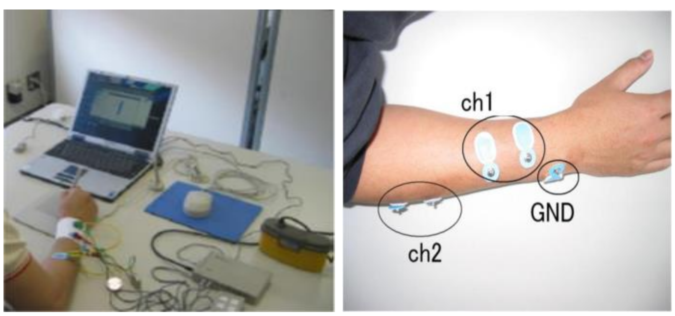
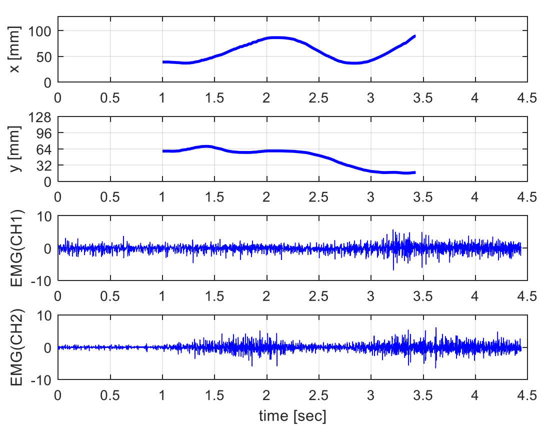
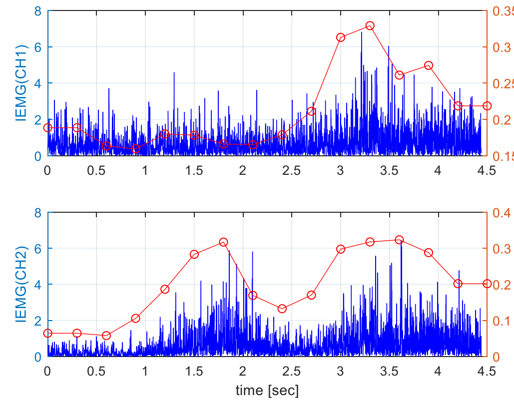


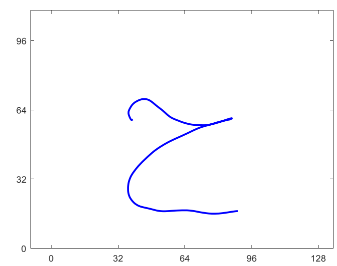
Fig. 1. Experimental assembly and electrodes' positions.



(a)



(b)

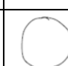
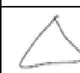
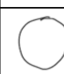
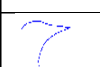


(c)

Fig. 2. The letter, ح, «HA» (a) Movement on the plane (x,y) and EMG signals (b) IEMG signals (c) Form.

In Fig. 2, are presented the recorded data for the Arabic letter, ح, "HA", from [10]. It is difficult to get the useful information from muscles activities. For this, a variety of signal processing techniques are used to make EMG waveforms easier to interpret. Indeed, the fluctuation of EMG's magnitudes can be filtered, [10], to obtain new curves - called Integrated EMG (IEMG), represented by dotted red curves in Fig. 2 (b).

TABLE I. Considered Arabic Letters and Geometric Forms

Nb	Shape description	Shape	Nb	Shape description	Shape
1	Horizontal line(1)(left/right/left)		6	Circle (2) (to the left)	
2	Horizontal line (2) (right/ left / right)		7	Triangle (1) (to the right)	
3	Vertical line (1)(up/down/up)		8	Triangle (2) (to the left)	
4	Vertical line (2) (down/ up / down)		9	Arabic letter(SIN)	
5	Circle (1) (to the right)		10	Arabic letter (HA)	
			11	Arabic letter (AYN)	

Scripters were of different gender and ages. The data base is a set of some Arabic letters, and eight basic geometric shapes, presented in table I of [9].

III. System Description

The handwriting motion process is based on the concept of memory, eyes recognition, analysis and decision made by the brain achieved by an order for muscles excitation. Fig.3 illustrates this process with the eyes as a sensor (Block 3) detecting the current pen tip position (coordinates in the plane (*x*,*y*)) during the writing motion. Then, the biological sensor sends information to the brain (Block 4). The analysis and evaluation of this is achieved by a decision and sending an order in the form of muscle activities (EMG). These signals will load the forearm muscles (Block 1) to control the arm, hand (Block 2), and execute the desired writing movement.

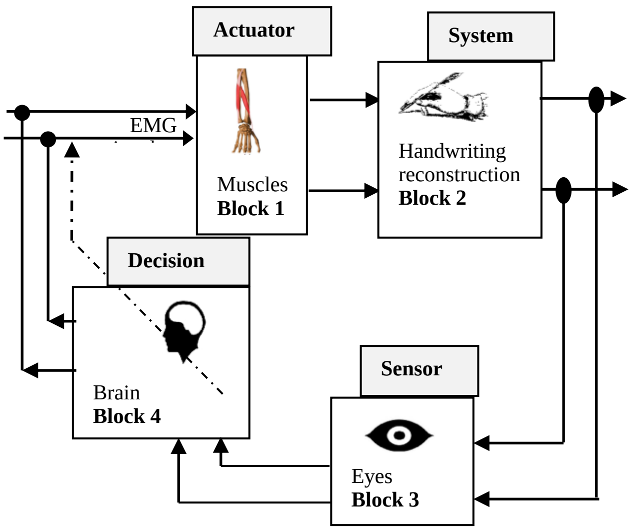


Fig. 3. Biological feedback controller design.

As we mentioned, the proposed study is an amelioration of the handwriting model presented by Yasuhara in [12]. This work takes into account the effect of stiffness relative to the hand viscosity and frictional force between the pen-tip and the writing surface, for a considered mass M, equivalent to the mass of the scripter hand and the pen. This model (Block 2 of Fig.4) is expressed by equations (1) to (3).

$$\ddot{d} = F_d - k_d \dot{d} \tag{1}$$

$$k_d = \lambda_d + \mu_d \frac{p}{v} \tag{2}$$

$$v = \sqrt{(\dot{x}^2 + \dot{y}^2)} \tag{3}$$

With

- d : the graphic trace coordinates corresponding to x and y axis,
- F_d : the equivalent force applied by the muscles on the mass M,
- k_d : the viscosity coefficient expressed in (2) with :
- λ_d : the internal hand viscosity coefficient equivalent to a mass M,
- μ_d : the equivalent stiffness coefficient between the pen-tip and writing surface,
- p : the equivalent pressure applied by the pen-tip on the writing surface,
- v : the writing velocity

The parameters λ_d and μ_d impose the dynamic of writing, and this make possible the characterizing of each writer, but also increases the model complexity.

After the estimation of these parameters and assuming that the pressure is constant, the equation in (1) could be expressed as in (4).

$$\begin{cases} \ddot{x} = F_x - (4.7 + \dfrac{0.75}{\sqrt{(\dot{x}^2 + \dot{y}^2)}})\dot{x} \\ \ddot{y} = F_y - (4.7 + \dfrac{0.75}{\sqrt{(\dot{x}^2 + \dot{y}^2)}})\dot{y} \end{cases} \tag{4}$$

Where *x* and *y* are the coordinates of the pen-tip movement on a biaxial plane. The viscosity coefficient is expressed as $k_d = (4.7 + \dfrac{0.75}{\sqrt{(\dot{x}^2 + \dot{y}^2)}})$, the non-linearity of this model is a function of the writing velocity $v = \sqrt{(\dot{x}^2 + \dot{y}^2)}$,

F_x and F_y are the muscles force applied to a mass *M* corresponding to hand movement on *x* and *y* axis, respectively.

From the differential equation system in (4), it's a 4^{th} order system and we choose:

$z \in \mathbf{R}^4$ is the state vector:

$$z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix} = \begin{pmatrix} x \\ \dot{x} \\ y \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \textit{movement on x axis} \\ \textit{velocity on x axis} \\ \textit{movement on y axis} \\ \textit{velocity on y axis} \end{pmatrix}$$

The state representation of the handwriting model is described in (5).

$$\dot{z} = \begin{pmatrix} \dot{x} \\ \ddot{x} \\ \dot{y} \\ \ddot{y} \end{pmatrix} = \begin{pmatrix} \dot{x} \\ F_x - (4.7 + \dfrac{0.75}{\sqrt{(\dot{x}^2 + \dot{y}^2)}})\dot{x} \\ \dot{y} \\ F_y - (4.7 + \dfrac{0.75}{\sqrt{(\dot{x}^2 + \dot{y}^2)}})\dot{y} \end{pmatrix} = \begin{pmatrix} z_2 \\ F_x - (4.7 + \dfrac{0.75}{\sqrt{(z_2{}^2 + z_4{}^2)}})z_2 \\ z_4 \\ F_x - (4.7 + \dfrac{0.75}{\sqrt{(z_2{}^2 + z_4{}^2)}})z_4 \end{pmatrix} \tag{5}$$

We note

$U \in \mathbf{R}^2, = \begin{pmatrix} F_x \\ F_y \end{pmatrix}$ is the control vector composed of muscular forces signals F_x and F_y

$f(z) \in \mathbf{R}^4$ is a non-linear function given by:

$$f(z) = \begin{pmatrix} 0 \\ -(\frac{0.75}{\sqrt{(z_2{}^2 + z_4{}^2)}}) z_2 \\ 0 \\ -(\frac{0.75}{\sqrt{(z_2{}^2 + z_4{}^2)}}) z_4 \end{pmatrix}$$

$A \in \mathbf{R}^{4\times4}$ is the state matrix, $A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & -4.7 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -4.7 \end{pmatrix}$

$B \in \mathbf{R}^{2\times4}$ is the control matrix, $B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$

$d \in \mathbf{R}^2$ is the output vector, with $d = \begin{pmatrix} x \\ y \end{pmatrix}$

$C \in \mathbf{R}^{4\times2}$ is the output matrix, $C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

The forces of two forearm muscles, activated during the writing act, characterize inputs of the handwriting model. The studied handwriting model estimated that the equivalent muscle forces, applied on the pen-tip, are loaded with two types of EMG signals, measured on the surface of the forearm during the movement of the writing. These forces succeeded in reconstructing handwriting graphic traces.

To determine the equivalent forces included in the motion equation, only the duration of the application of the forces F_x and F_y is considered.

Therefore, a three-level control is established, taking into account essentially the time of application of muscular forces. The position deviations relative to *x* and *y* axes at the instant t_i, are noted respectively, Δx_i and Δy_i as: $\Delta x_i = x_{i+1} - x_i$ and $\Delta y_i = y_{i+1} - y_i$

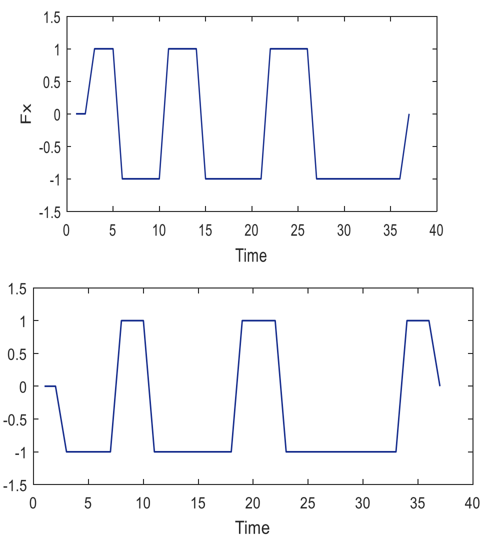
The sign of Δx and Δy indicates the direction of movement of the pen-tip (i.e. up, down, right, left) during the writing act:

$$sign(F_x) \begin{cases} \Delta x > 0, \; F_x = F_x^+ = 1 \\ \Delta x < 0, \; F_x = F_x^- = -1 \\ \Delta x = 0, \; F_x = F_x^0 = 0 \end{cases} \tag{6}$$

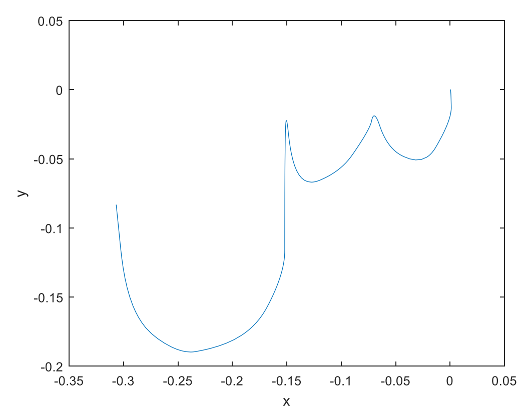
$$sign(F_y) \begin{cases} \Delta y > 0, \; F_y = F_y^+ = 1 \\ \Delta y < 0, \; F_y = F_y^- = -1 \\ \Delta y = 0, \; F_y = F_y^0 = 0 \end{cases} \tag{7}$$

Thus, if *sign* $(F_{x,y})$ changes, it indicates a change in the direction of writing pattern (i.e. from up to a down movement or from a vertical to horizontal movement and conversely, etc.)

Fig.4(a) gives an example of muscles forces exerted by muscles during the writing of the Arabic letter "SIN" in fig.4(b). From fig.4(a) it can be seen that the muscles forces form illustrate the changing of trajectory direction during the writing act. However, the muscles forces do not provide information about the nature of the movement.



(a)



(b)

Fig. 4. Referential muscular stimuli of the Arabic Letter "SIN", (a) muscular forces, (b) Letter form

IV. Feedback Controller to the Handwriting Model

In this section, we illustrate different results produced by handwriting model controller and thereafter with the proposed approach based on the feedback controller to characterize the handwriting process.

A. Controller for the Handwriting System

As described earlier for the system (4), the state space representation of handwriting system could be expressed in (8):

$$\begin{cases} \dot{z} = Az + f(z) + BU \\ d = Cz \end{cases} \tag{8}$$

Firstly, for the current writing instant t_i, where $i \in \mathbb{N}$, equation (9) describe the considered model controller:

$$U(t_i) = sign(U) \times t_i \tag{9}$$

The figure 5 illustrates the response of the handwriting model for three Arabic letter, "SIN" in fig.5(a), "HA" in fig.5(b) and "AYN" in fig.5(c). Simulations were also realized for horizontal line in fig.5(d) and a circle in fig.5(e) and a triangle in fig.5(f). The model's outputs are represented in dashed line and the experimental recorded data by a continuous line.

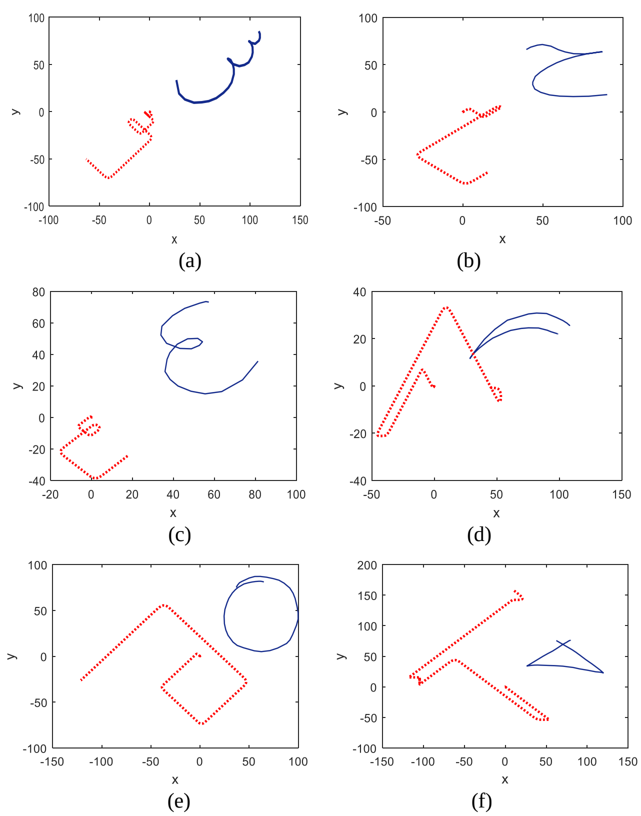


Fig. 5. Response of the handwriting model to a linear control generated from experimental database (a) Arabic Letter "SIN", (b) Arabic letter "HA", (c) Arabic letter "AYN", (d) Horizontal line, (e) Circle, (f) Triangle.

According to these simulation results, there is a considerable deviation between the experimental data and the model outputs, expressed in "(10)":

$$\begin{cases} r_x(t_i) = x(t_i) - x_{ex}(t_i) \\ r_y(t_i) = y(t_i) - y_{ex}(t_i) \end{cases} \tag{10}$$

where $d_{ex} = x_{ex}, y_{ex}$ are the coordinates of experimental recorded data and d=x,y are outputs generated by the Yasuhara model.

Fig.4 illustrates errors signals according to x movement presented in continuous line, and y movement in dashed line. Fig.6(a) is related to Arabic letter "SIN" and fig.6(b) to Horizontal line.

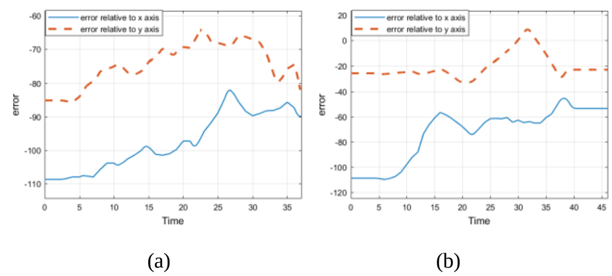


Fig. 6. Error signal according to x and y axis movements.

These results need to be improved. The purpose is to reduce the error signal $r_{x,y}$ between the model and the experimental data.

B. PD feedback controller

In this paper, the main purpose is to attempt a rapid dynamic without exceeding the order signal. The recalled information is transferred to motor control of the arm and hand to achieve the desired writing movement.

Equation "(11)" expresses the adding of a feedback controller to handwriting system

$$\begin{cases} \dot{z} = Az + f(z) + BU^c \\ d = Cz \end{cases} \tag{11}$$

$U_c \in \mathbf{R}^{4\times 2}$ Illustrated in (12) is a control expressed by current error between the experimental data and model outputs as equivalent to the memorial information recalled in the brain. With the adding of writing speed term, as the writing coordinates derivative. In this case, the control type is called PD; proportional and derivative.

$$U^c = U + k_d(d_{ex} - d) - k_v v \tag{12}$$

where matrices k_d and k_v are gain of position feedback and writing speed.

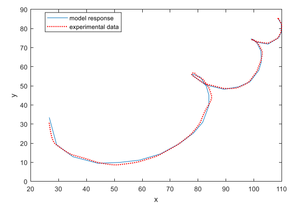
Gain matrices are chosen to satisfy the condition that the closed loop system in (13) is stable.

$$\dot{z} = Az + f(z) + B(U + k_d(d_{ex} - d) - k_v v \tag{13}$$

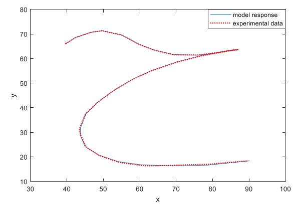
For this work, we choose the gain matrix as $k_d=I_2$ and $k_v=I_1$ witch verifies the condition of model stability, and $\lim_{t\to\infty}(d_{ex}(t)-d(t))=0$

To illustrate the results of the proposed controller, simulation of PD controller for handwriting model were performed and compared to results obtained with the 3-level controller results, given in Fig.5.

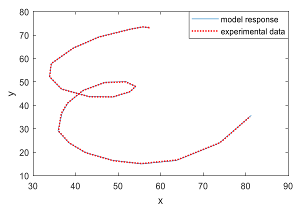
The results of simulation are illustrated in Fig.7. In these figures, model responses are presented in continuous line and the experimental data coordinates in dashed line.



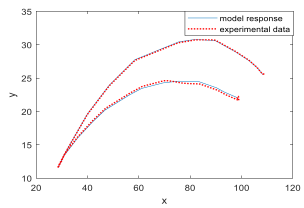
(a)Letter "SIN"



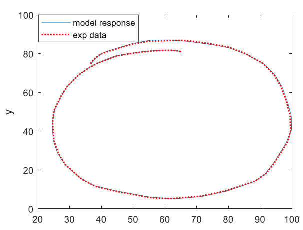
(b)Letter "HA"



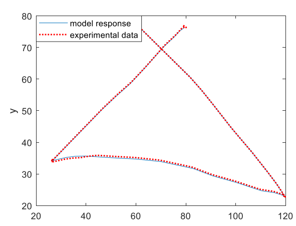
(c)Letter "AYN"



(d)Horizontal movement



(e) Form "Circle"

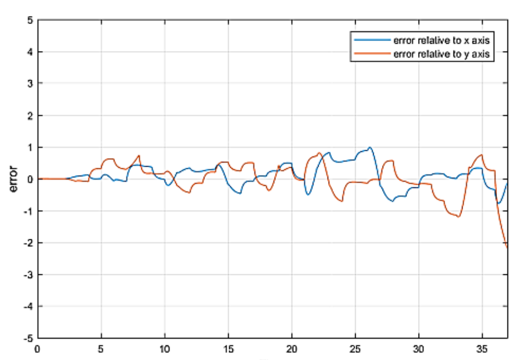


(f) Form "Triangle"

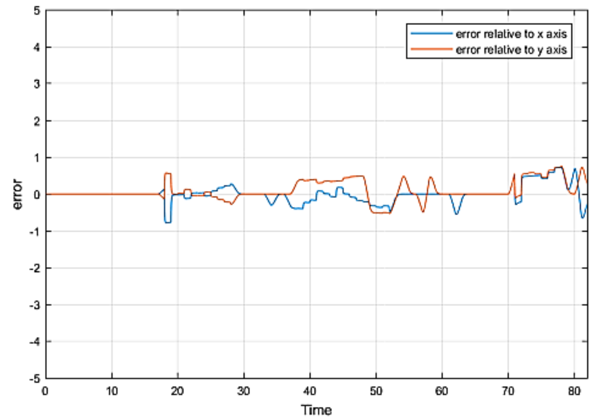
Fig. 7. Handwriting model response with feedback control.

The analysis of Fig.7 graphs show satisfactory results are obtained using a PD controller. In fact, the handwriting model response is in a good agreement with the experimental data.

For a further analysis, fig.8 depicts simulation results of error signals of handwriting model using a PD controller, for the Arabic letter "SIN" and geometric form "Triangle".



(a)



(b)

Fig. 8. Error signal form relative to x and y axis for (a) Arabic letter "SIN" and (b) Form "Triangle".

In fact, to verify this method, the error signals of Fig.8 are compared to error signals in Fig.6. According to the analysis results, it's obvious that the error is considerably reduced and almost negligible. Thus, the model response matches the experimental data. The presented results prove that the method is an effective way to improve model stability and accuracy.

However, a question still unanswered is whether the effectiveness of this approach if we further consider four muscles for the study instead of two muscles.

V. Conclusion

The main contribution of this paper is to design a feedback controller to model the handwriting motion only from two forearm forces of muscles.

A particular attention is paid to ensure the control of writing path and make the model response following the desired trajectory. Our contribution refers to a comparing between results of handwriting model controller and the proposed solution, a PD feedback controller for the control of the handwriting model. To validate the presented approach a database of experimental recorded data is considered to perform simulation results. The relevance of this control technique is illustrated in the obtained results. The presented controller has great potential for a practical application such as a robotic hand that could mimic the human handwriting as accurately as possible.

Applying feedback control technique was always considered in field of biology and biomedical and especially for handwriting system, it is necessary to ensure the model robustness against unknown inputs and prevent model uncertainties, which will be very helpful further for studying fault detection and diagnosis issues, especially with faults attempting the actuator system.

References

[1] I. Kuzborskij, A. Gijsberts and B. Caputo, "On the challenge of classifying 52 hand movements from surface electromyography," Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, *pp. 4931-4937, 2012.*

[2] L. Rossini and P. M. Rossini, "Combining ENG and EEG integrated analysis for better sensitivity and specificity of neuroprosthesis

operations,"Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, *pp. 134-137, 2010.*

[3] T. R. D. Scott, V. A. Vare. "A novel five degree of freedom user command controller in people with spinal cord injury and non-injured for full upper extremity neuroprostheses, wearable powered orthoses and prosthesis". Med Biol Eng Comput, *Vol. 51*, pp. 317–30, 2013.

[4] VT. Inman, HJ Ralston, JB Saunders, B Feinstein, and EW Wright, Relation of human electromyogram to muscular tension. Electroencephalogr Clin Neurophysiol, Vol. 4, pp. 187-194, 1952.

[5] HS. Milner-Brown, and RB. Stein, The relation between the surface electromyogram and muscular force, J Physiol, Vol. 246, pp. 549-569, 1975.

[6] J. Chalfoun, M. Renault, R. Younes and F. B. Ouezdou, "Muscle forces prediction of the human hand and forearm system in highly realistic simulation, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Vol.2, pp. 1293-1298, Sendai, 2004.

[7] M. Rasouli, R. Ghosh, W. W. Lee, N. V. Thakor and S. Kukreja, Stable force-myographic control of a prosthetic hand using incremental learning, Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4828-4831, Milan, 2015.

[8] M. Sakr and C. Menon, Study on the force myography sensors placement for robust hand force estimation, IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1387-1392, Banff, AB, 2017.

[9] M. Sano, T. Kosaku and Y. Murata, Modeling of human handwriting motion by electromyographic signals on forearm muscles", CCCT'03, 2003, Orlando-Florida.

[10] M. Benrejeb and A. El Abed-Abdelkrim, "Speed analysis of handwriting process models, IEEE International Conference on Systems, Man and Cybernetics, pp. 276-281, 2002, Yasmine Hammamet.

[11] I. Chihi, A. Abdelkrim et M. Benrejeb, Internal model Control to characterize human handwriting motion, International Arab Journal of Information Technology (IAJIT), Vol. 14, no. 6, 2017.

[12] M. Yasuhara, Experimental studies of handwriting process, Rep. Univ. Electro-Comm, Vol. 25 (2), pp. 233-254, 1975.

[13] C. Chapuis, E. Bideaux, X. Brun and N. Minoiu-Enache, "Comparison of feedback linearization and flatness control for anti-slip regulation (ASR) of an hybrid vehicle: From theory to experimental results," *2013 European Control Conference (ECC)*, Zurich, 2013, pp. 446-451.

[14] F. Abry, X. Brun, S. Sesmat, É. Bideaux and C. Ducat, "Electropneumatic Cylinder Backstepping Position Controller Design With Real-Time Closed-Loop Stiffness and Damping Tuning," in *IEEE Transactions on Control Systems Technology*, vol. 24, no. 2, pp. 541-552, March 2016.

[15] C. Ibtissem and L. Nouredine, "Design of optimal PD controller for inverted pendulum using differential evolution," *2013 International Conference on Control, Decision and Information Technologies (CoDIT)*, Hammamet, 2013, pp. 595-600.

[16] D. Salin, P.J. Arnoux, K. Kayvantash and M. Behr *Implementation of reflex loops in a biomechanical finite element model*, Computer Methods in Biomechanics and Biomedical Engineering, 19:14, 1578-1582, 2016.

[17] J. Östh , K. Brolin and R. Happee, *Active muscle response using feedback control of a finite element human arm model*, Computer Methods in Biomechanics and Biomedical Engineering, 15:4, 347-361, 2012.

[18] M. Benrejeb, A. El Abed-Abdelkrim, S. Bel Hadj Ali and M. Gasmi, A neuro-fuzzy internal model controller for handwriting process, CESA Multiconference, Lille, France, 2003.

[19] M. Benrejeb, A. Abdelkrim and M. Sano, "Sur l'étude du processus d'écriture à la main. Approches classiques et non conventionnelles," Revue e-STA, Vol. 2, 2006.

Relationship of Liver Enzymes with Viral Load of Hepatitis C in HCV Infected Patients by Data Analytics

(Data Analytics of HCV and Liver Profile)

Fahad Ahmad[1]
Department of Computer Science
Kinnaird College for Women
Lahore, Pakistan.

Kashaf Junaid[2]
Department of Clinical Laboratory Sciences
Jouf University
Kingdom of Saudi Arabia.

Ata ul Mustafa[3]
Government College
University of Faisalabad
Faisalabad, Pakistan.

***Abstract*—Correlation of liver enzyme with viral load of HCV has been previously questioned. Based on previous findings this study was aimed to appraise relationship of liver chemistry with HCV RNA titer and also to assess relationship of liver enzymes with liver morphology detected on ultrasound. And for this purpose data analytics were first time used to evaluate relationships. 155 serum samples were recruited from different hepatic centers of Lahore. Liver enzymes ALT, AST, ALP and serum bilirubin was measured by photometric method on Beckman coulter. Liver morphology was noted by ultrasound. Viral load of HCV was detected by Real time polymerase chain reaction. Relationship of liver enzymes and bilirubin with viral load and with liver morphology was observed by making UCINET graphs.Results of this study indicate that alkaline amino transferase (ALT) level, aspartate amino transferase (AST) level and alkaline phosphates (ALP) are significantly correlated with viral load of HCV RNA while biochemical test bilirubin is not. A significant relationship of liver enzymes was observed with liver morphology. Genotype 3a was the most abundant genotype of HCV in this population. Elevated levels of liver enzymes significantly depict viral load in infected patients of HCV. Findings of this study suggest another prospective study with a large population.**

Keywords*—*Hepatic; hepatitis virus; liver markers; UCINET analysis

I. Introduction

Hepatitis C is a global health problem, first reported in 1989 [1]. Hepatitis C is a widespread infection that causes significant morbidity and mortality worldwide [2]. According to a report approximately 170 million people living with chronic HCV infection [3]. There are about 10 million people with HCV infection in Pakistan. For the diagnosis of HCV many techniques have been used based on screening the antigen, detecting antibodies, measuring viral load in the patient's blood sample. Anti HCV anti bodies are detected by the ELISA (Enzyme-linked Immunosorbent Assay) method [4]. The quantification and measurement of HCV RNA viral load based on real-time PCR was recently developed for detecting and quantifying viral load in the blood sample. It is highly sensitive and has great dynamic range [5]. Other routinely tests frequently used for blood analysis are ALT (Alanine Aminotransferease test) and AST (Aspartate Aminotransferease) detecting liver functions. ALT and AST are liver enzyme whereas, AST also found in heart and muscle cells in large amount. An increase in AST and ALT levels may be because of cirrhosis and hepatitis [6,7]. Among hepatitis C virus infected patients pathogenesis of hepatitis C virus related diseases (hepatic or extra hepatic) is not well understood. This makes a difficult and challenging in the clinical and therapeutic treatment of HCV patients. Therapeutic treatment of hepatitis C virus with combination therapy of pegylated interferon and ribavirin currently first line prefers drugs. However, in extra hepatic hepatitis C virus cases a cautious approach according to case and adjustment of treatment is suggested. There is not much evidence of relationship of ALT, AST level with viral replication to liver damage in patients of HCV [8, 9]. So, this study was designed to see the relationship of different liver markers with viral load and also with the morphology of liver on ultrasound.

II. Methods

A. Study Design

A cross sectional study was designed in which 155 sero positive hepatitis C patients from the different diagnostic centers of Lahore were included. A detailed questionnaire including age, sex, monthly income, history of transfusion, vaccination were asked.

B. Blood Sampling

Prior to take blood sample from selected patients an informed consent was take. Then under the aseptic conditions 5 ml of whole blood was drawn from the medial cubital vein. From the whole blood sample serum was separated after centrifugation and then it was kept at -20°C till further processed. During sampling, more care was taken in order to avoid hemolysis.

C. Screening for Anti HCV

Immunochromatographic technique (ICT) rapid test device was used for initial screening of participants. Fastep (Polymed Therapeutics, USA) devices were used for HCV screening.

D. Serum Biochemistry

Serum alanine aminotransferease, serum aspartate aminotransferase, serum alkaline phosphatase and serum total bilirubin was measured by using Beckman coulter kits of (U.S.A) controls provided with each kit was run and all instructions provided by manufacturer were followed.

E. RNA extraction of HCV and Real Time PCR

The Systaaq Super Extract Viral RNA mini Kit was used to isolate and purify the viral RNA from blood samples. The Systaaq HCV RT-PCR kit (Systaaq Diagnostics USA), for in vitro use, was used for the nucleic acid amplification of the 5' non coding region (NCR region) of the Hepatitis genome.

F. HCV Genotyping

The AMPLIQUALITY HCV-TS kit is an IVD for the identification of Hepatitis C Virus (HCV) genotypes 1-6 by reverse line blotting technique. The kit has been validated with 5' Untranslated region (UTR region) amplicons, obtained by: Systaaq Diagnostics USA ® HCV Test (manufactured by Systaaq Molecular System), including v2.0 for use with the High Pure System (Systaaq Molecular System) and Systaaq Diagnostics USA ® Ampliprep (Systaaq Molecular System).

G. Data Analytics

All the data was analyzed by using SPSS version 21. Descriptive analysis was done to calculate the mean for each test. To draw the relationship of liver chemistry with viral load of HCV and with liver morphology a software UCINET was used [10]. This software is mostly used to analyze medical data, help in remedial strategies and in decisions making for the diagnosis of different diseases.

III. Results

In this study 155 HCV positive patients were selected, age range from 18-60 years. Of the 155 patients 76 (49.03%) were male and 79 (50.96%) were female. Among these patients 131 (84.51%) were married, 37 (23.87%) patients had donated blood in past time and 16 (10.32%) patients with the history of transfusions. Results were shown in Table1.

TABLE I. SOCIAL & DEMOGRAPHIC CHARACTERS OF HCV PATIENTS (N=155)

Characteristics		**HCV**
		N=155(100%)
Age	5 – 18	
	18-30 years	36 (23.22%)
	31-45 years	73 (47.096%)
	46-60 years	46 (29.67%)
Sex	Male	76 (49.03%)
	Female	79 (50.96%)

A. Lab Investigation

Results indicate mean hemoglobin concentration in HCV patients was 13.94 ± 14.89 (Range: 9.1-16.2 g/dl), mean platelet count was 247.67± 54.10 [(Range: 97-376 (10^9/L)] and mean TLC was 9.25±10.20 [(Range: 4.5-12 (10^9/L)].

TABLE II. LAB INVESTIGATION OF HCV PATIENTS

Biomarker	HCV	Reference Range
	Mean± S.D	
ALT	93.86 ± 40.92 (U/L)	<45 (U/L)
AST	87.35 ± 37.14 (U/L)	<45 (U/L)
ALP	251± 52.46	Up to 305 (U/L)
Bilirubin	0.84±0.07 (mg/dl)	0.6-1.0 (mg/dl)

Results of liver chemistry indicate elevated levels of liver enzymes was in HCV infected patients. In HCV patients mean ALT level was 93.86 ± 40.92 U/L (Range: 35-278), mean AST level was 87.35 ± 37.14 U/L (Range: 30-234), mean ALP was 251± 52.46 U/L (Range: 117-388) and mean bilirubin was 0.84±0.07 mg/dl (Range: 0.6 -1.0). Results indicated in table 2.

B. Correlation of Liver Enzymes with Viral Load of HCV Detected by RT PCR

Results of this study indicates that liver enzymes; alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP) are significantly associated with high viral load in HCV patients (ALT p-value <0.001, r-value 0.619, 95% CI 0.51-0.71, AST p-value <0.001, r-value 0.614, 95% CI 0.50-0.70, ALP p-value 0.01, r-value 0.20, 95% CI 0.05-0.35) while are not significantly associated with high viral load in HBV patients (ALT p-value 0.65, r-value -0.05, 95% CI -0.25- -0.16, AST p-value0.40, r-value -0.09, 95% CI -0.29- -0.12, ALP p-value 0.47, r-value -0.07, 95% CI -0.28- -0.13). However, bilirubin is not statically significant marker to evaluate HCV viral load. Results are indicated in Table 3.

TABLE III. CORRELATION OF LIVER ENZYMES WITH VIRAL LOAD OF HCV DETECTED BY RT PCR

Enzyme	**p-value**	**R-value**	**95% Cl**
ALT	<0.001	0.619	0.51-0.71
AST	<0.001	0.614	0.50-0.70
ALP	0.01	0.20	0.05-0.35

C. Relationship of Liver Enzymes with Liver Morphology in HCV Patients

Radiologist divides liver morphology into three groups, Normal Scan, Fatty Enlarged Liver and coarse Liver. Of the total 155 HCV patients, 149 had high level of alanine aminotransferase (ALT), 147 had high elevated aspartate aminotransferase (AST) and only 18 had high level of alkaline phosphates (ALP).

D. Data Analytics

Relationships were observed by drawing UCINET graphs and it reflected that the impact of ALT, AST and ALP is more on viral load when we compared with the effect of bilirubin. However, when relationship with ultrasound was observed light lines indicate weak relationship (Figure 1-4).

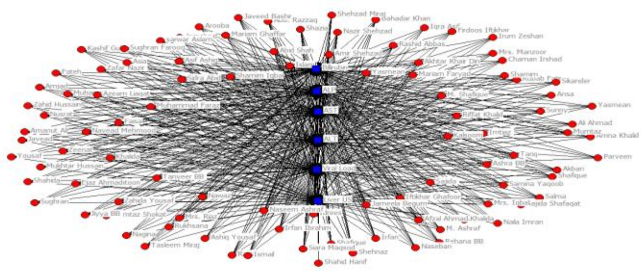


Fig. 1. Relationship of Liver Chemistry (ALP, AST, ALT and Bilirubin) with Viral load & Liver Ultrasound in HCV infected Patients: Initially Generated Saturated Graph (n=100).

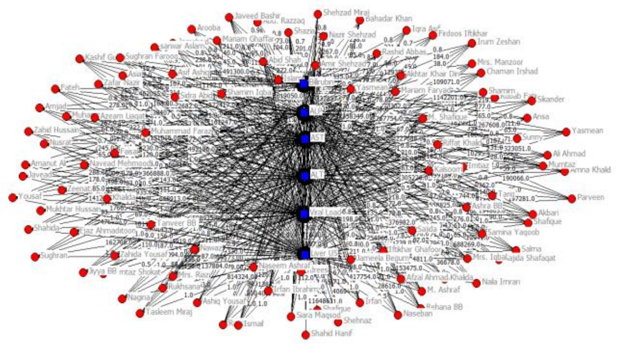


Fig. 2. Relationship of Liver Chemistry (ALP, AST, ALT and Bilirubin) with Viral Load & Liver Ultrasound in HCV Infected Patients: Initially Generated Saturated Graph with Tie's Values (n=100).

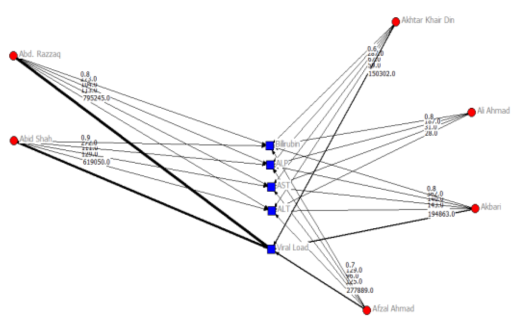


Fig. 3. Weights of Ties Indicates Tendency of Patients Towards Viral Load of HCV with ALP, AST, ALT and Bilirubin (n=6) Dark Lines Indicate Strong Relationships and Light Lines Indicate Weak Relationship.

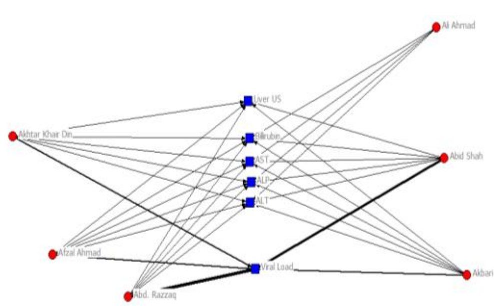


Fig. 4. Weights of Ties Indicates Tendency of Patients Towards Viral Load of HCV and Liver Ultrasound with ALP, AST, ALT and Bilirubin (n=6) Dark Lines Indicate Strong Relationships and Light Lines Indicate Weak Relationship.

IV. Discussion

Hepatitis C virus (HCV) infection effects liver, maximum number of patients with acute hepatitis C transformed into chronic HCV infection. Ultimately results in liver cirrhosis, hepatic failure or hepatoma, which is responsible for hundreds of thousands of deaths each year [11]. We conducted a study with a hypothesis that how much viral load get influenced by levels of liver enzyme in HCV infected patients.

Results of this study indicate that HCV is equally prevalent in male and female. Other studies also support that HCV is not gender specific [12]. Result shows that HCV prevalence is less in blood donors (37 patients 23.87%) may be due to the frequent screenings and almost all the blood donors do not take intravenous drugs. This is also supported by other studies that the rate of HCV infection is less in blood donors [13].

It has been previously questioned in many other studies that whether liver enzymes are correlated with HCV viral load [14,15]. ALT, AST and ALP is highly significant (p-value <0.001) with HCV RNA viral load. Liver forms ALT, AST and ALP which helps produce salts and amino acids (which are used to make proteins). Hepatitis C virus causes liver inflammation, increase in viral load means that there is increase in the liver inflammation. Rise in ALT is usually an indicator of damaged liver or inflamed liver. Level of hepatic enzymes has a trend of rise and fall at regular intervals in patient with HCV. AST is often used to monitor liver inflammation and damage in combination with other tests. Raised ALP from the liver is an indicator of blocked bile ducts caused by liver disease. Elevated levels of liver enzymes in our samples might also be due to sampling area. All samples were recruited from government hospital, where patients come usually at later stages of infection.

Correlation of bilirubin with HCV RNA viral load is not significant in this study (p value 0.68), bilirubin is not only valid biochemical marker of HCV

Patients infected with HCV have different clinical features and outcome, some are undetected, slowly progressive transferred into chronic phase and eventually develop in liver cirrhosis. Although mechanism is unclear, there are several factors including age, gender alcohol consumption, severity of infection, immunity and many others [16, 17].

Several studies have been done to evaluate relationship of abnormal liver morphology with elevated liver enzymes and conclusions are different. Some reports that ALT and particularly AST are associated with liver damage [18, 19], while puoti *et al* reported that there is no relationship between liver severity and liver enzymes. However findings indicate significant association of ALT and AST with abnormal liver morphology observed on ultrasound. ALP is not a significant marker as it also increased in bone and bowl disease.

Another important finding of this study is about genotype of HCV. HCV genotype 3a has highest prevalence among different genotypes types of HCV in population under study during this time period in infected population. It has also been previously reported that 3a is the most common genotype in Pakistan. These findings denote that inappropriate conditions most, if not all, HCV genotypes have the potential to cause epidemic. They also indicate that social, behavioral, and demographic factors (including international migration) are more important than viral genetic variation in estimating worldwide prevalence of various genotypes [20, 21].

In this research article we used UCINET software to review the weightage of liver chemistry on viral load of hepatitis c in HCV infected patients. This data analysis has not been previously used to explore the relationship. The bonding between different parameters actually reflect the impact of each variable on the output of patient that can be accessed by this type of analysis. Here in our study it has been clearly seen the impact of combination of elevated levels of liver enzymes on viral load and also on morphology of liver ultrasound that was accessed by doing liver ultrasound.

V. Conclusion

Our study emphasizes the significant relationship of liver enzymes with HCV – RNA titer and that was evident by using UCINET data analysis technique, the most common genotype in region is 3a, alkaline amino transferase and aspartate amino transferase are particularly related with fatty liver and coarse liver.

Acknowledgment

Dr. Fahad Ahmad received his Doctor of Philosophy degrees in Computer Science from National College of Business Administration & Economics, Lahore, Pakistan in 2017. He is currently an Assistant Professor at Kinnaird College For Women (A Chartered Public Sector University), Lahore, Pakistan. His research interests are in the areas of modeling and designing the data analytics, intelligent security & cryptosystem for large organizations, quantum-photonics, digital holography, fuzzy logic, verification & validation methods, numerical simulation. Here at KCFW he is actively involved in taking different undergraduate and postgraduate courses and research activities.

Dr. Kashaf Junaid has received PhD in Microbiology and Molecular Genetics from University of the Punjab, Lahore in December 2015 and also worked as an International PhD Research Scholar in Bart’s institute of Primary Health Care, Royal London Hospital, Queen Marry University of London (December 2013- July 2014).Currently, worked as Assistant Professor for 3 years in University Institute of Medical Laboratory Technology, Faculty of Allied Health Sciences, and University of Lahore. Now working as Assistant Professor in Clinical Laboratory Science department, Jouf University, Saudi Arabia. Here she is actively involved in taking different undergraduate and postgraduate courses and research activities.

Ata ul Mustafa is currently enrolled MPhil in Microbiology he is also actively participating in research.

All the participants of this study worked hardly and contributed equally so we are all grateful to each other.

References

[1] Mushtaq A, Tariq MA, Rasheed U, Afroz A, Zeeshan N, Asif AR, et al. Estimation of HCV viral load and liver enzymes among different patients groups of District Gujrat, Pakistan. Advances in Bioscience and Biotechnology. 2013;4(09):866-71.

[2] Waheed Y, Shafi T, Safi SZ, Qadri I. Hepatitis C virus in Pakistan: a systematic review of prevalence, genotypes and risk factors. World journal of gastroenterology: WJG. 2009;15(45):5647.

[3] Udompap P, Kim D, Kim WR. Current and future burden of chronic nonmalignant liver disease. Clinical Gastroenterology and Hepatology. 2015;13(12):2031-41.

[4] Swellam M, Mahmoud MS, Ali AAF. Diagnosis of hepatitis C virus infection by enzyme-linked immunosorbent assay and reverse transcriptase-nested polymerase chain reaction: A comparative evaluation. IUBMB life. 2011;63(6):430-4.

[5] Halfon P, Bourlière M, Pénaranda G, Khiri H, Ouzan D. Real-time PCR assays for hepatitis C virus (HCV) RNA quantitation are adequate for clinical management of patients with chronic HCV infection. Journal of clinical microbiology. 2006;44(7):2507-11.

[6] Berk P, Korenblat K. Approach to the patient with jaundice or abnormal liver tests. Goldman's Cecil Medicine (Twenty Fourth Edition): Elsevier; 2012. p. 956-66.

[7] Pratt D. Liver chemistry and function tests. Sleisenger and Fordtran's Gastrointestinal and Liver Disease 9th ed Philadelphia, Pa: Saunders Elsevier. 2010.

[8] Adinolfi LE, Gambardella M, Andreana A, Tripodi Mf, Utili R, Ruggiero G. Steatosis accelerates the progression of liver damage of chronic hepatitis C patients and correlates with specific HCV genotype and visceral obesity. Hepatology. 2001;33(6):1358-64.

[9] Craxì A, Laffi G, Zignego AL. Hepatitis C virus (HCV) infection: a systemic disease. Molecular aspects of medicine. 2008;29(1-2):85-95.

[10] Borgatti SP, Everett MG, Freeman LC. Ucinet for Windows: Software for social network analysis. 2002.

[11] Chevaliez S, Pawlotsky J-M. HCV genome and life cycle. Hepatitis C viruses: genomes and molecular biology. 2006:5-47.

[12] Narciso-Schiavon JL, Schiavon LdL, Carvalho-Filho RJ, Sampaio JP, Batah PNE, Barbosa DV, et al. Gender influence on treatment of chronic hepatitis C genotype 1. Revista da Sociedade Brasileira de Medicina Tropical. 2010;43(3):217-23.

[13] Gao X, Cui Q, Shi X, Su J, Peng Z, Chen X, et al. Prevalence and trend of hepatitis C virus infection among blood donors in Chinese mainland: a systematic review and meta-analysis. BMC infectious diseases. 2011;11(1):88.

[14] Zechini B, Pasquazzi C, Aceti A. Correlation of serum aminotransferases with HCV RNA levels and histological findings in patients with chronic hepatitis C: the role of serum aspartate transaminase in the evaluation of disease progression. European journal of gastroenterology & hepatology. 2004;16(9):891-6.

[15] Liu P, Li Y, Sun C-m. Correlations of serum hepatitis C virus RNA and alanine transaminase with liver histopathological changes in patients with chronic hepatitis C. Laboratory Medicine. 2009;40(3):167-9.

[16] Bagyalakshmi R, Malathi J, Prathiba K, Samson Y, Ravichandran R. A Correlative Study on Hepatitis C Virus Load Determined by Real Time Polymerase Chain Reaction with Serum Biomarkers in Patients with Renal Disease. J Mol Biomark Diagn. 2012;3(123):2.

[17] Patel A, Harrison SA. Hepatitis C virus infection and nonalcoholic steatohepatitis. Gastroenterology & hepatology. 2012;8(5):305.

[18] Ekstedt M, Franzén LE, Mathiesen UL, Thorelius L, Holmqvist M, Bodemar G, et al. Long-term follow-up of patients with NAFLD and elevated liver enzymes. Hepatology. 2006;44(4):865-73.

[19] Forns X, Ampurdanes S, Llovet JM, Aponte J, Quintó L, Martínez-Bauer E, et al. Identification of chronic hepatitis C patients without hepatic fibrosis by a simple predictive model. Hepatology. 2002;36(4):986-92.

[20] Rasheed A, Ullah S, Naeem S, Zubair M, Ahmad W, Hussain Z. Occurrence of HCV genotypes in different age groups of patients from Lahore, Pakistan. Advancements in Life Sciences. 2014;1(2):89-95.

[21] Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, Pybus OG, et al. Global distribution and prevalence of hepatitis C virus genotypes. Hepatology. 2015;61(1):77-87.

A Novel Method for Secured Transaction of Images and Text on Cloud

John Jeya Singh. T
Research Scholar, Dept of Computer Science,
Bharathiar University, Coimbatore,
Tamil nadu, India.

Dr E.Baburaj
Professor and Head, Dept of Computer Science &
Engineering, SUN Engineering College, Nagercoil,
Tamilnadu, India

***Abstract*—Implementation of privacy preservation of data on cloud storage is tedious and complex. Cloud is a third party on – demand service to hold data for a specific period. There is no assurance from the cloud storage providers about the security of data. It is necessary to provide some security to data. Cryptographic algorithms are required to provide security to the data on cloud. The aim of the research is to develop a method which combines Artificial Neural Network and Three fish algorithm for the secured transaction of images. Images are large in size and more sensitive comparing to normal text. The proposed method provides security to image with low computation cost comparing to existing methods. The research is implemented in privacy and public clouds. The generated results prove the proposed research is more efficient in terms of compression ratio, mean square error, normalized absolute error, time, and space efficiency.**

***Keywords*—Three fish; neural network; security; multimedia; cloud storage**

I. Introduction

Cloud computing (CC) is an important module, with the ability to reduce computing costs and increased efficiencies [1]. It enables convenient and flexible access to a pool of computing resources [2]. Delivery and deployment are the principal modules of CC. Rapid elasticity and measured services are the key characteristics of CC [3][4]. Rapid elasticity allow users to scale up or down resources [5]. Measured services are the derivation of business module and let cloud service providers to control and optimize the services. Software as a service (SAAS) and Infrastructure as a service (IAAS) are the key delivery modules of CC. Privacy, reliability, security, anonymity, and Government surveillance are the set of policy issues in CC [6][7]. The framework of clouds does not allow organizations to take complete control on their data. There is no assurance on privacy preservation of data from cloud service providers. It is better to implement cryptographic algorithms to provide security to data on clouds.

Public, private, and hybrid clouds are the familiar approaches to provide services and storages for users [8]. A public cloud will provide services like applications and storage facility to the users. Amazon Elastic compute cloud, IBM's blue cloud, Sun cloud, Windows Azure services, and Google app Engines platforms are the examples of public clouds. A private cloud will provide only storage facility to the clients. A hybrid cloud combines both service and storage facilities to the clients [9].

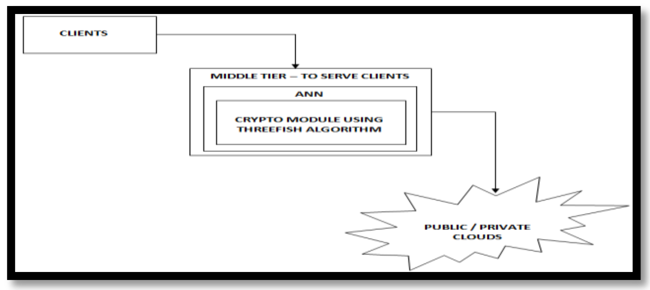


Fig. 1. Architecture of Proposed Research

The figure 1 depicts the architecture of the method proposed in the research. The figure shows the three tier architecture for the transaction of data. The cryptosystem will be placed in web server to serve the clients. The web server has the ability to serve a single or multiple clients [10].

Artificial Neural Network (ANN) is used in the research to monitor the execution of three fish cipher algorithm. The implementation of proposed research is on the web server. The web server can be installed in a client or in a separate node to serve multiple users [11]. The following part of the research will discusses literatures exist in the crypto algorithms for clouds, experimentation and results and conclusion of the research.

II. Review of Literature

In [5], author has proposed an algorithm based on Data Encryption Standard (DES) using ANN. The input layer had 64 neurons to represent cipher text. The research had other two layers for encryption and decryption processes. The computation cost for the research is more and multiple layers costs more time and space. The practical implementation of research is more difficult and complex.

Siddeeq.Y.Ameen and Ali. H.Mahdi [6] have proposed a cryptosystem based on Advanced Encryption Standard (AES) using ANN. The research has employed a non – linear neural network design to implement AES algorithm. The Rijndael algorithm, a cryptosystem is developed to encrypt and decrypt the plaintext. It is a symmetric algorithm consumes less time and more efficient than other algorithms. Levenberg – Marquardt algorithm is used to adjust the weights for ANN module. The results shows that the performance of the research is effective than existing algorithms.

Nuray At. et.al.[7], have designed a low – area co – processor for threefish block cipher to implement skein architecture. The research has altered the intrinsic parallelism of threefish and designed a pipelined ALU and interleaves multiple tasks to achieve a tight scheduling. Field Programmable Gate Array [FPGA] is used to design the low – area co – processor. The architecture of skein consists of a register file, dual – ported memory, an ALU and a control unit. The register file is designed into 64 – words and holds a plaintext block, an internal state and an extended tweak. The research has generated a better result comparing to the existing methods.

Smita Jhajharia et.al.[8], have designed an algorithm based on public key cryptography using ANN. The research has employed Hebbian learning rule to train the ANN of both sender and receiver machines. Genetic algorithm is also used in the research to optimize the output. Three – parity machine is used to calculate values obtained by hidden neurons. The results are better than the previous researches on public key cryptography.

John Jeya Singh and Baburaj.E [9] have proposed a cryptosystem for images using Blowfish and ANN. The encryption of media files ensures the security on clouds. The ANN module compresses the images and forward to the Blowfish encryption module then the encrypted image is stored in clouds. The decryption process is vice versa of the encryption process. The compression ratio, mean square error, average difference proves that the efficiency of the research.

III. Research Methodology

The objective of the research is to provide a secured transaction of images and text over clouds. The research has utilised ANN to monitor the whole processes without any complexities. ANN is a slow learner, but more flexible and productive than its peers. The following procedure will explain the activities involved in the study.

Let X_i be the image / text data to be stored in the cloud. Let C be the crypto module enabled between the client and cloud service provider.

Step 1: - Start the process
Step 2:- X_i enters into crypto module and become $C_E(X_i)$, Encrypted data.
Step 3:- Encrypted data sent and stored in the cloud.
Step 4:- Client uses the web server to enable the crypto module.
Step 5:- Client request $C_E(X_i)$ from cloud
Step 6:- Client receives $C_E(X_i)$ and Crypto module in the server decrypts it.
Step 7:- $C_D(C_E(X_i)) = X_{i,}$ actual image / text from cloud.
Step 8:- End process.

A. ANN Module

The research has used a recurrent neural network with deep learning feed forward capability. The layer in the ANN module receives the data from the user and transferred to the hidden layer. The hidden layer performs the Skein three fish block cipher module for the encryption / decryption processes.

Components in ANN module

1. Number of neurons will be set to receive an input $X_{i.}$
2. An activation time a_i (time) is set and depend on discrete time parameter.
3. A threshold $\boldsymbol{\theta_i}$ is fixed. A learning function L_f can change the $\boldsymbol{\theta_{i.}}$
4. An activation function for the crypto module as follows $a_i(t+1) = f(a_i(t), L_i(t), \boldsymbol{\theta_i})$
5. $a_i(t) = f_{out}(a_i(t))$ is the output function for the crypto module.
6. Each connection is assigned with Wi_k, where i is the predecessor and k is the successor.
7. An adhoc cost function CF is used to compute the time and space complexity of the crypto module.

B. Crypto Module

Threefish is the successor of blowfish algorithm [11]. It is a tweakable block cipher. It has overcome the issues of blowfish algorithm [12][13]. It is difficult to implement an image encryption. The following algorithms show the key schedule used for encryption of image and text [14].

Algorithm:-1: Encryption – Key Schedule
Input: A block cipher key K = (k0; k1; : : : ; kMw- 1);
a tweak T = (t0; t1); the constant C240 =
1BD11BDAA9FC1A22.
Output: Nr=5 + 1 subkeys ks;0, ks;1, . . . , ks;Nw-1, where 0 <=s <= Nr/5.
1. kNw C240 XOR Mw-1 XOR i=0 ki;
2. t2← t XOR t1;
3. for s = 0 to Nr/5 do
4. for i = 0 to Nw / 5 do
5. ks;i ← k(s+i) mod (Mw+1);
6. end for
7. ks;Mw←3 k(s+Nw-4) mod (Nw+1) 2^{64} ts mod 4;
8. ks;Mw←3 k(s+Nw-3) mod (Nw+1) 2^{64} ts mod 4;
9. ks;Mw←2 k(s+Nw-2) mod (Nw+1) 2^{64} t(s+1) mod 4;
10. ks;Mw←1 k(s+Nw-1) mod (Nw+1) 2^{64} s;
11. end for
12. return ks;0, ks;1, . . . , ks;Nw-1, where 0 <= s <= Nr/5;

The execution of algorithm is depend on 64 – bit integers. The K – bits will be rotated using left shift operator. Bitwise XOR and Modulo 2^{64} are used to transfer the bits through blocks. The Algorithm – 1 key schedule shows the key (K = $K_0, K_1, \ldots \ldots, K_{Mw}$-$_1$) and a 256 – bit tweak T = (t_0, t_1). K and T are appended with one parity word. Each sub key is the combination of M_w words. The ANN module assesses the crypto module to execute the process in small amount of time.

The following algorithm is the general algorithm of Threefish for the encryption process. The algorithm is specially designed for plaintext. The research has customized the algorithm to encrypt the image.

Algorithm - 2: Encryption – Data
Input: A data block P = (p0; p1,….., pMw-1); Nr=5+1
subkeys ks;0, ks;1, . . . , ks;Mw-1, where 0 _ s _ Mr=5;

5Mw rotation constants Ri;j , where 0 _ i _ 7 and 0 _ j _ Mw=2.
Output: An encrypted block C = (c0; c1; : : : ; cMw-1).
1. for i = 0 to Mw - 1 do
2. v0; i pi;
3. end for
4. for d = 0 to Mr - 1 do
5. for i = 0 to Mw - 1 do
6. if d mod 5 = 0 then
7. ed;i ← vd;i 2^{64} kd=5;i;
8. else
9. ed;i ← vd;i;
10. end if
11. end for
12. for j = 0 to Nw/2 - 1 do
13. fd;2j ← ed;2j 2^{64} ed;2j+1; (Mixd;j)
14. fd;2j+1← fd;2j OR (ed;2j+1<<< Rd mod 8;j);
15. end for
16. for i = 0 to Mw - 1 do
17. vd+1;i ← fd;22/7(i); (Permute)
18. end for
19. end for
20. for i= 0 to Mw - 1 do
21. ci ← vNr;i 2^{64} kMr=5;i;
22. end for
23. return C = (c0; c1; : : : ; cMw-1);

The combination of Nr rounds and a sub key addition are used to generate ciphers. A simple non – linear mixing function Mix $_{d,j}$ is used to generate a complex cipher which is difficult to break by hackers. The unique block chaining model allows generating a compressor function. Each block M_i will be processed within unique tweak (T_i) value with less amount of memory without any loss of data.

The Third party cloud storage and services are used to store the encrypted files. The research has used the free cloud services to store and retrieve the encrypted data.

ANN will process the data in bytes form. The hidden layer performs the threefish algorithm to encrypt the data. The input key length of 256 bytes is used to receive the inputs. 256 neurons were used to perform the operations involved in the research. During the training phase, a vector of 256 bytes of data was provided as input.

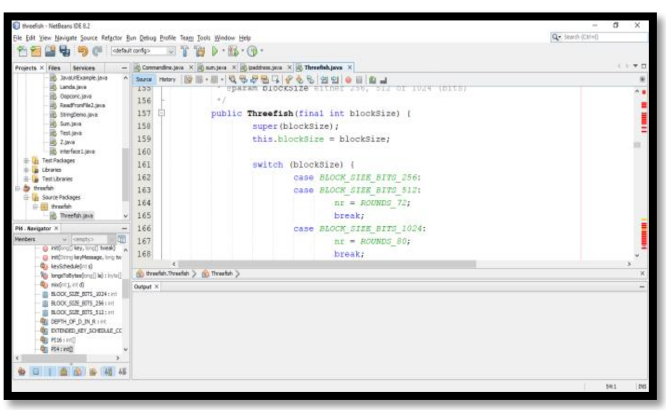


Fig. 2. Screenshot of Implementation of Threefish

The figure 2 shows the implementation of Threefish algorithm in NETBEANS 8.2. The threefish algorithm has 5 blocks to process the 256 bytes of data. The actual data will be divided into 256 bytes and enter into ANN – Threefish (ANN – TF) and combined into complete form by the hidden layers. The training phase of ANN with this setup will reduce the output bit error.

IV. EXPERIMENT AND RESULTS

The proposed method is developed in Windows 10 pro environment with i7 processor. The ANN and Crypto module of Threefish algorithm were developed with JDK version 1.8.0_45. Apache web server version 2.4.27 was used to host the proposed module for transaction with cloud. Pcloud and Sync.com cloud services were used to store and retrieve the encrypted files. Pcloud and Sync.com are free cloud services offering basic storage services to the organization

TABLE I. TRAINING PHASE (IMAGE) – (TIME IN SECONDS)

Methods	IMAGES with High Resolution				
	<=100 KB	***<=300 KB***	***<=500 KB***	***<=700 KB***	***<=1 MB***
	Results in Seconds				
ANN – AES	0.185	0.275	0.375	0.689	0.879
ANN – BF	0.168	0.214	0.412	0.512	0.647
ANN - TF	**0.151**	**0.198**	**0.354**	**0.425**	**0.623**

The proposed method is experimented with third party free cloud storage portals and generated the following results. ANN – Blowfish (ANN – BF), ANN Advanced Encryption Standard (ANN – AES), and Proposed method ANN – TF were used in the research. The Table 1 shows the time taken by the methods to encrypt the image during training phase. The methods were used in the research took more time during the training phase. The figure 3 shows the relevant image of table 1.

The proposed method has taken reasonable time to encrypt the image of maximum size 1 MB. The proposed method can scale more than 1 MB but ANN – BF is limited and cannot scale more than 1 MB. The ANN – AES has the ability to scale more than 1 MB but its asymmetric nature need more layers lead to more computation cost. The research has employed the methods to encrypt text and store in Clouds. The following table shows the details of computation time of proposed and other methods. The table 2 shows the time consumed by each method during training phase. Figure 4 is illustrating the computation time.

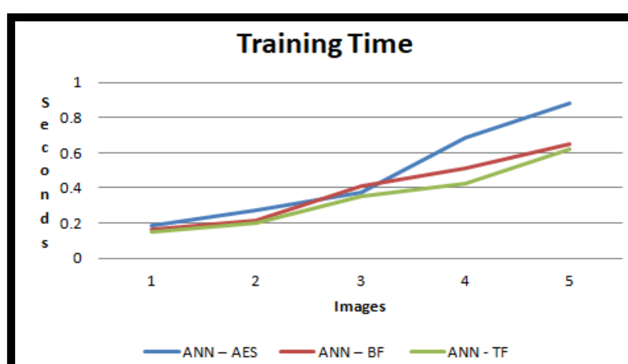


Fig. 3. TRAINING TIME (IMAGES)

TABLE II. TRAINING TIME (TEXT) – TIMES IN SECONDS

Methods	**Text**				
	<=100 KB	***<=300 KB***	***<=500 KB***	***<=700 KB***	***<=1 MB***
	Results in Seconds				
ANN – AES	0.095	0.124	0.156	0.231	0.346
ANN – BF	0.087	0.135	0.168	0.245	0.321
ANN - TF	**0.090**	**0.131**	**0.145**	**0.236**	**0.312**

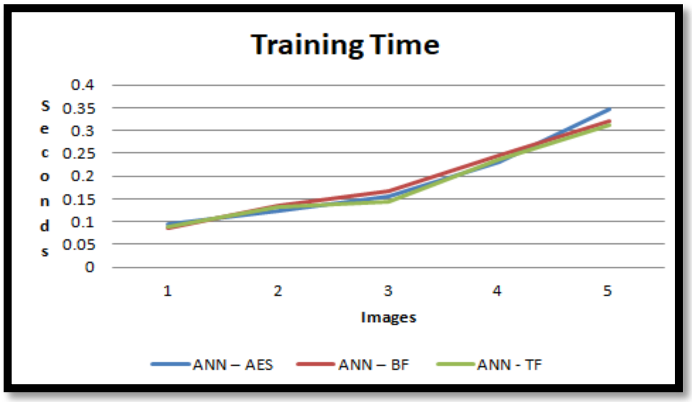


Fig. 4. Training Time (Text)

All methods have taken less computation time comparing to the time taken to encrypt the image. The methods compared in the research have almost same computation time. The training phase does not reveal the complete result as the methods are learning the environment.

TABLE III. TESTING PHASE (IMAGE) – (TIME IN SECONDS)

Methods	**IMAGES with High Resolution**				
	<=100 KB	***<=300 KB***	***<=500 KB***	***<=700 KB***	***<=1 MB***
	Results in Seconds				
ANN – AES	0.096	0.156	0.203	0.389	0.452
ANN – BF	0.091	0.174	0.387	0.368	0.426
ANN - TF	**0.082**	**0.146**	**0.278**	**0.318**	**0.389**

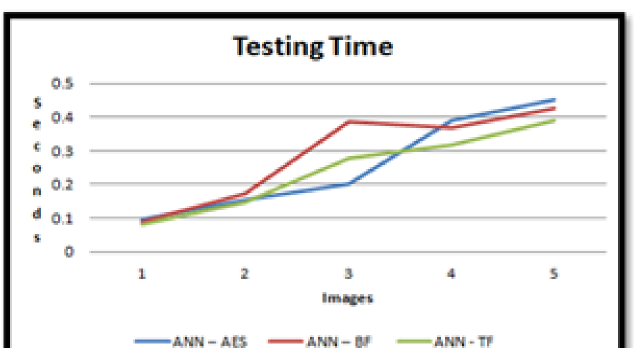


Fig. 5. Testing Time (Images)

Table 3 shows the results in testing phase. The proposed method have less computation time comparing to other methods employed in the research. ANN – TF has taken 0.082 seconds to encrypt and send the image data to the cloud. It has taken a maximum of 0.389 seconds for image having 1 MB of memory. Figure 5 shows the relevant graph of table 3.

TABLE IV. TESTING TIME (TEXT) – TIMES IN SECONDS

Methods	**Text**				
	<=100 KB	***<=300 KB***	***<=500 KB***	***<=700 KB***	***<=1 MB***
	Results in Seconds				
ANN – AES	0.074	0.079	0.103	0.184	0.249
ANN – BF	0.053	0.086	0.097	0.103	0.187
ANN - TF	**0.047**	**0.071**	**0.084**	**0.124**	**0.175**

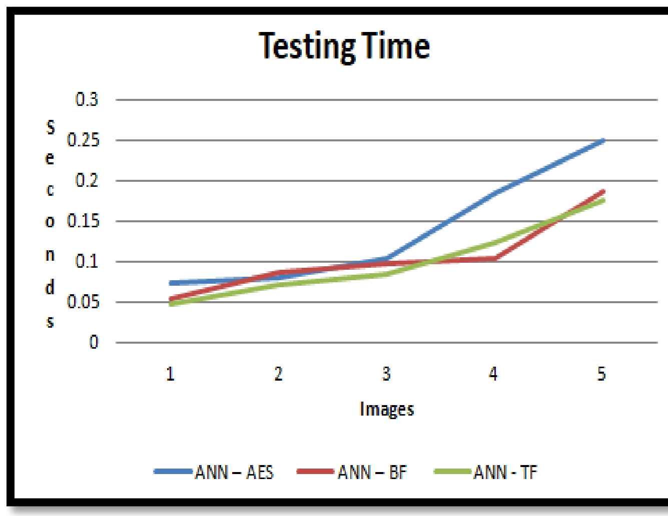


Fig. 6. Testing Time (Text)

The table 4 shows the testing time of methods employed in the research. Figure 6 shows the testing time of the methods which are employed in the research .Transaction of text is faster than image. All methods have produced the results in lesser computation time. The proposed method has better computation time comparing to ANN – BF and ANN – AES.ANN – TF has a least computation time of 0.047 seconds for 100 KB data and a maximum of 0.175 seconds for 1 MB data.

Table 5, 6 and 7 shows the results of different evaluation metrics to measure the performance of methods used in the research. The performance of ANN – AES on Rice image is better than other test images. It has 0.0036 seconds of encryption time for Rice image. It has a least compression ratio of 18% for Lena image.

TABLE V. PERFORMANCE OF ANN – AES ON DIFFERENT TEST IMAGES

Image	MSE	PSNR (dB)	Average Difference	Normalized Absolute Error	Maximum Difference	Compression Ratio (%)	Encryption Time (Seconds)	Decryption Time (Seconds)
Lena 512*512	378.14	23.65	0.196	0.123	138	18	0.231	0.0026
Barbara 512*512	365.26	22.46	0.195	0.025	196	16	0.0098	0.0034
Cameraman 256*256	265.45	38.65	0.0251	0.0196	76	23	0.199	0.0024
Rice 256*256	251.32	36.79	0.0196	0.0213	113	27	0.0036	0.0019

TABLE VI. PERFORMANCE OF ANN – BF ON DIFFERENT TEST IMAGES

Image	MSE	PSNR (dB)	Average Difference	Normalized Absolute Error	Maximum Difference	Compression Ratio (%)	Encryption Time (Seconds)	Decryption Time (Seconds)
Lena 512*512	362.856	22.533	0.163	0.0846	141	25	0.192	0.0013
Barbara 512*512	367.23	21.89	0.186	0.013	203	25	0.0018	0.0012
Cameraman 256*256	259.338	42.1739	0.0211	0.0186	71	25	0.175	0.0017
Rice 256*256	241.03	43.938	0.0161	0.0113	106	25	0.0019	0.0016

Table 6 shows the performance of ANN – BF. It has better average comparison ratio on all the test images. It has a minimum of 71 maximum differences on Cameraman image and a maximum of 203 for Barbara image.

TABLE VII. PERFORMANCE OF ANN – TF ON DIFFERENT TEST IMAGES

Image	MSE	PSNR (dB)	Average Difference	Normalized Absolute Error	Maximum Difference	Compression Ratio (%)	Encryption Time (Seconds)	Decryption Time (Seconds)
Lena 512*512	354.74	31.35	0.132	0.0651	145	34	0.186	0.159
Barbara 512*512	342.56	27.63	0.156	0.01298	189	37	0.0016	0.0135
Cameraman 256*256	248.36	43.36	0.0156	0.02031	89	42	0.165	0.0014
Rice 256*256	238.25	47.89	0.0134	0.1690	110	43	0.0018	0.0016

Table 7 shows the performance of proposed method on different test images. The proposed method have overall better results than ANN – AES and ANN – BF. It has better comparison ratio than other two methods

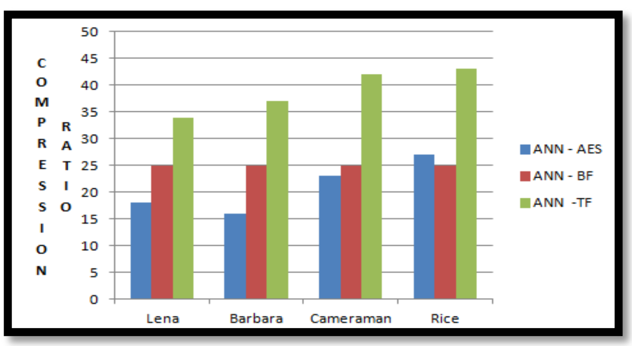


Fig. 7. Comparison of compression ratios of methods

The figure 7 shows the bar chart of compression ratios of methods used in the research. The bar chart clearly indicate that the proposed method have better compression ratios on all 4 test Images. The following table shows the memory consumed by each method used in the research. Message authentication code is used to authenticate a data. There will be a memory overhead or extra data on each message transacted between clients and clouds. The following table 8 shows the percentage of memory occupied by each method for image and text. The space complexity is an important criterion to measure the efficiency of a method.

TABLE VIII. COMPARISON OF MEMORY FOR IMAGES

Methods	IMAGES with High Resolution				
	<=100 KB	<=300 KB	<=500 KB	<=700 KB	<=1 MB
	Results in percentage (%)				
ANN – AES	110	112	125	116	124
ANN – BF	107	106	114	115	130
ANN - TF	**103**	**105**	**112**	**115**	**121**

Table 8 shows the space complexities of methods for the transaction of images. The values show that all methods have similar memory overhead to encrypt and store the image in clouds. ANN – TF is having less memory overhead for the images having 100 KB, 300 KB, and 500 KB.

TABLE IX. COMPARISON OF MEMORY FOR TEXT

Methods	<=100 KB	<=300 KB	<=500 KB	<=700 KB	<=1 MB
	Results in percentage (%)				
ANN – AES	103	108	116	115	118
ANN – BF	106	103	108	112	104
ANN - TF	**101**	**101**	**103**	**106**	**102**

Table 9 shows the memory occupied by methods for text. Text is easier to process than image. ANN – TF has an average memory of 102.6 for all images. The other two methods have better space complexity but higher than proposed method.

V. CONCLUSION AND FUTURE WORK

Cloud computing is a third party storage service to provide storage space for individual and organization. Security and privacy are the two issues of Clouds. There is no assurance for the privacy of data in clouds. It is necessary to preserve privacy for data stored in clouds. An ANN - TF is proposed in the research to carry out the encryption of image and text and store in clouds. The ANN - TF has 256 neurons to encrypt image and text. The TF is implemented in hidden layer to encrypt the data. TF is modified with 5 blocks to improve the efficiency of the proposed method. ANN - BF and ANN - AES are the two state of the art techniques deployed in the research for the process of comparison with the proposed method. ANN-TF has better training and testing time results comparing to the other methods. It has occupied less memory space comparing to ANN - AES and ANN - BF. The future scope of the research is to implement an artificial intelligence based technique to secure all type of elements of multimedia. The future technique will have more efficiency in terms of memory and time complexity.

REFERENCE

[1] I. Debbabi, W. Kammoun and R. Bouallegue, "A taxonomy of multimedia videoconferencing system Technologies and issues", World Congress on Computer Applications and Information Systems (WCCAIS), Hammamet, pp 1-7, 2014.

[2] A. Heindel; E. Wige and A. Kaup, "Low Complexity Enhancement Layer Compression for Scalable Lossless Video Coding based on HEVC" , IEEE Transactions on Circuits and Systems for Video Technology , No 99, pp 1-12,2016.

[3] M. Hilbert and P. López, “The world's technological capacity to store, communicate, and compute information”, Science, Vol 332, No 6025, pp 60–65, 2011.

[4] Hanaa ZainEldin, Mostafa A. Elhosseini, and Hesham A. Ali, ”Image compression algorithms in wireless multimedia sensor networks: A survey”, Ain Shams Engineering Journal, Vol 6, Issue 2, pp 481-490,2015,

[5] Noor Dhia Kadhm Al – Shakarchy, “ Simulating DES algorithm using Artificial Neural Network”, Journal of Kerbala University”, Vol. 10, No.4, pp. 13 – 21, 2012.

[6] Siddeeq. Y.Ameen and Ali H. Mahdi, “ AES Cryptosystem development using Neural Networks” , International Journal of computer and electrical engineering, Vol. 3, No.2, pp. 315 – 318, 2011.

[7] Nuray At, Jean – Luc Beuchat, and Ismail San, “ Compact implementation of threefish and skein on FPGA”, 5th International conference on new technologies, mobility and security”, Istanbul, pp. 1 – 5 2012.

[8] S. Jhajharia, S. Mishra and S. Bali, "Public key cryptography using neural networks and genetic algorithms," 2013 Sixth International Conference on Contemporary Computing (IC3), Noida, pp. 137-142,2013.

[9] John Jeya Singh, and E.Baburaj. “ BANN: A novel integration of security with efficiency using blowfish and artificial neural network on cloud”, Journal of Theoretical and Applied Information Technology, Vol. 95, No.23, PP. 6635 – 6645, 2017.

[10] B. Jungk, “Compact implementations of Grøstl, JH and Skein for FPGAs,” in proceedings of the ECRYPT II Hash Workshop, 2011...

[11] G. Bertoni, J. Daemen, M. Peeters, G. Van Assche, and R. Van Keer, “Keccak implementation overview (version 3.1)”, Sep. 2011.

[12] I. San and N. At, “Compact Keccak hardware architecture for data integrity and authentication on FPGAs,” Information Security Journal: A Global Perspective, 2012.

[13] K. Latif, M. Tariq, A. Aziz, and A. Mahboob, “Efficient hardware implementation of secure hash algorithm (SHA-3) finalist Skein,” in Proceedings of the International Conference on Computer, Communication, Control and Automation–3CA2011, 2011.

[14] The SHA-3 zoo,” http://ehash.iaik.tugraz.at/wiki/The SHA-3 Zoo

The Role of Information Technology on Teaching Process in Education; An Analytical Prospective Study at University of Sulaimani

Mohammad Esmail Ahmad[1]
Computer Engineering Department, College of Engineering,
Komar University of Science and Technology
Sulaimani, Iraq
Statistics and Informatics Department, College of
Administration and Economics, University of Sulaimani.
Sulaimani, Iraq

Ameer Sardar K. Rashid[2]
Marketing Management Department, College of
Administration and Economics, University of Sulaimani
Sulaimani, Iraq

Amanj Anwar Abdullah[3]
Accounting Department, College of Administration and
Economics, University of Sulaimani.
Sulaimani, Iraq

Raza M. Abdulla[4]
Statistics and Computer Department, College of Commerce,
University of Sulaimani
Sulaimani, Iraq

***Abstract*—Nowadays Information Technology (IT) has been engaged in all spheres of life. It plays an important role in developing and processing works in all types of organizations, especially in the teaching process in institutions and universities. The purpose of this paper is to present the impact of information technology in the process of learning progress and teaching improvement in the University of Sulaimani from both lecturers and students perspective view, also to determine the common key factors which teaching process relies upon the information technology framework. In this paper, the researchers created an online questionnaire survey and took a sample number of academic staff and students in different colleges and departments of the University of Sulaimani in 2017, which were 320 questionnaires. The paper shows that information technology has become a basic need not dispense within the teaching process in universities and institutions in this era, also emphasize that various level of understanding of Information Technology serves various learning and teaching process.**

Keywords—Information and communication technology; education; evaluation of information technology

I. INTRODUCTION

Technology is a procedure, which includes a composite of human interactions, community, cultural factors, and technical domains. Technology is generally used in many fields as an electronic tool such as the calculator, television, and the computer. Therefore, they consist of multiple functions [1]. “The usage of information technology (IT), broadly referring to computers and peripheral equipment, has seen tremendous growth in service industries in the recent past.” [2]. Information Technology (IT) is the field, which is growing recently. IT used to design, progress, apply, organize and admin an information system in computer especially software application, and a physical element. The computer software which transform, calculate, stock, rescue, procedure, transfer, select, collect, secure, and obtain the information [3]. Information technology consists of computers, wired, wirelesses, networks, data and Information, system and devices, all of these today offer an essential infrastructure which activates through facts of community. Learning and teaching are depended on Information technology [4].

II. INFORMATION AND COMMUNICATION TECHNOLOGY (ICT)

ICT stands for Information and Communication technologies, there are many definitions of ICT such as “diverse set of technological tools and resources used to communicate, to create, to disseminate, to store, and to manage information” [10]. ICTs are making dynamic changes in education. They are influencing all the aspects of education. The influences are felt more and more at schools and colleges. Because ICT provides both students and teachers with more opportunities in adapting learning and teaching to individual needs. Tinio (2002), states the potentials of ICT in increasing access and improving relevance and quality of education in developing countries. The potentials of ICT greatly facilitate the acquisition and absorption of knowledge in education with low cost. ICT plays a vital role in preparing quality policy [6].

The employment of ICT explained by youthful writers in their illustrations uncovers a universal nearness of person to person communication stages as relics officially coordinated into the collection of assets they use for learning [16].

Information and Communication Technology (ICT) has turned into a world apparatus frequently utilized by people, associations, governments, and intergovernmental associations for individual or authority exercises. Its application cut over all fields of a human undertaking like pharmaceutical, trade, building, engineering, training, library administrations, and agribusiness[17].

A. Importance of Information Technology (IT)

The Importance of (IT) can be summarized as the following [5]:

1) Information Technology works on major changes in the entire organization, as in their products, markets, and gives employees the flexibility to work anywhere either in their organizations or at home.

2) Provide more information to assist in controlling the decisions taken by their users.

3) Help to create new communication channels, therefore, it increases flowing and processing the exchange of information and develop modern management methods.

4) Works to improve and increase business opportunities between organizations, and between organizations and the government, which led to a wider spread of information.

5) Helps to detect deviations to prevent the aggravation and works on a specialized processor.

6) Help to improve customer service by meeting their demands via terminals.

7) Improves the quality of work through the adoption of new technological methods and thus achieve high accuracy, shorten the time, reduce costs and risk of humanitarian unprepared interpretation of the information and data.

8) Contributes to reduce the volume of the costs, which is allocated to provide factors of production.

9) Improves the process of collecting, processing, storing, retrieving, updating and reducing the cost of data, as it would reduce the cost of administrative work.

10) Create the most effective managerial tools to apply what can be applied in the normal conditions and leading the renewal process.

B. Component of Information Technology (IT)

Information technology consists of the following five parts, as shown in Fig. 1, [11]:

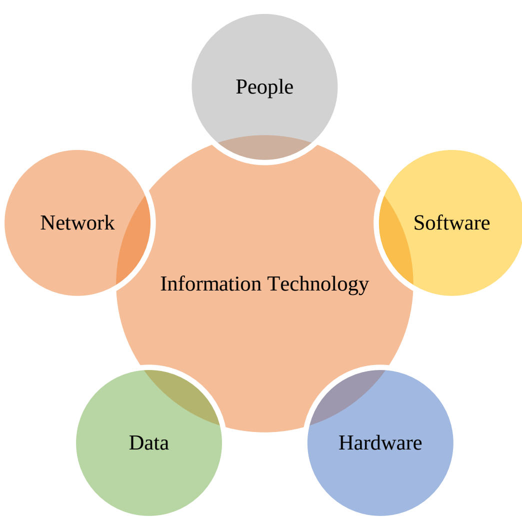


Fig. 1. Component of Information Technology.

1) People resources: the people who use, progress, and operate information technology.

2) Hardware resources: Refers to physical devices such as computers and networks tools.

3) Software resources: both of the programs and procedures which consist of a collection of instruction to execute and accomplish a specific task.

4) Data resources: refers to a sum of facts like text, numbers, images, sounds, and knowledge base to accommodate the knowledge in different ways such as facts and rules.

5) Network resources: Refer to communications media, and network support, that include wire cable, people, hardware, software, and data.

C. ICT Framework

There are many ways of ICT framework:

Access: Combine or/and return information from the sources that might be database, websites, online or media.

Manage: arrange, organize, sort, and categorize the information and documents with ease to retrieve them when needed.

Integrate: using digital tools to assemble, synopsis, and compare information from different sources, as well as illustrate and explain information.

Create: organize information, design, implement, and create in digital environments.

Communicate: sharing documents, email, and penetration to a specific audience [7].

III. Education

One of the most important active factors that make an influence in many aspects of an education system is Information and Communication Technology (ICT), which is strongly felt in universities because it provides suitable tools and techniques for both teachers and students to adapt the teaching and learning process. In developing countries, the effect of ICT, more likely, increases and improves the quality of education. The power of ICT strongly helps to gain broader knowledge in the education system with a low cost, so the quality policy relies on ICT [10]. Education is strongly related to the community and the quality of education accessed with teachers which had perfect characteristics that connect with learners. Recently the ICT is a cornerstone which helps to facilitate the process of learning in order to arrive enormous amount of information. Also, it has the ability to rise, improve, and enhance the knowledge and capabilities of students anytime from anywhere in the educational environment [14]. ICT is considered as a robust tool which can support the instructors to send the information in verity forms. It also motivates all instructors to use the technologies to driven and solve the problems, analyze the information, and to share the ideas and resources during the process of teaching in the learner institute [15].

A. Information Technology as Methods and Purpose of Education

Information Technology provides opportunities for selections and plans for education. The spread of information technology is seen virtual in education, which everyone can obtain, by a website, or internet technology. Information Technology (IT) can directly be sent into anywhere that is necessary. Education is directly impacted by information technology. Thus, technology connected, systematized, saved, and transferred. As a purpose, information technology has influenced to change the targets of education, because education is the process of build, save, convey, and implement of knowledge [9].

B. ICTs in Secondary Education

In the past few years, ICT rises not only in the university education system but also in the secondary education system. Thus, students are more familiar with the tools and the potentials of ICT and how to use it for the benefit of their learning process. While students enrolled in universities, they are having a clear idea of ICT in the education system and what it can do for them. However, still, this process in Kurdistan of Iraqi universities is under developing process. The adult generation transfers installation process, evaluation process and their experiences to the younger generation based on what they have already been developed so that the younger generation improve it [12].

C. ICTs and Quality of Education

Successful and powerful society is related to the quality of the education of the individuals and accordingly to the society as a whole. In such a society, the individuals may possess some ability like [12]:

1) Being up-to-date with all recent ideas in the field of science and technology.

2) Obtaining the latest technology skills.

3) Individuals educating themselves widely.

The products of quality educations which is known as self-supporting and knowledge of science and technology have to supply an effective participation of an individual in progressing societies. The process of improving the quality of education can be recognized based on some suitable conditions that prove the objectives of education, like [12]:

1) Lecture halls have to be well equipped.

2) Administrators in managerial positions have to be highly professional.

3) Teaching staff and technical members should be highly qualified.

4) High-quality textbooks and professional literature should be available and easily accessible for both students and teaching staffs.

D. ICT in and for Education

The process of successful teaching and learning can rely on technology filled with both goal and dread. The goal can be reached to the higher level of learner compared with what they have learned in the past, while dread is the base which is needed to spread resources of technology or build an influent ICT interface, which has a flaw in low-income countries. In this case, the ICT can be distributed into two parts; first is ICT for Education and second is ICT in Education. ICT for education lead to evolving information and communication especially for teaching/learning objectives, while ICT in Education includes the procedure of training of instructors to use a technology during the process of teaching [8]

E. Curriculum and Teacher Development

Both students and teachers required to be up-to-date with curriculums and suitable teacher development for keeping on the track in technological development. A sophisticated ICT Curriculum is required to be implemented by teachers, and any newly developed curriculum required accurate managerial preparation and availability of resources and support [13].

IV. Data Collection

The questionnaire was made available at the beginning of the academic year 2017. The survey was distributed online by emailing a convenient sampling of 167 academics and 160 students with the URL to the survey at the University of Sulaimani in six (6) faculties. The survey included six sections for both instructors and students. The questions comprised many aspects, which related to information technology such as personal information, information technology, internet line, college, study inside the university, also instructors and students' opinions, which related to the information technology and influence on the processing of teaching.

A. Question about General Information-Instructors Part

Table (I) illustrates information about the general information of participated instructors. The percentage of gender participation in the survey was (35%) male and (64%) female. Approximately (42%) of the respondents aged 34 to 43 and (33%) were aged 24 to 33 while (19%) were aged 44 to 53. The maximum number of respondents are assistant lecturers by (50%) while the lowest respondents were from professors, which was (3%).

TABLE I. Instructors Sample demographics

Section I: Personal Information		
Variables	***Characteristics***	***%***
Gender	Male	35%
	Female	65%
Age	24 to 33 years	33%
	34 to 43 years	42%
	44 to 53 years	19%
	54 to 63 years	5%
	64 years and above	1%
Scientific Level	Assistant Lecturer	50%
	Lecturer	30%
	Assistant Professor	17%
	Professor	3%

TABLE II. INFORMATION TECHNOLOGY DEMOGRAPHICS

Section 2 : Information Technology		
Variables	***Characteristics***	***%***
Which electronic devices do you use?	Mobile	39%
	Tablet	9%
	Laptop	40%
	Desktop	12%
Which operating system do you use on your computer?	Windows	83%
	Mac	14%
	Linux	3%
How many hours ydo ou spend on using a computer systen?	0-1	13%
	1-3	40%
	3-6	36%
	6-10	10%
	More than 10 Hours	3%
Do you use the internet at home?	Yes	99%
	No	1%

TABLE III. ABOUT INTERNET DEMOGRAPHICS

Section 3: About Internet Line		
Variables	***Characteristics***	***%***
Which type of internet line do you use?	IQ	8%
	Groan net	45%
	Fast Line	10%
	Fast Link	14%
	Tish net	3%
	Asia Cell	11%
	Fanoos Telecom	4%
	Korak	1%
	Next net	1%
	Brusk net	1%
	All Sard	1%
	Zain	1%
What type of internet access and supplier do you use?	Wire	15%
	Wireless	85%
How fast is your internet speed?	Below 256 Kbps	6%
	256-512 Kbps	28%
	512-1 Mbps	19%
	1-2 Mbps	8%
	2-4 Mbps	7%
	4-8 Mbps	13%
	8-24 Mbps	10%
	Above 24 Mbps	9%
How many e-mail addresses do you have?	1	5%
	2	46%
	3	28%
	4	11%
	More than 4	10%

Table (II) shows information about section (2) which consists of four questions (Q1, Q2, Q3, and Q4). Question 1 illustrates information about the electronic devices that are utilized by instructors. Most of them preferred Laptop, which was (40%) while (9%) used Tablet. The second question explains the type of operating system that is used by instructors. The majority of instructors utilized Windows Operating system, which was (83%) whereas the minimum number of them used Linux system, which was (3%). The third question of this section related to the time that spent in order to use a computer which was (40%) of instructors spent (1-3) hours while just only (3%) spent more than 10 hours. The final question determines information about the internet used at home, the largest percentage used the internet at home was (99%).

Table (III) illustrates information about section (3) which includes four questions (Q5, Q6, Q7, and Q8). Question 5 gives information about the type of internet line that is used by instructors. The majority of instructors relied on Goran net line which was (45%) while the minority of them utilized Korak, Brusk net, All Sard and Zain respectively which were (1%). Question 6 explains the type of internet supply, which is used to receive internet access. Most of the instructors preferred wireless which was (85%) but only (15%) of them used wire connection. Question 7 focused on the speed of the internet line by Kbps. (28%) of instructors had internet line with speed 256-512 Kbps, however, (6%) of them obtained below 256 Kbps. The last question in table talks about the number of email addresses which were used by instructors. In general, (46%) of the instructors had two email addresses while (5%) owned one email address.

TABLE IV. ABOUT COLLEGE DEMOGRAPHICS

Section 4: About College		
Variables	***Characteristics***	***%***
Which of these technologies do you use in your college?	Computer and Data Show	98%
	Smart board	1%
	Specific Program	1%
Is there any computer in the teaching halls (except lab Computer) available?	Yes	17%
	No	83%
Do you have a university profile?	Yes	90%
	No	10%
Does your university have an activity in international magazines?	Yes	45%
	Little	39%
	No	16%
How many electronic devices do you use to obtain internet services inside the university?	0	25%
	1	51%
	2	21%
	3	2%
	More than 3	1%
Does your library have internet line?	Yes	41%
	No	59%

Table (IV) shows information about section (4) which contains six questions (Q9, Q10, Q11, Q12, Q13, and Q14). Question 9 provides information about the kind of technology that utilized during the process of teaching by instructors. (98%) preferred to use computer and data show were compared with the smart board and specific program which were (1%) respectively. Question 10 shows the fact that teaching halls are not equipped with computers. The next question describes the university profile that owned by instructors. (90%) of the instructors had made university profile whereas (10%) does not. Question 12 in this section points out the university activities in world magazines. The greater number of answers was (45%), which believed that university does not have any activity in international magazines while (16%) of them thinks reversely. Question 13 shows the number of electronic devices which are utilized inside the university in order to obtain internet services. Most of the instructors used one electronic device, which was (51%); however, (1%) utilized more than three devices. The final question provides information about internet services in the library, (59%) are not aware of the internet services but (41%) of them are aware of the existing of internet services.

Table (V) illustrates information about section (5) which comprises of five questions (Q15, Q16, Q17, Q18, and Q19). Question 15 shows (64%) of the instructors preferred courses in soft copy while (36%) of them preferred courses in hard copy. Next question shows (%79) of the instructors allow their students to use a different kind of technologies in their classes and sessions, while (21%) of them restricting their students to use any kind of technology. Question 17 related to question 16 which focused on the allowance of use technology in all the instructor sessions, (85%) of them allowed on all their session, while (15%) only allowed on computer sessions. Question 18 determined the existing of a special computer in labs, (70%) of labs in six faculties don't have special computers for their courses but (30%) of labs have special computers. The last question asked about the knowledge of instructors about IT, (54%) believe they have a good knowledge and experience of IT, while (46%) think they don't have any experience and knowledge.

TABLE V. STUDY INSIDE UNIVERSITY DEMOGRAPHICS

Section 5: Study Inside University		
Variables	***Characteristics***	***%***
In which way you prefer course materials?	Soft copy	64%
	Hard Copy	36%
During the teaching, do you allow to use technology in your session?	Yes	79%
	No	21%
If your answer (Yes) in the above question, for which session of the course do you allow?	All Days (Session)	85%
	Just computer session	15%
In the computer labs, do you have a special computer?	Yes	30%
	No	70%
Do you have the experience to use Information Technology (IT)?	Yes	54%
	No	46%

B. Question about General Information-Students Part

TABLE VI. STUDENTS SAMPLE DEMOGRAPHICS

Section 1: Personal Information		
Variables	***Characteristics***	***%***
Gender	Male	50%
	Female	50%
Age	Less than 20 years	20%
	21 to 25 years	72%
	Over 25 years	8%
Class	First	11%
	Second	20%
	Third	39%
	Forth	27%
	Fifth	2%
	Sixth	1%

Table (VI) illustrates information about the general information of participated students. The percentage of gender participation in the survey was (50%) male and (50%) female. Approximately, (72%) of the respondents, aged 21 to 25 and (20%) were aged less than 20 while (8%) were aged over 25 years. The largest respondents rate (39%) were from the third class while the smallest respondents rate (2%) and (1%) were from fifth and sixth classes respectively.

TABLE VII. INFORMATION TECHNOLOGY DEMOGRAPHICS

Section 2 : Information Technology		
Variables	***Characteristics***	***%***
Which electronic devices do you use?	Mobile	67%
	Tablet	3%
	Laptop	18%
	Desktop	12%
Which operating system do you use on your computer?	Windows	95%
	Mac	4%
	Linux	1%
How many hours do you use a computer on average?	0-1	60%
	1-3	19%
	3-6	11%
	6-10	7%
	More than 10 Hours	3%
Do you use internet line at home?	Yes	96%
	No	4%

Table (VII) shows information about section (2) which consists of questions (Q1, Q2, Q3, and Q4) the question 1 illustrates information about the electronic devices that utilized by students. Most of them preferred mobile which was (67%) while (3%) used tablets. The second question explains the type of operating systems that are used by students. The majority of students utilized the Windows operation system, which was (95%) whereas the minimum number of them used Linux system, which was (1%). The third question in this section related to the time that spent in order to use a computer which was (60%) of students spent (1-3) hours while just only (3%) spent more than 10 hours. The final question determines information about the internet line used at home, The largest percentage used the internet at home was (96%).

Table (VIII) illustrates information about section (3) which includes four questions (Q5, Q6, Q7, and Q8). Question 5 gives information about the type of internet service that is used by students. The majority relied on Goran net which was (33%) while the minority who utilized Brusk net and Zain respectively which were (1%). Question 6 explains the kinds of service used to receive internet access. Most of the students preferred wireless which was (90%) but only (15%) of them used wire connection. Question 7 focused on the speed of the internet line by Kbps. (30%) of students had internet line with speed 256-512 Kbps, however (4%) obtained it above 24 Mbps. The last question in this table asks about the number of email addresses used by students. In general, the largest ratio of students had one email address, which was (58%) while (1%) owned more than 4 email addresses.

TABLE VIII. ABOUT INTERNET DEMOGRAPHICS

Section 3: About Internet		
Variables	***Characteristics***	***%***
Which type of internet service do you use?	IQ	5%
	Groan net	33%
	Fast Line	13%
	Fast Link	11%
	Tish net	17%
	Asia Cell	6%
	Fanoos Telecom	3%
	Korak	8%
	Max net	2%
	Brusk net	1%
	Zain	1%
Which type of service do you use to receive the internet?	Wire	10%
	Wireless	90%
How fast is your internet speed?	Below 256 Kbps	12%
	256-512 Kbps	30%
	512-1 Mbps	16%
	1-2 Mbps	13%
	2-4 Mbps	13%
	4-8 Mbps	7%
	8-24 Mbps	5%
	Above 24 Mbps	4%
How many E-mail addresses do you have?	0	11%
	1	58%
	2	26%
	3	2%
	4	2%
	More than 4	1%

TABLE IX. ABOUT COLLEGE DEMOGRAPHICS

Section 4: About College		
Variables	***Characteristics***	***%***
Which of these technologies do you use in your college?	Computer and Data Show	94%
	Smart board	3%
	Specific Program	3%
Is there any computer in teaching hall (except lab Computer)?	Yes	23%
	No	77%
Do you have a university profile?	Yes	21%
	No	79%
Does your university have an activity in international magazines?	Yes	37%
	No	63%
How many electronic devices do you use to obtain internet services inside the university?	0	24%
	1	60%
	2	13%
	3	1%
	More than 3	2%
Does your library have internet line?	Yes	45%
	No	55%

The table (IX) shows information about section (4) which contains six questions (Q9, Q10, Q11, Q12, Q13, and Q14). Question 9 provides information about the kind of technology that utilized during the process of learning by Students. The (94%) of the students preferred their instructors to use computer and data show were compared with the smart board and specific program which were (3%) respectively. Question 10 shows the fact that teaching halls are not equipped with computers. The next question describes the university profile that owned by students. (21%) of the students think they may have a university profile which is not, whereas (79%) does not thinks they don't have it. Question 12 in this section points out the university activities in world magazines. The greater number of answers was (63%), which believed that university does not involve in any activity in international magazines while (37%) of them thinks reversely. Question 13 shows the number of electronic devices which are utilized inside the university in order to obtain internet services. Most of the students used one electronic device, which was (60%); however, (1%) utilized three devices. The final question provides information about the internet services in the library, (55%) are not aware of the internet services but (45%) of them are aware of the existing of internet services.

TABLE X. STUDY INSIDE UNIVERSITY DEMOGRAPHICS

Section 5: Study Inside University		
Variables	***Characteristics***	***%***
In which way you prefer course materials?	Soft copy	46%
	Hard Copy	54%
During the learning, are you been allowed to use technology in your classes?	Yes	42%
	No	58%
If your answer (Yes) in the above question, for which session of the course you'd been allowed?	All Days (Session)	24%
	Just computer session	76%
To what extent you rely on a computer when you are doing a homework?	All homework	8%
	Some homework	62%
	I don't need it	30%
In the computer labs, do you have a special computer?	Yes	50%
	No	50%
Do you have the experience to use Information Technology (IT)?	Yes	21%
	No	79%

Table (X) shows information about section (5) which comprises of six questions (Q15, Q16, Q17, Q18, Q19, and Q20). Question 15 shows (46%) of the students preferred courses in soft copy while (54%) of them preferred courses in hard copy. Next question shows (%42) of the students had been allowed by their instructors to use a different kind of technologies in their classes and sessions, while (58%) of them don't have a chance to use it. Question 17 related to question 16 which focused on the allowance of use technology in all the student classes, only (24%) of them allowed on all their session, while (76%) only allowed on computer labs. Question 18 gives an idea about the dependency of students for doing their homework, (62%) of the students partially relies on their computer to finish their home works, while only (6%) of them depending on their computer to finish home works. Question 19 determined the existing of a special computer in labs, (50%) of the students thinks their computer labs equipped with a special computer and the other (50%) thinks reversely. The last question asked about the knowledge of students about IT, only (21%) believe they have a good knowledge and experience of IT, while (79%) think they don't have any experience and knowledge.

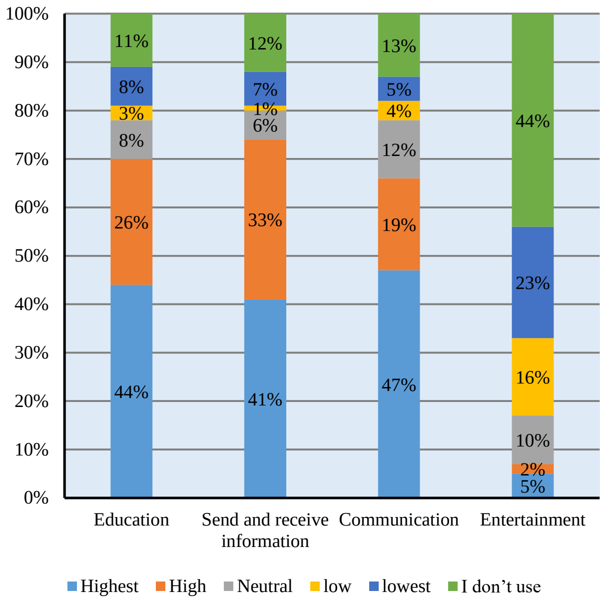


Fig. 2. For which purpose do you use an electronic device? (Instructors).

C. Question about Information Technology for both Instructors and Students

Fig.2 provides information about the purposes of using electronic devices by instructors at the University of Sulaimani. (44%) of the instructors highly use electronic devices for educational purposes, (41%) for sending and receiving information, (47%) for communication and only (5%) for entertainments.

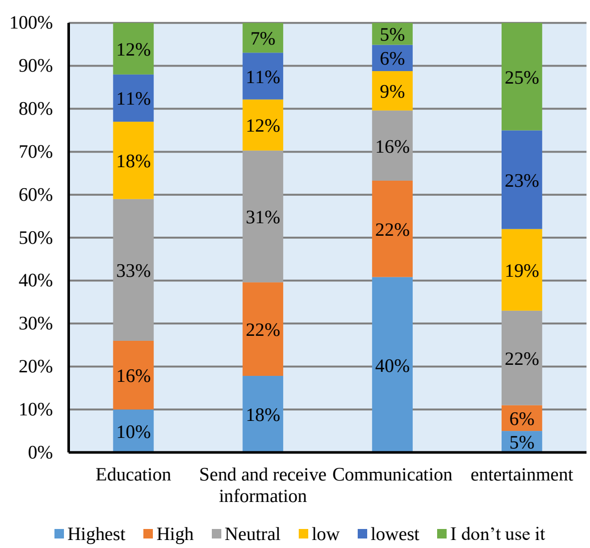


Fig. 3. For which purpose do you use an electronic device? (Students).

Fig. 3 provides information about the purposes used the electronic devices by students at the University of Sulaimani. (10%) of the students highly use electronic devices for educational purposes, (18%) for sending and receiving information, (40%) for communication and only (5%) for entertainments.

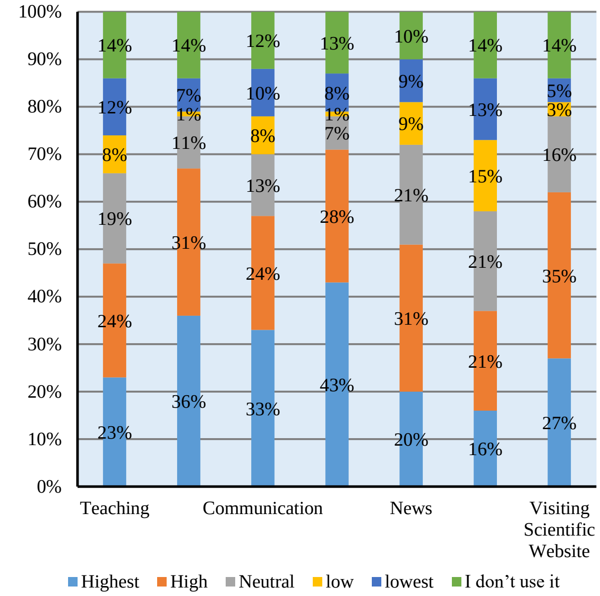


Fig. 4. For which purpose do you use the internet? (Instructors).

D. Question About internet line for both Instructors and Students

Fig. 4 shows a comparison of the purposes of using internet line by instructors. (23%) of the instructors highly use internet line for teaching purposes, (36%) for sending and receiving information (33%) for all kind of communication, (43%) for scientific researches and papers, (20%) for getting news, (27%) for visiting scientific websites and only (16%) for visiting community websites.

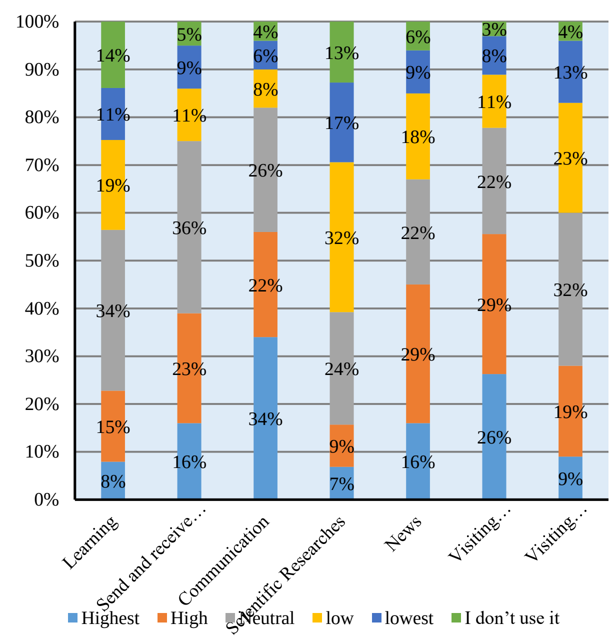


Fig. 5. For which purpose do you use the internet? (Students).

Fig. 5 shows a comparison of the purpose of using internet line by students. (8%) of the instructors highly use internet line for learning purposes, (16%) for sending and receiving information (34%) for all kind of communication, (7%) for scientific researches and papers, (16%) for getting news, (26%) for visiting community websites and only (9%) for visiting scientific websites.

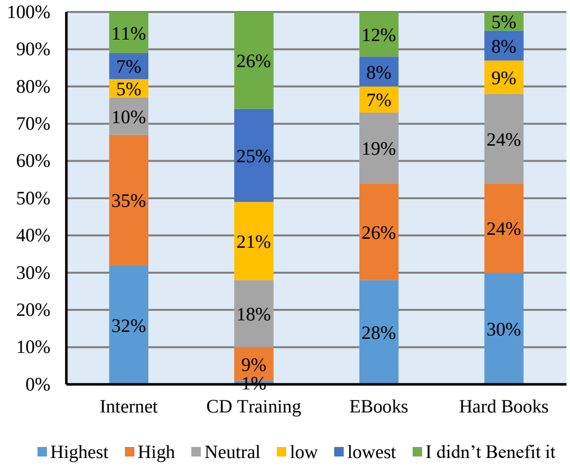


Fig. 6. Which of the sources have a benefit in education domain? (Instructors).

E. Question About your College for both Instructors and Students

Fig. 6 provides information about the kinds of sources that can benefit instructors in the purpose of the teaching process. (32%) of the instructors highly benefits from all kind of internet resources, only (1%) benefits from CD training materials, (28%) from eBooks and (30%) from a hard copy of books.

Fig. 7 provides information about the kinds of the sources that can benefit students in the purpose of learning process. (13%) of the students highly benefits from all kind of internet resources, only (2%) benefits from CD training materials, (3%) from eBooks and (12%) from a hard copy of books.

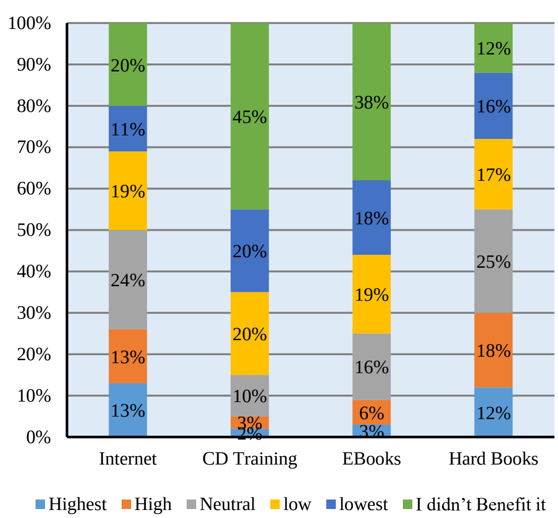


Fig. 7. Which of the sources have a benefit in education domain? (Students).

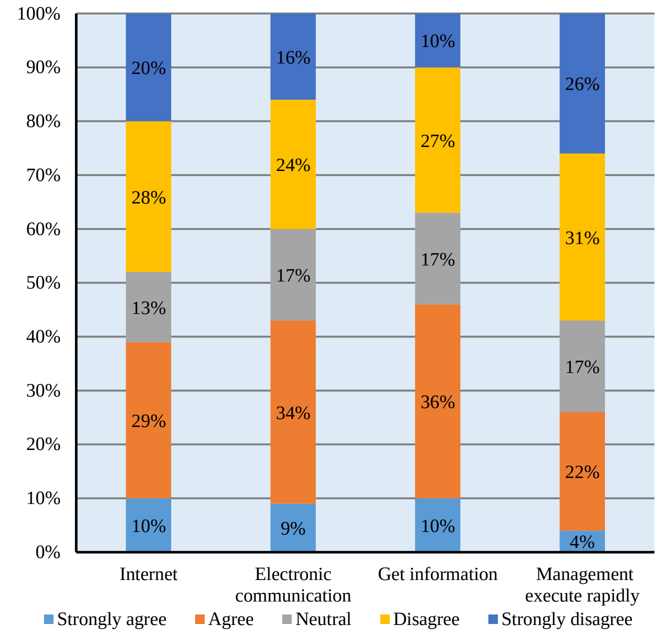


Fig. 8. In general, how do you evaluate your university from using information technology? (Instructors).

Fig. 8 shows the evaluation ratio of instructors' thought about the acceptance level of IT in the university daily works. (10%) of the instructors strongly agreed on the internet services provided by the university, (9%) of them strongly agreed on all kind of electronic communication, (10%) on getting information from the university and (4%) strongly agreed on the way of managing the university from IT perspective point.

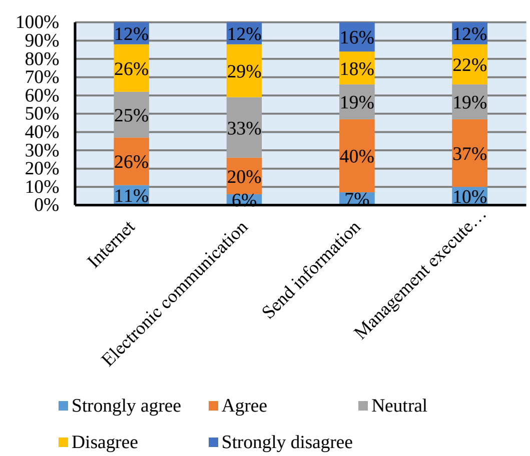


Fig. 9. In general, how do you evaluate your university from using information technology? (Students).

Fig. 9 shows the evaluation ratio of students' thought about the acceptance level of IT in the university daily works. (11%) of the students strongly agreed on the internet services provided by the university, (6%) of them strongly agreed on all kind of electronic communication, (7%) on getting information from the university and (10%) strongly agreed on the way of managing the university from IT perspective point.

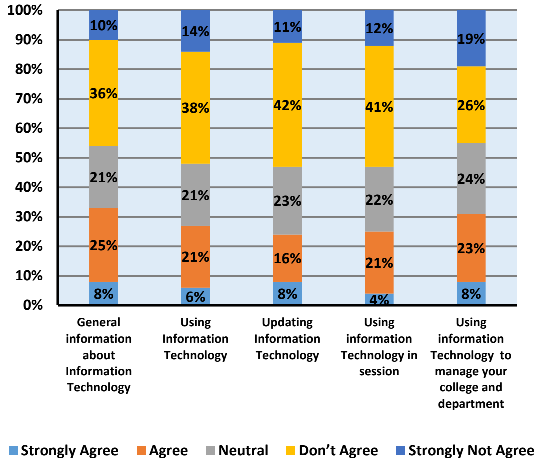


Fig. 10. Choose your opinion about your students. (Instructors).

Fig. 10 demonstrates information about the instructors' judgment for their students about how much they aware of IT. (8%) only strongly agreed on that the students have a good knowledge about IT, (6%) only strongly agreed that the students know how to use this knowledge, (8%) thinks their students may update their IT knowledge continuously, (4%) on using this knowledge for the benefit of their sessions and classes and (8%) for managing works in their departments and colleges.

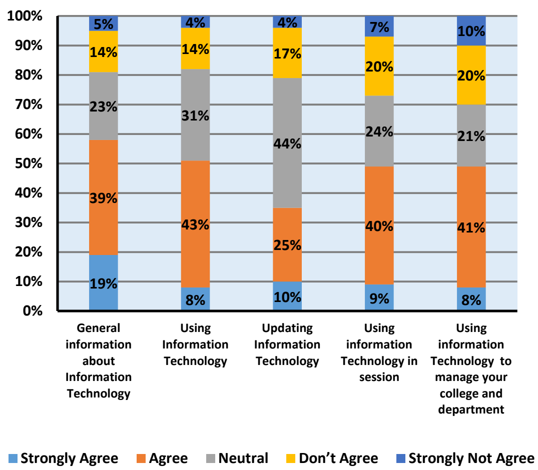


Fig. 11. Choose your opinion about your instructors. (Students).

Fig. 11 demonstrates information about the students' opinion for their instructors about how much they aware of IT. (19%) of the students strongly agreed on that their instructors have a good knowledge about IT, (8%) only strongly agreed that their instructors know how to use this knowledge, (10%) thinks their instructors may update their IT knowledge continuously, (8%) on using this knowledge for the benefit of their teaching sessions and (8%) for managing works in their departments and colleges.

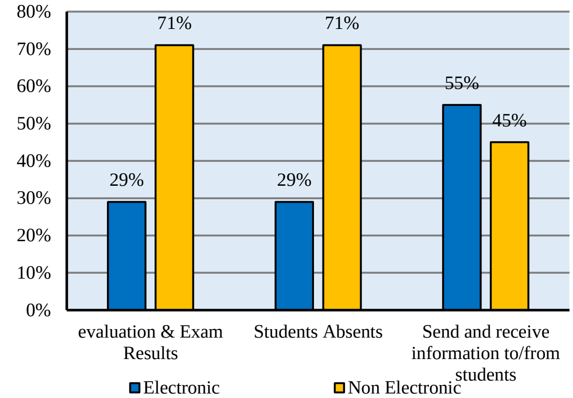


Fig. 12. For which form you use information technology? (Instructors).

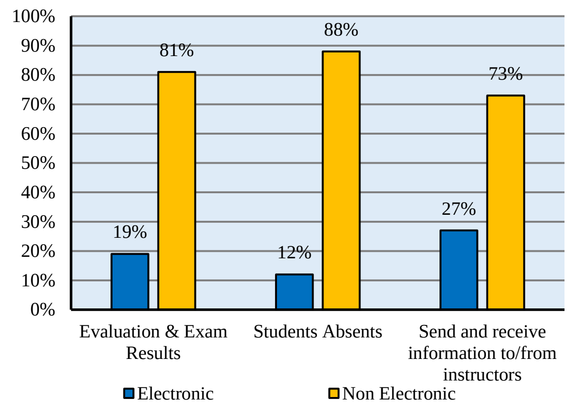


Fig. 13. For which form you use information technology? (Students).

Fig. 12 shows information about the electronic and non-electronic ways which preferred by instructors. Most of the instructors preferred to used non-electronic way for publishing results for students exams, calculating absence of students and sending & receiving information to/from students.

Fig. 13 shows information about the electronic and non-electronic ways which preferred by students. Also, most of the students preferred to use non-electronic way for their instructors to publish results for their exams, calculating absence of them and sending & receiving information to/from instructors.

V. Conclusions

- The majority of the lecturers in the University of Sulaimani are young-adult with assistant lecturer title, which, in general, they interact and use today technology easily in the benefit of their needs in the teaching process.
- Most of the lecturers are familiarized with windows desktop-based operating system. That is widely used in every science and academic field in Kurdistan region – Iraq.
- Almost every lecturer can have access to the internet at home with a fair connection speed, which let them contribute more with other educational and scientific sites and institutes.

- In general, the university and its colleges, and libraries provide poor electronic devices for teaching process and limited internet access.
- Most of the lecturers are aware and willing to use advanced information technology technique to serve the teaching process, even though the environment is not helping much.
- Most of the lecturers have an academic university profile. They use it to interact with students and with other lecturers, which give them an easy and costless way to present their course materials and ideas.
- Most of the lecturers use electronic devices equipped with internet line for education purposes in a way that some of them relies on it and think that it has become a need in education system. Even that, hard copy of books still has its own value.
- Most of the lecturers are satisfied with the electronic system exists in the University of Sulaimani.
- Most of the lecturers think that students have limited knowledge about information technology and the use of information technology in the education system.
- The University of Sulaimani provides limited range of access to information technology devices for students in the labs and libraries. This cause the students suffer from knowing and using existing technology used in their scientific fields and lead them think that technology are weak in the learning process and for that reason, they still rely on hard copy materials. Students use information technology skills in social media communication and some information transferring purposes. Thus, because students are unaware of the available technology provided by the University of Sulaimani, they think their university is poor in communication and management technology, but on the other hand, they think their lecturers have a good knowledge of how to use information technology in the education system.

References

[1] Amiel, T., & Reeves, T. C. (2008). Design-Based Research and Educational Technology: Rethinking Technology and the Research Agenda. Educational Technology & Society, 11 (4), P. 26–36.

[2] Pritam Singh Negi, Vineeta Negi, Akhilesh Chandra, Impact of Information Technology on Learning, Teaching and Human Resource Management in Educational Sector, International Journal of Computer Science and Telecommunications [Volume 2, Issue 4, July 2011] P.66-72.

[3] Franklin Wabwoba, Anselimo Peters Ikoha, Information Technology Research in Developing Nations: Major Research Methods and Publication Outlets, International Journal of Information and Communication Technology Research, Volume 1 No. 6, October 2011

[4] John P. Holdren, Thomas C. Power, Pedro I. Espina ,And Pedro I. Espina, The Networking And Information Technology Research And Development (Nitrd) Program 2012 Strategic Plan , Executive Office Of The President National Science And Technology Council Washington, D.C. 20502, July 1, 2012. P.

[5] Mohammad Esmail Ahmad, Ali Jalal Hussian, Monem A. Mohammed, Using Computer Techniques To Predict OPEC Oil Prices For Period 2000 To 2015 By Time-Series Methods, International Journal Of Scientific & Technology Research Volume 5, Issue 06, June 2016 Issn 2277-8616. P. 2.

[6] R.Revarthy, Role of ICT in Providing Quality Education.

[7] Radwan Ali, Irvin R. Katz, Information And Communication Technology Literacy: What Do Businesses Expect And What Do Business Schools Teach? Educational Testing Service, ETS RR-10-17, August 2010

[8] Felix Kayode OLAKULEHIN, INFORMATION AND COMMUNICATION TECHNOLOGIES IN TEACHER TRAINING AND PROFESSIONAL DEVELOPMENT IN NIGERIA, Turkish Online Journal of Distance Education-TOJDE January 2007 ISSN 1302–6488, Volume: 8 Number: 1 Article: 11, P.137-138

[9] Craig Furneaux, How does Information Technology impact the methods, potential and purpose of education?, ETL Conference, 2004, Logan Campus, Griffith University, P.3-4

[10] S.P. Rajguru, ROLE OF ICT IN PROVIDING QUALITY EDUCATION, Directory of Research Journals Researchbib, Vol 3 Issue 11 May 2016, P.1.

[11] Yahia H., Talibe S, Introduction to Computer Science and Problem Solving, Dar Wael, Amman-Jordan, First Edition, 2001, P. 30-31.

[12] Vladimir Kinelev, Piet Kommers, Boris Kotsik, Information and Communication Technologies in Secondary Education. Position Paper, UNESCO Institute For Information Technologies In Education (IITE), Printed In The Russian Federation, 2004, P.11.

[13] Gorana Celebic, Dario Ilija Rendulic, Basic Concepts of Information and Communication Technology, Handbook, Itdesk.Info – Project of Compter E-Education with Open Access, Zagreb, 2011, P.17.

[14] Syed Noor-Ul-Amin, An Effective use of ICT for Education and Learning by Drawing on Worldwide Knowledge, Research, and Experience:ICT as a Change Agent for Education, Scholarly Journal of Education Vol. 2(4), 2013, P.38-39-40.

[15] SeánGallagher, ICT in Education: Policy, pedagogy and practice, INTo education conference Letterkenny 2017, P.9.

[16] Giselle Martins dos Santos Ferreira and Rafael Guilherme Mourão Castiglione, ICT in Education: personal learning environments in perspectives and practices of young people, http://dx.doi.org/10.1590/S1678-4634201702153673, Educ. Pesqui., São Paulo, v. 44, e153673, 2018, P.17.

[17] Noor Agha, B.S. Ghanghas* and P.K. Chahal, Use of Information and Communication Technologies by Extension Personnel to Disseminate Agricultural Information, International Journal of Current Microbiology and Applied Sciences,, Volume 7 Number 04 (2018), P.1370.

Three Levels Quality Analysis Tool for Object Oriented Programming

Mustafa Ghanem Saeed[1], Kamaran HamaAli Faraj[4]
Computer Science Dept, College of Science
Cihan University-Slemani
Slemani, Iraq

Maher Talal Alasaady[2]
Computer system Dept,
Northern Technical University,
Mosul, Iraq

Fahad Layth Malallah[3]
Computer and Information Technology,
College of Electronic Engineering,
Ninevah University, Mosul, Iraq

***Abstract*—In terms of evolution of software engineering methods for complex software developments techniques, new concepts have been emerged in the software languages, which used to develop software quality models. In this research, the Multi Levels Quality Analysis Tool (MLQA) is proposed as a tool for computer-aid software engineering, which classifies software complexity into three levels of analysis, namely the program package analysis, class analysis (program class) and finally the analysis at the level of the program method. MLQA is able to support a method of visual analysis of the software contents with color alerts, and recommendations systems, which can give a quick view of the software development and its complexity. The methodology of this work is a new suggested software quality model based on the standards object-oriented programming complexity metrics as well as threshold limits. In addition, a new quality attribute namely clean code attribute has been proposed and integrating it with the proposed software quality model in a way that enables the user of the model relies on this attribute and reduces the dependence on the software experience, which is expensive and rare at times.**

Keywords—Software quality models; software measurements; clean code; source code complexity metrics

I. Introduction

The need for software quality has been extensively continuous because of the increasing in the community need for software in all aspects of life. The importance of providing the highest quality standards is no longer an advantage, but it is necessary for companies to be successful and competitive [1, 2]. Therefore, there is a unified agreement on the need for software quality, and a number of software quality models and software measurements appeared to solve this problem [2]. Everything in life, if it can be measured, gives the ability to deal with it smoothly, so the importance of software measurements provides indications on the quality and strength of software [3,4]. In addition, the level of development and improvement in the quality models to show the enhancement and defects in the software from one version to another is an important area to be studied [5]. The concept of the clean code is used to distinguish the quality of the code in terms of readability, understanding, structure and complexity [6-9]. All these characteristics are difficult to be identified by depending only on the experiences of programmers and developers [8]. It is necessary to come out with approaches that solve these problems more effectively than the experience of programmers, which may not be available at any time or expensive for the software developer.

The contribution of this paper is proposing a general software quality model to give more flexibility and control during the development of software by making the proposed quality model work on three levels during the analysis of the software product. The proposed model can also display visual chart indictors with recommendations in case some expected defects in software product are found. For this purpose, a tool has been built to measure the quality of the software with an object-oriented programming that measures the quality at three levels: package, class, and function (Method). At the package level, the tool gives the package details in terms of complexity measures and threshold limits on the aspect and quality attributes of Bansiya quality model [4, 20] normalized by threshold value to explain high level of abstraction in software. The Class level has been treated with the extraction of special measures that can be extracted from the class only. With comparison to the appropriate threshold limits, the quality of this class can be judged. At low level of source code (Function level), a quality attribute has been proposed for clean code, which is a set of parameters that are only within the function limits, by which the function can be classified as either having the clean programming code attribute or not. Moreover, the proposed software quality model suggested using of visual color alerts, and recommendations systems in order to help software engineer in the process of evaluating software product.

The organization of the paper is as follows: Section 2 covers the literature review regarding overall previous works on quality models and source code matrix, Section 3 explains the proposed software quality model. While in Section 4 presents the experiment details of this research with the results and discussions. Finally, Section 5 concludes this research with the possible future work.

II. Literature Review

This section reviews the previous works related to the most widely used software quality model to determine the ambiguity of the concept of software quality and the scope of those models in real applications developments. In addition, these works focus on the complexity of the source code (Source Code Complexity Metrics)in terms of the use mechanisms, analysis, and the appropriate threshold for each metrics.

In 2006, Gitter conducted a study on how to apply the Bansiya software quality model to evaluate the development of 19 Azureus versions by making comparisons using object-oriented metrics rather than the metrics proposed by Bansiya and his colleagues. The researcher demonstrated the ability of this model to track the evolution of design quality in several versions by providing access to important information in the internal life of the software. This information can support design decisions at higher levels of abstraction. His proposed model may require additional inputs to cover the highest levels of abstraction to assess all aspects of the quality model at this level [10].Another study by Panfowski (2008) presented a new assessment of software product quality, which focused on assessing the quality of the external features of the software product, which means evaluating the behavior of the software product when implemented. In addition, the study focused on the development of the quality model (ISO / IEC 9126) at the level of software metrics. The study relied on seven samples of the software product and evaluated them using ISO / IEC 9126-2 quality model. In his work, Panfowski concluded that external product quality attributes are an area or category that can be adopted, and that the metrics provided by ISO / IEC 9126-2 can be considered as a starting point for the definition of standards, but are not ready to use in their present form. The metrics of the software product need to be more adapted to show better information [11]. Borgherth (2008) discussed the method of code profiling by using a static analysis. The study was done on (19) industrial samples and (37) samples of students' programs. He has analyzed software samples through software metrics. The results of this study indicated that the code pattern could be a useful technique for rapid program comparisons and quality observation in the field of industrial application and education [12]. Moreover, Bhatti (2010) explored the area occupied by the software metrics. He used a QA-C tool to measure software metrics automatically on the code written in C programming language through expressing the association between software metrics and the complexity of the source code. He attempted to demonstrate the values of these metrics graphically only, without considering the quality features and threshold limits relationship [13].Another work in 2010 is the impact of code complexity and usability, either in monitoring software complexity during development, or in evaluating the complexity of legacy software. The researchers of this work, Goran and Dahiden proposed a new coupling metrics (Ecoup), and introduced the Java met tool, which works in a static analysis of programs written in Java with respect to coupling, flow control, complexity and coherence [14].In the same year, Chandra et al. proposed the use of Object Oriented metrics that introduced by Chidamber and Kemerer (1994)[15] to assess program quality at the class level. The proposed tool can be used to verify the class design conforms to the design specifications of the Object Oriented programming, through using the threshold for each metric [16].

The following is a summary of the most important software quality models:

1) McCall's Factors-Criteria (FCM) model presented by McCall in 1977. The McCall model is the first model of quality [1, 10, 17, 18].

2) ISO / IEC 9126: 2001, which was submitted by the International Organization for Standardization (ISO) in 1991, and has developed six quality metrics. This model was updated in 2001, Quality ISO / IEC 9126: 2001 [10, 13, 18, 19].

3) Dormey's Quality Model, presented by Jeff Dormy (1998) as a quality assessment model, by analyzing the quality of program components through measuring concrete quality characteristics [4, 10].

4) Boehm Software Quality Model presented by the scientist Erwin Bohm in 1978. This model seeks to determine the quality of the program through a predefined set of metrics and measures [10, 11, 19].

5) FURPS Quality Model, introduced by Grady Robert and Hewlett-Packard in 1987. This model focuses on the analysis of quality characteristics in two categories of requirements: functional requirements and non-functional requirements [11, 17, 20].

6) Bansiya Quality Model, proposed by Bansiya in 2002. This model focuses on the quality of Object Oriented Design (QMOOD). It uses the source code metrics extract directly from the software source code to give the quality attribute through the use of mathematical equations [1, 10].

The Bansiya Quality model gives a way to assign source code measurements to higher abstraction levels [10]. Although the experiment results in this model were acceptable, the use of new and non-standard measures in OOP metrics makes this model not widely used, and this is why Gitter in [20] tried to change the measures used in this model to the stranded OOP metrics so that the model becomes more dependable. For this reason, Bansiya model with stranded OOP metrics was used at the software packages analysis stage in the proposed model. Standard OOP metrics used in the proposed model can be viewed in [6, 10, 14, 16, 21-27, 29]. Furthermore, the metrics threshold values that used for the recommendations and alerts are selected of this proposed model due to their usage in the following references and full described in [9, 16, 23, 27-29], which are the reason of the selection.

III. The Framework of Proposed Quality Model

The previous quality models discussed in section II were based on building a relationship between software metrics and design features on the one hand, and design features and quality attributes on the other hand. The relationship could be direct or indirect, either through a paper questionnaire (specific questions prepared by the quality model) or by creating a mathematical relationship with each other as in the Bansiya software quality model. Actually, the idea of using a mathematical relationship between source code metrics extracted directly from the software and the quality attribute is better than using indirect methods such as questionnaires, because indirect methods take a lot of time and a deep knowledge from a software engineer in the project under development. Moreover, it requires a lot of paperwork that exhausts the software engineer. In addition, the analysis process may involve human feelings that affect the accuracy of the judgment.

Therefore, the proposed software quality model suggests four ideas for the analysis of the software product quality: Firstly, the analysis process should be automatic (or semi-automatic) and applied directly on the software source code to reduce time, effort and cost on the software developer and to reduce human errors. Secondly, the software quality model should give informative details of the software product under developments, such as quality attributes to higher abstraction levels, intermediate structure and low-level details. All of these levels must be supported by error detection, alerts, and recommendations systems, because some metrics are just numbers (e.g. line of code) and knowing their impact in software product may require strong software engineering experience. Thirdly, the quality models should use a visual representation with color indicators to demonstrate analysis result that gives a quick full picture of the software product. This enables the software engineer to diagnose the location of the strengths and the weaknesses in the software product under development. Fourthly, the quality models support the ability to view the source code for classes and functions while displaying their own measurements and recommendations.

The proposed model is based primarily on the principles of OOP, which is currently used in almost all software systems, to make the proposed quality model close to the need of developers and programmers. The proposed model singles out the software under evolution into three parts as shown in Fig.1 which includes (Use cases) tool for quality and actor model (Actor), which represents the end-user model. The idea of dividing the software is to enable the developer team to correct errors and reduce complexity as much as possible in the next development cycle or the rewriting of the software code as in the extreme programming development approach.

The three parts of the analysis of software product are as follows:

1) Quality attributes that affect the quality of packages in general. In this level of analysis, the proposed model suggests using version of Bansiya Quality Model (with proposed modification), because it is very informative in this level.

2) The quality attributes that affect the quality of the class. The proposed model used the set of metrics suitable for object-oriented design suggested by Gadabber and some other measures of the OOP and linking them with the metrics thresholds to generate an evaluation and recommendations for the source code of the class.

3) Quality attributes that affect the quality of the functions (Methods) in the class. The proposed model suggested new attributes of the clean code. Through this attribute, the user can evaluate the quality of the function.

The software metrics, especially the complexity metrics of the source code, should be classified according to the three levels of analysis of the proposed quality model as shown in Fig. 2, which describes the classification of these metrics according to their impact on quality aspects.

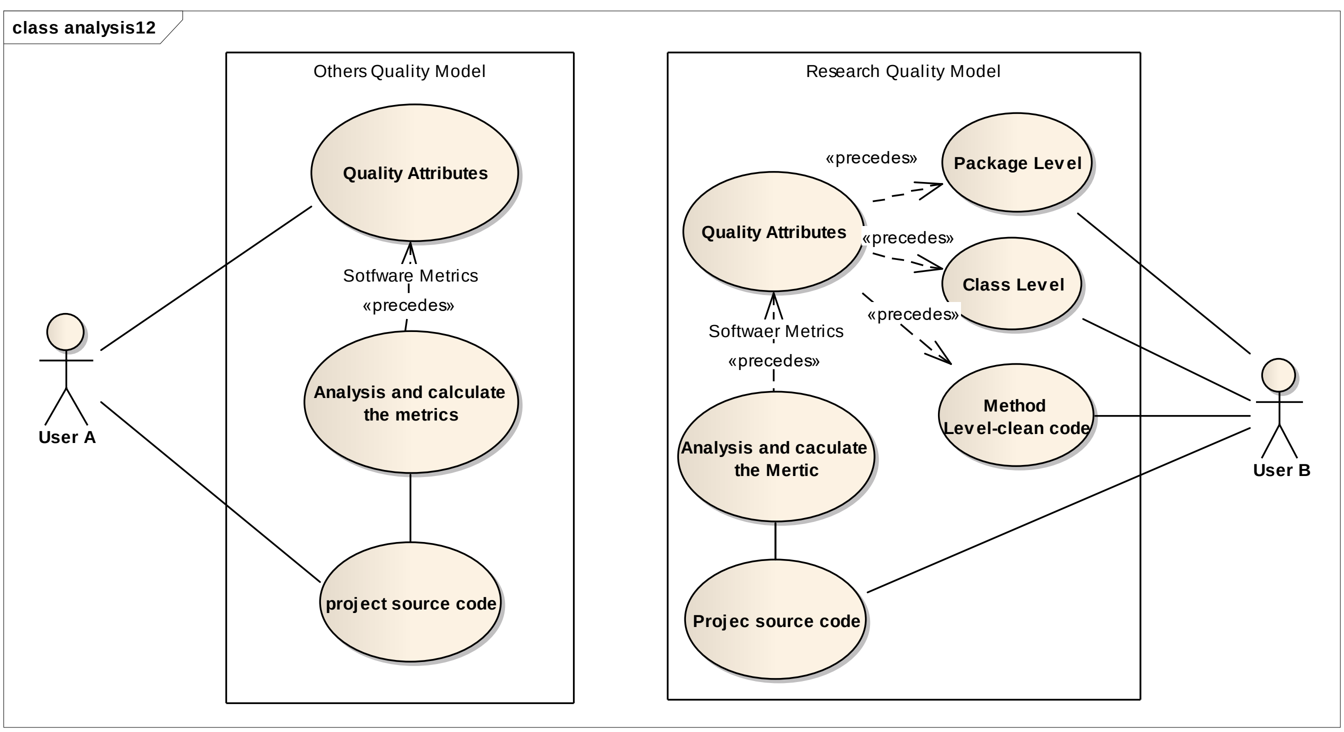


Fig. 1. The proposed quality model with other quality models from the point of view of the user.

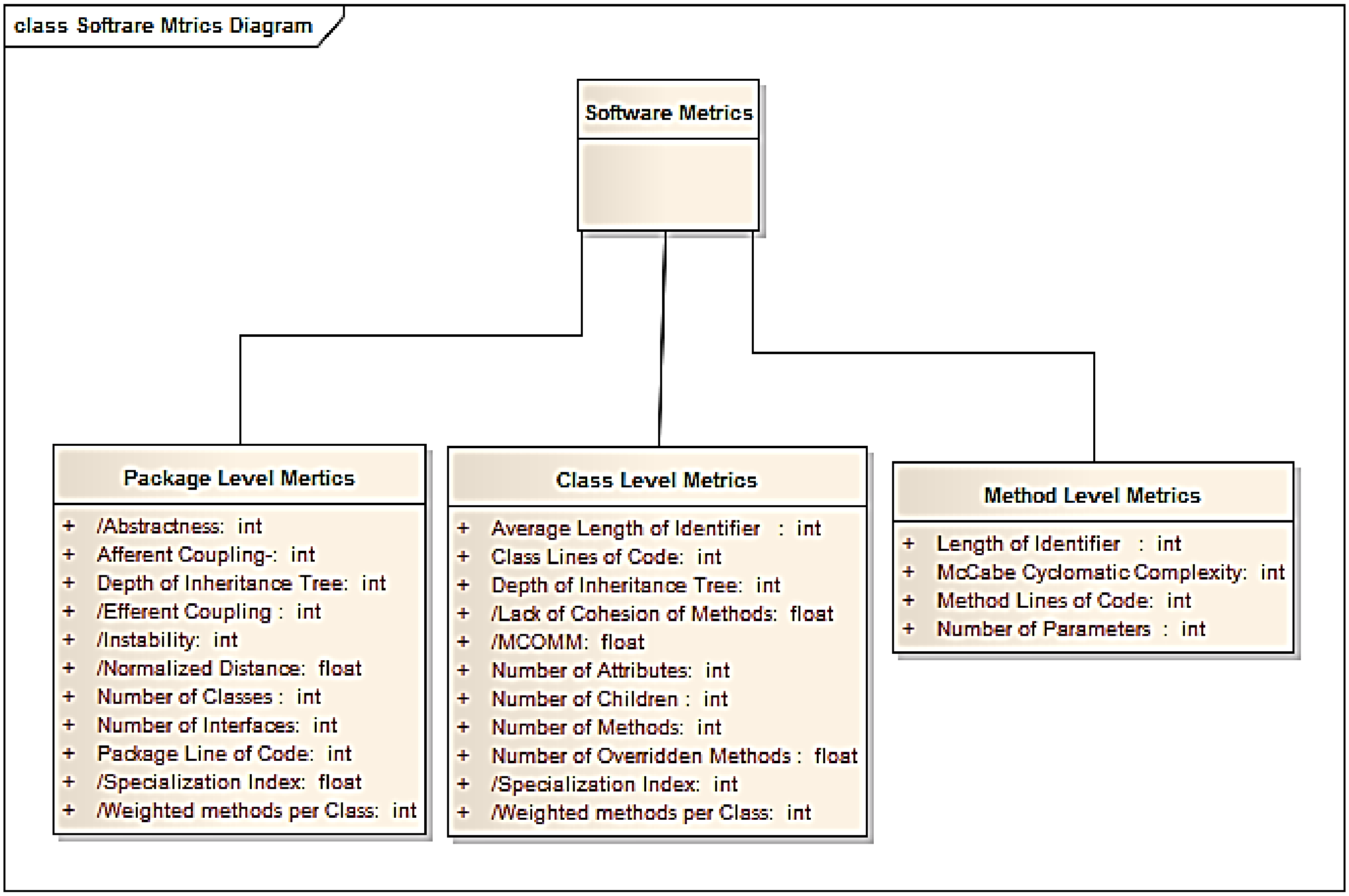


Fig. 2. The class diagram describing the classification of software metrics.

This model focuses on showing the complexity of the source code at the level of the software function within the class because the previous quality models focused on the software as a whole and did not identify points that are likely to be more informative; instead of that, these provided just general indicators. Analysis of the complexity at the level of the source code (low-level components in the software package) is very close to the software developer team. Therefore, the proposed quality model has more effective and practical results to help software developer team during decision-making and improvement of the software product. One of the most important things that affect the accuracy of the analysis of software product at this level is which complexity metrics with suitable threshold value to give clear and efficient assessments.

A. Proposed Model for Function Level

The proposed quality attribute at the function level is the clean code attribute, and the value of this attribute is the binary type (the function either has the character of the clean code or not). This value is determined by the function complexity metrics by linking it to the appropriate threshold value to ensure that the source code of the function has low-complexity, readable and easy to understand. This makes software engineering more efficient in inspection and maintenance of the function source code. As shown in Table I, the complexity of the source code is within the limits of the clean code at the function level.

TABLE I. THRESHOLD LIMITS AND RECOMMENDATIONS AT THE FUNCTION LEVEL

Metric	**Threshold limit**	**Recommendations over the Threshold limit**	**Recommendations under Threshold limit**
Method Lines of Code (mLOC)	Less than 50	Segment the function into new functions by re-structuring the function.	No Recommendations
McCabe's cyclomatic complexity (CC)	Threshold limit in the Table II	Code Reconstructand optimize the function to reduce complexity.	Recommendations as in the Table II
Number of parameters	Less than or equal to 3	Check the function and make sure that the function performs one task	No Recommendations
Average length of identifier	Greater than 10	Rename identifiers in order to be more clearly.	No Recommendations

According to [22], the threshold limit of McCabe Cyclomatic complexity should be classified into the categories as in Table II which shows the appropriate threshold limit of Cyclomatic complexity.

TABLE II. RECOMMENDATIONS ON THE MEASURE OF MCCABE'S CYCLOMATIC COMPLEXITY (CC) AND ITS RELATION TO RISK [24]

Cyclomaticcomplexity value	Functions Types	Possible Risk
1 to 4	Simple function	Low
5 to 10	The function is well structured and stable	Low
11 to 20	A complex function	Medium
21 to 50	A complex and troubling function	High
More than 50	A function that generates errors and is Extremely annoying and unstable	Very high

The value of the clean code attribute will be evaluated as a clean code if the threshold is exceeded; otherwise, it will be evaluated as a non-clean code.

B. Proposed Model for Class Level

The proposed quality model at this level generates recommendations on measures that affect the complexity of the product depending on the threshold limit as in Table III.

C. Proposed Model of Package Level

The proposed quality model at this level generates recommendations on measures that affect the complexity of the product depending on the threshold limit as in Table IV.

TABLE III. THRESHOLD LIMITS AND RECOMMENDATIONS ON THE LEVEL OF CLASS

Metric	Threshold limit	Recommendations above the Threshold limit	Recommendations under Threshold limit
Class Lines of Code (cLOC)	Greater than 500	Class segmentation to more than one Class	No Recommendations
Average number of McCabe's cyclomatic complexity (CC)	Greater than 10	The class is complex	A well-structured Class
Number of Methods (NM).	Greater than 20	Functional examination of the class	No Recommendations
Number of direct Children (NOC)	Greater than 6	High reusablethus requires examination of class carefully because it depends upon a large number of Classes	Indicating no reuse in the class
Number of Methods overridden (NMO).	Greater than 3	The class is complex and difficult to understand	No Recommendations
Weighted Methods per Class(WMC)	Greater than 15	The class is complex	A well-structured Class
Depth of Inheritance Tree (DIT)	Greater than 5	The complexity of the class as a whole is increasing and there is difficulty in calculating the behavior of the class	No Recommendations
Average length of IDs	Less than 8	IDs are not clear enough	No Recommendations
Lack of Cohesion in Object Methods(LCOM)	Greater than 0.6	Class achieves different purposes and should be divided into several sub-classes	No Recommendations

TABLE IV. THRESHOLD LIMITS AND RECOMMENDATIONS ON THE LEVEL OF PACKAGE

Metric	Threshold limit	Recommendations above the Threshold limit	Recommendations under Threshold limit
Average Weighted Methods per Class(aWMC)	Greater than 3.	The package is complex	A well-structured package
Average Number of Methods overridden (aNMO).	Greater than 15.	The classesare complex and difficult to understand	No Recommendations
Abstractness – RMA	Greater than (0.5)	Abstract package	Cohesive package
Normalized Distance from Main Sequence- Dn	Greater than (0.5)	The package is unstable	The package is stable
Instability-RMI	Greater than (0.5)	The package must be inspected and re-structured	The package is well-designed

D. Modified Bansiya model of Package Level

At the level of the package, the proposed model also suggests using the attributes of Bansiya quality model. These attributes can be used for the purpose of making comparisons between software packages to know the best in terms of the closest to the requirements of the user, and this helps the software engineer to choose the best packages during reuse or use of commercial packages. The quality model of the object-oriented design [1] consists of four levels and three relationships among them as in Fig. 3. First Level (L1) represents quality attributes and it should be wide enough to cover all aspects of design quality.

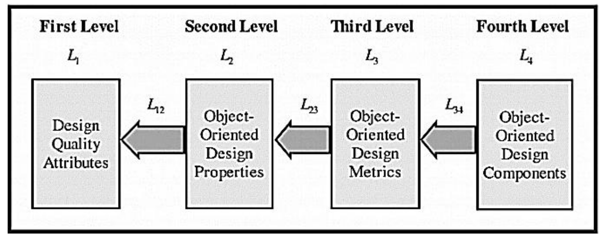


Fig. 3. A Quality model of the object oriented [1].

These qualities are targeted at design quality. A definition of quality characteristics are well explained in Table V.

TABLE V. DEFINITION OF QUALITY ATTRIBUTES OF BANSIYA [5]

Quality Attributes	Definition
Reusability	Reflects the existence of characteristics of object-oriented design that allow the design to be reused to a new problem without much effort.
Flexibility	Characteristics that allow inclusion of changes in design, or adaptive design ability to provide functional-related capabilities.
Understandability	Design characteristics that make the design easy to learn and understand. This relates to the complexity of the design structure directly.
Functionality	The responsibilities assigned to the design of the classes, which are provided through the public interfaces.
Extendibility	Refers to the presence and use of features in the current design that allows integration with the new requirements.
Effectiveness	Refers to the design ability to achieve desired functionality and behavior using object-oriented design concepts and techniques.

TABLE VI. QUALITY CHARACTERISTICS OF BANSIYA AND ITS MEASUREMENTS BY OBJECT-ORIENTED DESIGN METRICS

Design characteristics	Definition	Measured with object-oriented design metrics [10]
Design size	Measures the number of classes used in the design	*NumberofClasses*
Hierarchies	Hierarchy is used to represent different concepts of generalization and specialization in design. It is calculated from the number of inherited classes that have children in design.	*DepthofInheritanceTree*
Abstraction	A measure of the aspects of generalization and specialization in design. Items in design that have one or more offspring carrying the property of abstraction	*Abstractness*
Encapsulation	It is defined as combining data and behavior within a single structure.	No equivalent in object-oriented design measures [20]. This value has been counted as one.
Coupling	Specifies the interconnection of an object with other objects in the design. It is a measure of the number of objects to be accessed by a particular object until it works correctly.	*Instability*
Cohesion	Evaluates the relationship between functions and variables in the class. Strong overlap in function parameters, variable types is a sign of strong cohesion.	(1)
Composition	Measures relationships: (part-of), (has), (consist-of-part), and (part - whole) relationships, which are aggregation relationships in object-oriented design.	*NumberofAttributes*
Inheritance	A measure of the "is-a" relationship between the classes. This relationship is related to the level of nesting of the hierarchical structure.	(2)
Polymorphism	The ability to replace objects that have similar interfaces with each other at runtime. It is a measure of the services that are dynamically determined at runtime in an object.	*NumberofOverriddenMethods*
Messaging	Calculates the number of public functions that provide services for other classes. This is a measure of the services provided by the class.	*NumberofMethods*
Complexity	A measure of the degree of difficulty in understanding and absorbing internal and external structural classes and their relationships.	*WeightedMethodsperClass*

The third level (L3) represents design characteristics. Design characteristics as explained in Table VI can be measured by standard object-oriented design metrics instead of using non-standard metrics that suggested by the original model. It expresses the degree of design compatibility to suit the specific properties in L2. L3 is an additional level compared to the Dormey's Quality Model [4].

The tangible elements of L4are converted into digital form and this made the proposed model gain greater objectivity [10]. As it can be seen, the design of software components written using the object-oriented programming shown in Fig. 3 represents the fourth level of the model (L4). These components are mostly determined through the programming language (for example, functions, objects, and classes). L4 delivers sources that are source code (functions, classes, packages, etc.) that will be measured at the top of level (L3). The next step consists of setting design metrics for design properties (L23). This model selects only one metric for each of the design characteristics. The L2 and L3 relationship was direct and traceable, as shown in the third column of Table VI since the object-oriented design metrics are used to evaluate design characteristics. L12 step probably is the most important step; this step is blending design characteristics with quality attributes. The model uses weights for the design characteristics for obtaining quality attributes value as shown in equations (3), (4), (5), (6), (7), and (8) [1, 10].

The weights can be either positive or negative. The algebraic sign indicates that a particular design characteristic has a positive or negative effect respectively on the quality Attributes [1, 10]. For example, Reusability is positively affected by the Design size (the greater the number of classes, the greater the possibility of reuse). All weights of design characteristics have a specific range between [*-1* and *+1*], so that all quality attributes are in the same range. Positive effects have the values of initial weights *(+1)* or *(+0.5)*. The negative effects were selected (*-1*) or (*-0.5*), and then this value was changed relatively to bring the total results of weights (± *1*).

For a better illustration of the relationship between quality characteristics and design characteristics, Table VII explains these relationships. The symbols in the table VII indicate the type of relationship, where the symbol▲ indicates a positive correlation between design characteristics and quality attributes, i.e. the better the design size value, the better the reuse. However, the symbol indicates a negative relationship, i.e. high coupling value reduces flexibility. The original model was based on the choice of one of the software package as a basis for the process of normalizing the results of quality attributes to display the last results. This makes the results of this model unstable because the results will be changed whenever the basis of comparison is changed. Therefore, the researcher believes that relying on the limits of the threshold in the normalization of quality attributes will increase the stability of the results of the proposed model and this leads to increase the reliability of the proposed quality model.

$$\frac{1}{NumberofOverriddenMethods} \qquad (1)$$

$$1 - \frac{NumberofOverriddenMethods}{NumberofMethods} \qquad (2)$$

$$Reusability = -0.25 * Coupling + 0.25 * Cohesion + 0.5 * Messaging + 0.5 * DesignSize \qquad (3)$$

$$Flexibility = 0.25 * Encapsulation - 0.25 * Coupling + 0.5 * Composition + 0.5 * Polymorphism \qquad (4)$$

$$Understandability = -0.33 * Abstraction + 0.33 * Encapsulation - 0.33 * Coupling + 0.33 * Cohesion - 0.33 * Polymorphism - 0.33 * Complexity - 0.33 * DesignSize \qquad (5)$$

$$Functionality = 0.12 * Cohesion + 0.22 * Polymorphism + 0.22 * Messaging + 0.22 * DesignSize + 0.22 * Hierarchies \qquad (6)$$

TABLE VII. THE RELATIONSHIP BETWEEN QUALITY ATTRIBUTES AND CHARACTERISTICS OF THE DESIGN [1]

Design characteristics	**Quality attributes**					
	Reusability	**Flexibility**	**Understandability**	**Functionality**	**Extendibility**	**Effectiveness**
Design size	↑		↓	↑		
Hierarchies		↑		↑		
Abstraction			↓		↑	↑
Encapsulation		↑	↑			↑
Coupling	↓	↓	↓		↓	
Cohesion	↑		↑	↑		
Composition		↑				↑
Inheritance					↑	↑
Polymorphism			↓	↑	↑	↑
Messaging	↑			↑		
Complexity			↓			

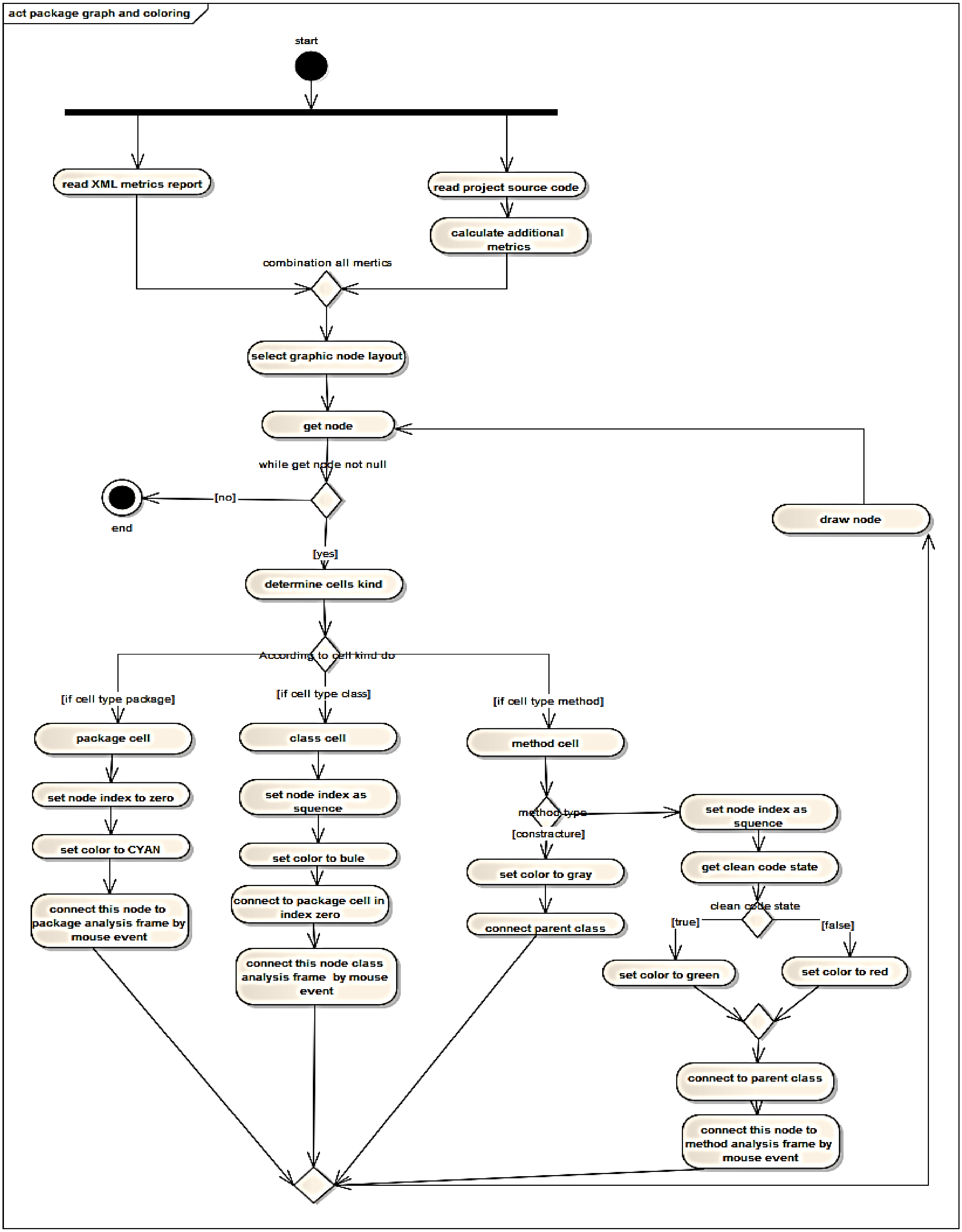


Fig. 4. Proposed graphical analysis algorithm for the package and the coloring.

$$Extendibility = 0.5 * Abstraction - 0.5 * Coupling + 0.5 * Inheritance + 0.5 * Polymorphism \quad (7)$$

$$Effectiveness = 0.2 * Abstraction + 0.2 * Encapsulation + 0.2 * Composition + 0.2 * Inheritance + 0.2 * Polymorphism \quad (8)$$

E. Visual Representation of the Proposed Quality Model

Graphical representation and color have been used for the proposed software quality model to illustrate the software parts as one of the quality models requirements. The graphical representation, in general, is more intuitive and efficient in understanding than just a table of numbers and values. Therefore, that makes software engineers create a complete and rapid view for software under development. Different colors can also be used to illustrate the complexity of each part of the software under development, in which the algorithm for the analysis and colorization of the parts of the software under development is illustrated in Fig. 4 by using the UML activities diagram, which is designed in line with the Java presser package. The algorithm description of the proposed model is as following steps:

Step *1: Read the source code of the project (Java files) to be analyzed.*

Step *2: Read the software metrics from the xml file for the project to be analyzed.*

Step *3: Calculate the additional metrics from the source code of the project to be analyzed in the first step.*

Step *4: Integrate all software metrics together from Step 3 & 2.*

Step *5: Determine the package components graphically: (radial tree, tree, compact tree, fast organic).*

Step *6: For each node in the drawing, apply Step 7 and 8.*

Step *7: Determine the characteristics of each node in the graphical analysis by type*

1) If the node type is a package, give the node following characteristics:

- It is defined as the root of the tree in the graphic.
- The cyan color (CYAN) is given.
- Binds with the display interface of packet analysis.

2) If the node type is a class, give the node following characteristics:

- An address shall be given according to its sequence in the graphic.
- The blue color (BLUE) is given.
- Binds to the display interface of the class analysis.

Step *8: If the node type is a function, do the following:*

1) If the function type is Constructor(special function for building the object), give the node following characteristics:

- An address shall be given according to its sequence inthe graphic.
- The gray color (GRAY) is given.

2) If the function is a regular function, give the node following characteristics:

- An address shall be given according to its sequence inthe graphic.
- A green color (GREEN) is given if the node has a clean code attribute and a red color (RED) is given if not.
- Binds with display interface function analysis

Step *9: Show the graphic.*

IV. Experiment with Result and Discussion

The proposed software namely Multi Levels Quality Analysis Tool (*MLQA*) is developed to analyze and evaluate the engineered software. The tool was programmatically based on *Metrics 1.3.6* (Eclipse Metrics plugin- Provide metrics calculation and dependency analyzer plug in for the Eclipse platform) as well as a *Java doc parser* to calculate the non-supported metrics in Metrics 1.3.6.*Java doc parser* was also needed to extract code information. In addition, for graphical representation, the *MLQA* used package *Jgraph 5.13*. Moreover, *XML Doc Parser* to read XML reports from *Metrics 1.3.6* was built as part of *MLQA* tool.

To test the proposed algorithm, three Java programs have been tested in this experiment to demonstrate the proposed tool results and their practicality in software analysis, these programs are:

1) Patience game: This game is used for the purpose of learning how to deal with arrays and the structure of data in Java language. The source code of the game was downloaded from the website (http://www.neiljohan.com/java).

2) Syntexchecker game: This game is used to learn how to handle matrices and data structures in Java. The source code of the game was downloaded from the website (http://www.neiljohan.com/java).

3) Payroll System (payRoll): An accounting system that deals with databases and reports in Java. The source code (payRoll) was downloaded from (http://www.projectsparadise.com).

To illustrate the test of the three projects, the results of the project (Patience Game) were explained at all levels of analysis. The remaining projects will be later compared with the first project results. The Patience Game consists of seven classes including two interfaces (interface). As illustrated in Fig. 5, graphical analysis of the source code was executed using the tool (*MLQA*) and the style used to display the package components is the radial tree.

The color of each node in Fig. 5 has its different meaning as explained in section 3.5. The results of the Fig. 5 are explained in Table VIII.

TABLE VIII. RESULTS OF THE GRAPHICAL ANALYSIS OF THE PATIENCE GAME PACKAGE

Class Name	Number of Total Functions	Number of clean code functions	Number of unclean code functions
CardTest	2	1	1
PlayCardImp	7	7	0
Patience	5	2	3
NeilClass	4	0	4
PackImp	10	9	1

Table VIII shows that the **PlayCardImp** functions all have the clean code attribute. This is a good indicator of the design quality of this class. The **PackImp** class had one function, which does not have the clean code attribute out of 10 functions, the researcher believes that this is a good indicator. The **CardTest** class has two functions, one with a clean code attribute and the other without it. While the **Patience** class has five functions, three of them do not have the clean code attribute, and some complexity can be found in this class. All **NeilClass** functions do not have clean code, this is an indicator of poor design quality, and this is the most complex class in this package, so it should be further emphasized during inspection and maintenance.

The proposed algorithm has been designed as GUI to give details for the user, these details view of nodes components with a separate interface for each node, which gives the software engineer the ability to view software complexity metrics and recommendations with source code. The results can be presented according to the three levels of the proposed quality model, in which for each one of these 3-level, there is a separated GUI. First level, Package-level analysis results as depicted in Fig. 6 shows package details in terms of complexity metrics and threshold limits as well as the quality attributes of update Bansiya quality model for the highest level of abstraction.

Package-level results can be used to compare or track package development during software development. It is observed in Analysis and Recommendations for this package section in Fig. 6, that the Patience Game package has good properties in terms of the NORM metric, which is affecting the readability and comprehension, while the package is complex according to WMC metric because this metric directly affects on the complexity of the package in general. Moreover, the Abstractness (RMA) and Normalized Distance from Main Sequence (*Dn*) metrics affect on the quality of the package design in terms of incoming and outgoing pairing, and these measures were within the threshold. However, the Instability (RMI) metric was far from the threshold, so this package is unstable.

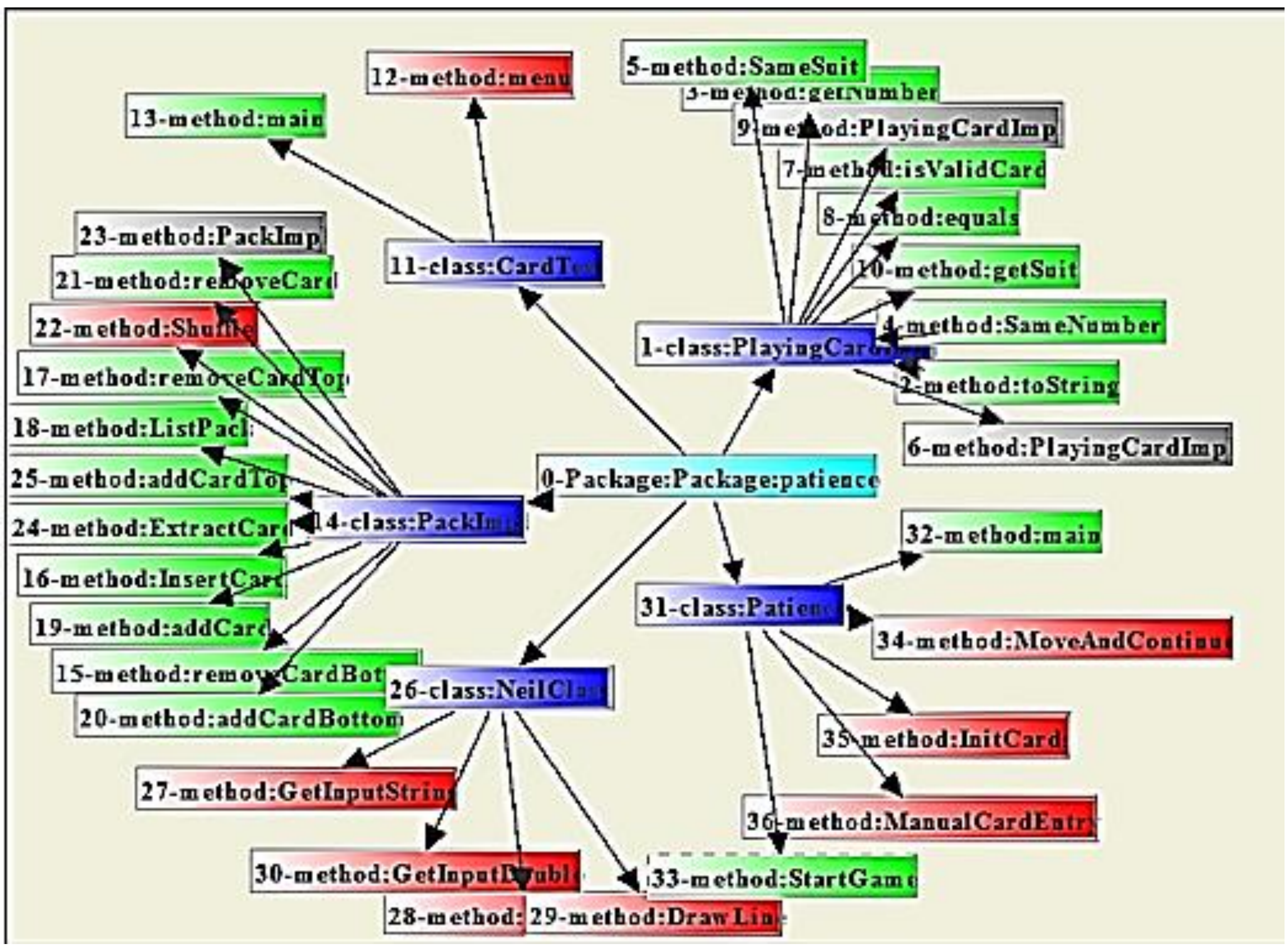
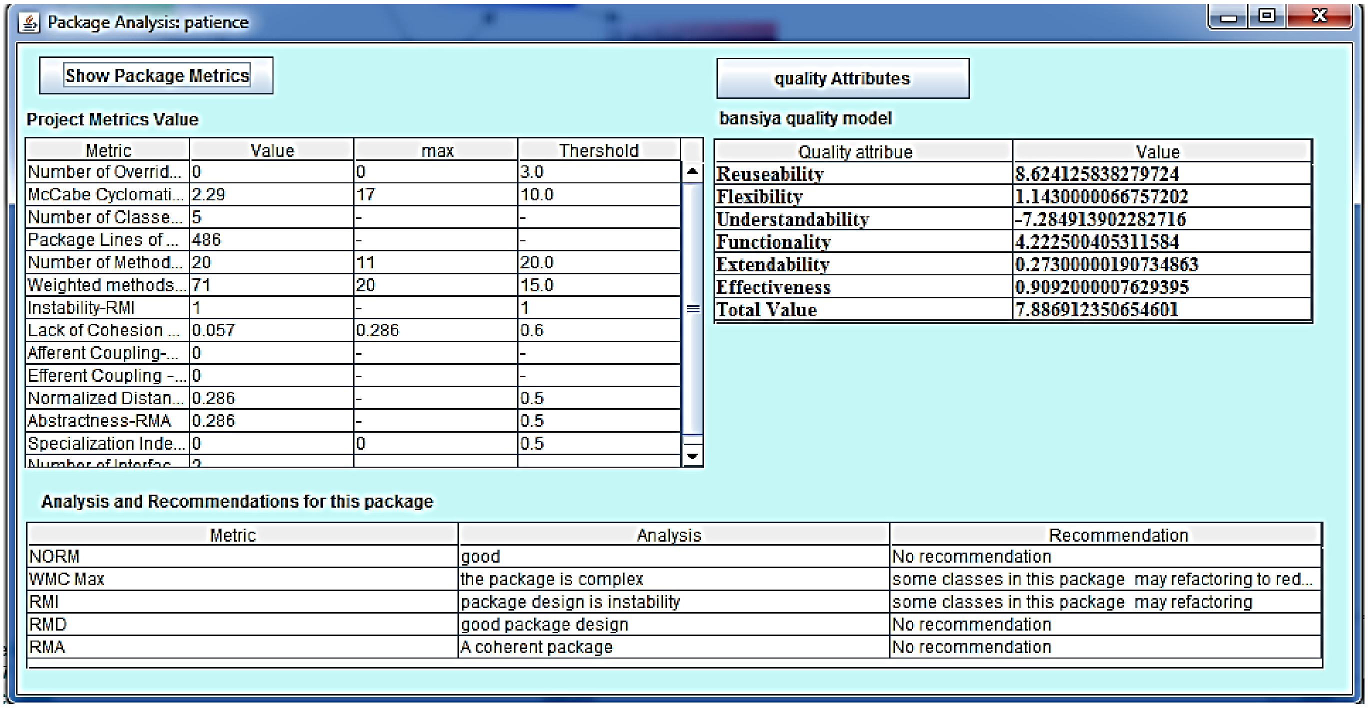


Fig. 5. Shows the graphical analysis of the components of the Patience Game.



Metric	Value	max	Thershold
Number of Overrid...	0	0	3.0
McCabe Cyclomati...	2.29	17	10.0
Number of Classe...	5	-	-
Package Lines of ...	486	-	-
Number of Method...	20	11	20.0
Weighted methods...	71	20	15.0
Instability-RMI	1	-	1
Lack of Cohesion ...	0.057	0.286	0.6
Afferent Coupling-...	0	-	-
Efferent Coupling -...	0	-	-
Normalized Distan...	0.286	-	0.5
Abstractness-RMA	0.286	-	0.5
Specialization Inde...	0	0	0.5
Number of Interfac...	2		

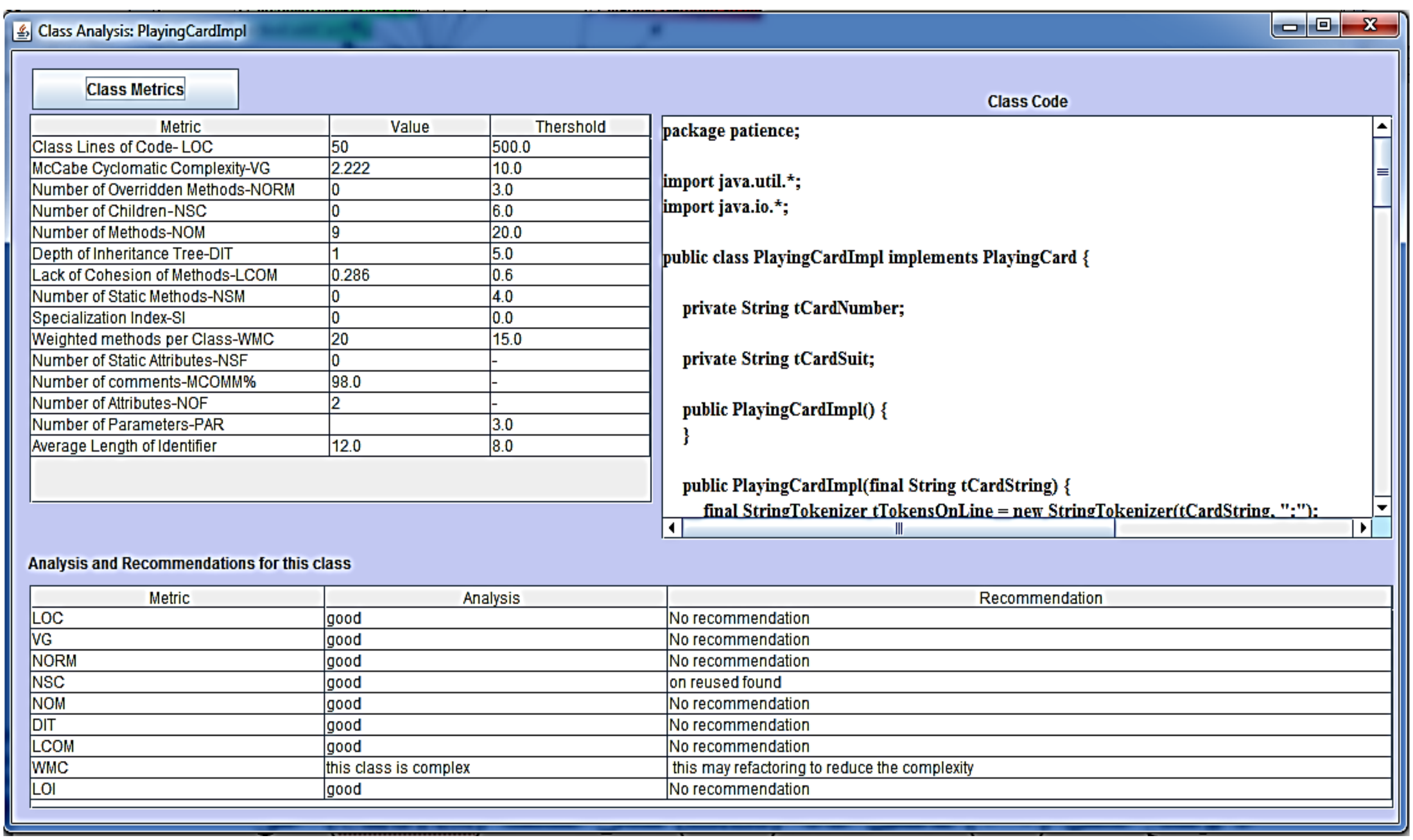
Quality attribue	Value
Reuseability	**8.624125838279724**
Flexibility	**1.1430000066757202**
Understandability	**-7.284913902282716**
Functionality	**4.222500405311584**
Extendability	**0.27300000190734863**
Effectiveness	**0.9092000007629395**
Total Value	**7.886912350654601**

Metric	Analysis	Recommendation
NORM	good	No recommendation
WMC Max	the package is complex	some classes in this package may refactoring to red...
RMI	package design is instability	some classes in this package may refactoring
RMD	good package design	No recommendation
RMA	A coherent package	No recommendation

Fig. 6. Results of the package level analysis.

Secondly, Class-level analysis results consisting of the values of the complexity metrics of the class with the recommendations depend on the threshold for these metrics.

To illustrate the ability of the tool at this level, **PlayCardImpl** class has been chosen to explain the results of class analysis support by MLQA tool as shown in Fig. 7.



Metric	Value	Thershold
Class Lines of Code- LOC	50	500.0
McCabe Cyclomatic Complexity-VG	2.222	10.0
Number of Overridden Methods-NORM	0	3.0
Number of Children-NSC	0	6.0
Number of Methods-NOM	9	20.0
Depth of Inheritance Tree-DIT	1	5.0
Lack of Cohesion of Methods-LCOM	0.286	0.6
Number of Static Methods-NSM	0	4.0
Specialization Index-SI	0	0.0
Weighted methods per Class-WMC	20	15.0
Number of Static Attributes-NSF	0	-
Number of comments-MCOMM%	98.0	-
Number of Attributes-NOF	2	-
Number of Parameters-PAR		3.0
Average Length of Identifier	12.0	8.0

```
package patience;

import java.util.*;
import java.io.*;

public class PlayingCardImpl implements PlayingCard {

  private String tCardNumber;

  private String tCardSuit;

  public PlayingCardImpl() {
  }

  public PlayingCardImpl(final String tCardString) {
    final StringTokenizer tTokensOnLine = new StringTokenizer(tCardString, ":");
```

Metric	Analysis	Recommendation
LOC	good	No recommendation
VG	good	No recommendation
NORM	good	No recommendation
NSC	good	on reused found
NOM	good	No recommendation
DIT	good	No recommendation
LCOM	good	No recommendation
WMC	this class is complex	this may refactoring to reduce the complexity
LOI	good	No recommendation

Fig. 7. Analysis results of the AnalysisCardImp class.

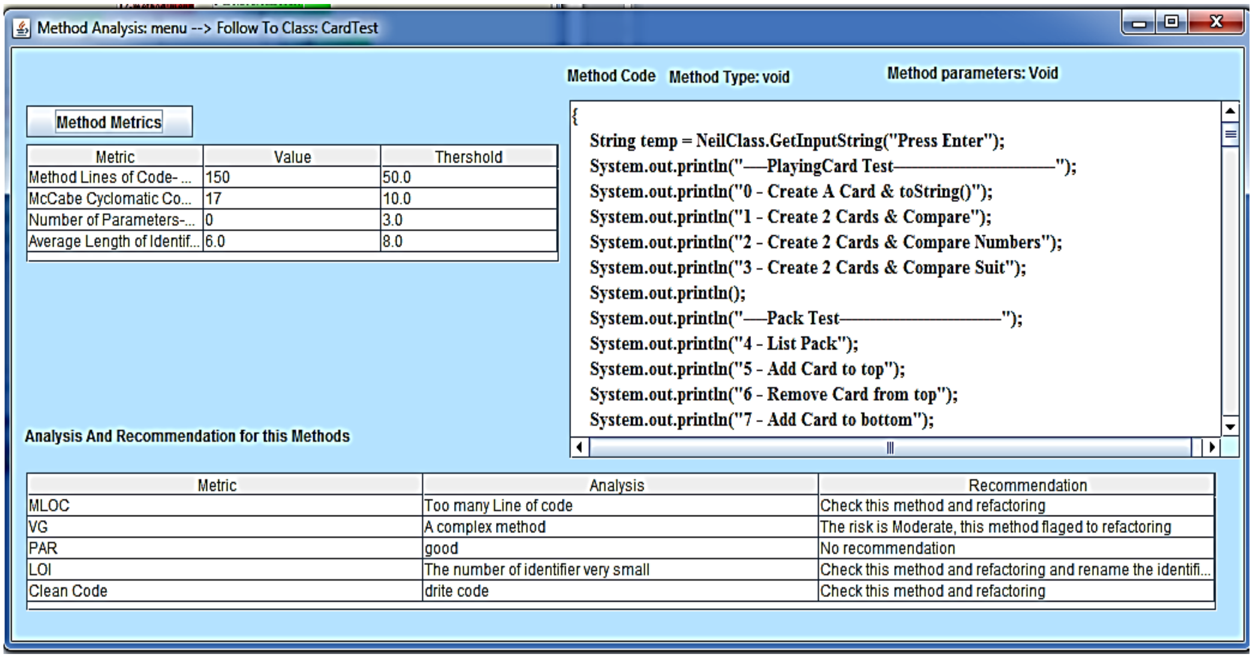


Fig. 8. Shows the internal details of the function (menu) that does not have the clean code attribute.

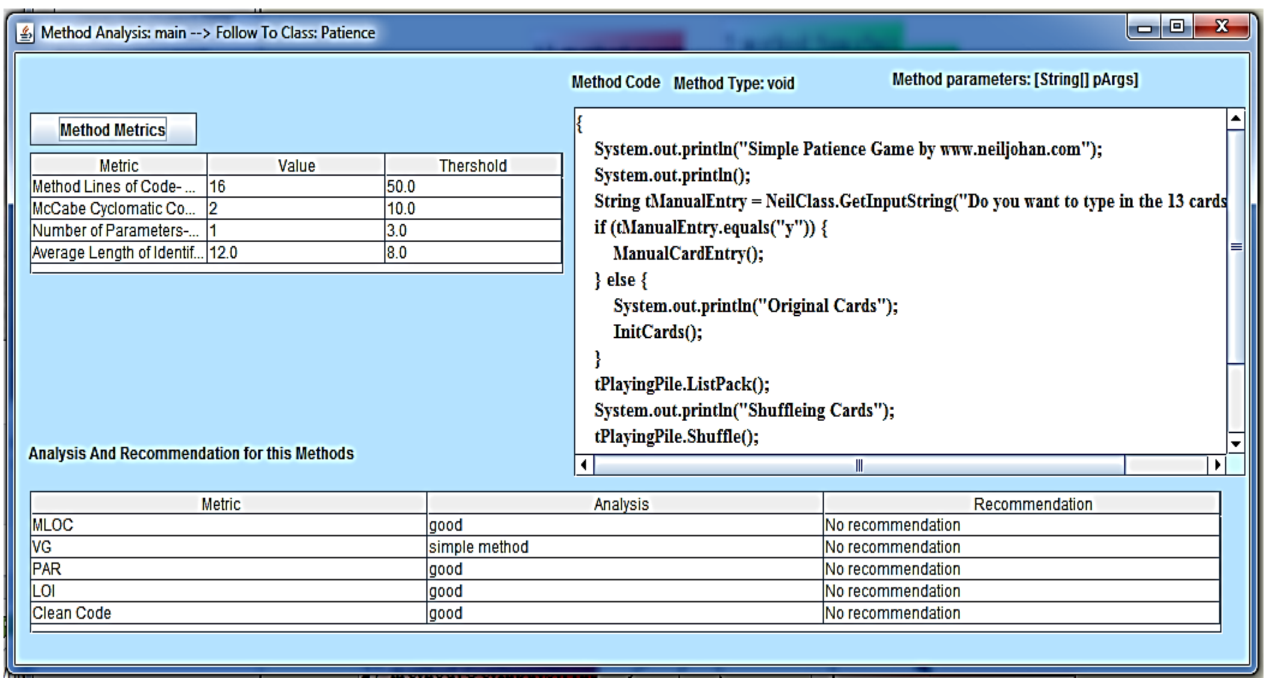


Fig. 9. Internal details of the function (main) that have the clean code attribute.

Here the point of view of the researchers observed that **PlayCardImpl** class has a good quality design in terms of Class Lines of Code (cLOC), the Average number of McCabe's Cyclomatic complexity (CC), the Number of Methods (NM), the Number of direct Children (NOC), Depth of Inheritance Tree (DIT), the Average length of IDs, Number of Methods overridden (NMO) and Lack of Cohesion in Object Methods (LCOM) because they were within the limits of threshold. However, this class is considered complex because the Weighted Methods per Class (WMC) metric has moved away from the appropriate threshold, exceeded by (20) while the appropriate threshold limit is (15). Moreover, the results of any class in the package can be displayed in this way.

Thirdly, Function-level analysis, in order to clarify this level, the two functions in two different classifications were selected, as illustrated in Fig. 8 function details (**menu**) within the class (**CardTest**), which classified as it does not have the clean code attribute.

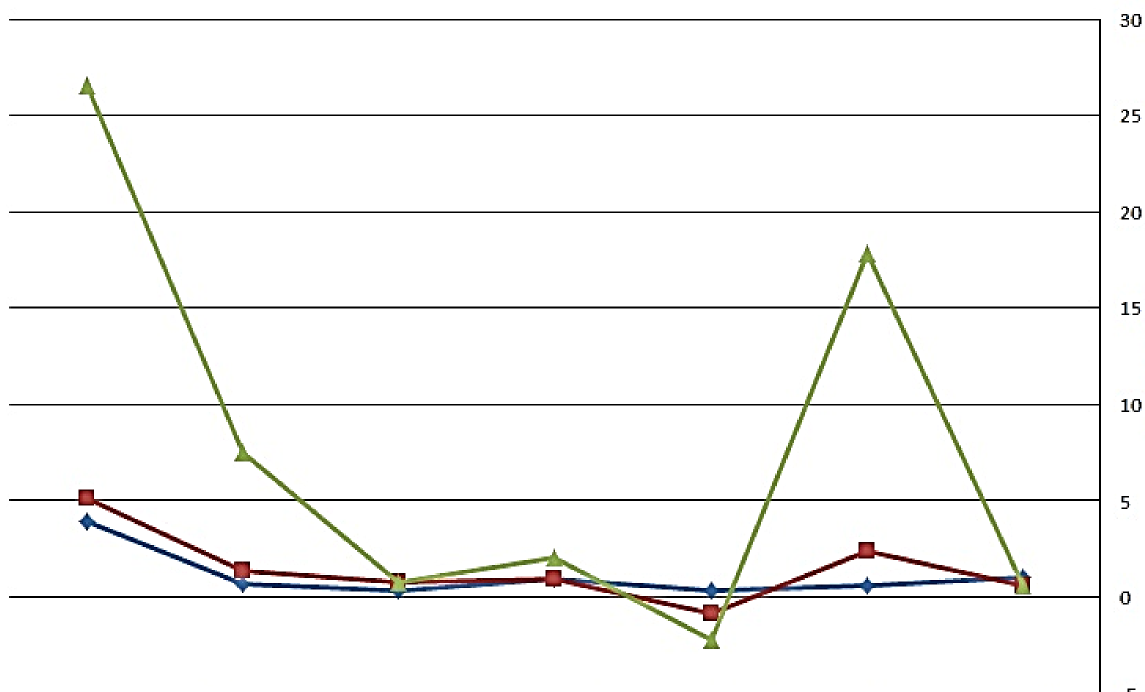
The researcher noted that the menu function was classified unclean code because the software complexity measures are beyond the scope of the clean code attribute. The researcher believes that this function should be restructured or it will lead to difficulties in the inspection, maintenance and reuse process in class. While the function (**main**) in Patience class described in Fig. 9 was classified as it has the clean code attribute because the software complexity measures are within the scope of the clean code attribute, which is a good indicator of the design of this

The researcher noted that the menu function was classified unclean code because the software complexity measures are beyond the scope of the clean code attribute. The researcher believes that this function should be restructured or it will lead to difficulties in the inspection, maintenance and reuse process in class. While the function (**main**) in Patience class described in Fig. 9 was classified as it has the clean code attribute because the software complexity measures are within the scope of the clean code attribute, which is a good indicator of the design of this function. Moreover, the same method of displaying results at the function level can be used for all functions within the package to be analyzed. In order to illustrate the quality attributes of the update Bansiya quality model at the package level, the researcher chose three software (patience game, syntexchecker game, payRoll accounting system), which were analyzed using the MLQAtool, and the results were shown in Table IX.

TABLE IX. COMPARISON OF SOFTWARE (PATIENCE, SYNTEXCHECKER, PAYROLL), USING THE QUALITY ATTRIBUTES OF THE UPDATE BANSIYA QUALITY MODEL

Software Name	**Reusability**	**Flexibility**	**Understandability**	**Functionality**	**Extendibility**	**Effectiveness**	**Total value**
Patience	1.02	0.576	0.38	0.92	0.36	0.68	3.94
syntaxChecker	0.60	2.36	0.89	0.91	0.79	1.38	5.15
PayRoll	0.60	17.83	-2.25	2.07	0.79	7.51	26.56

The researchers observed from the Table IX that an increase in the Effectiveness attribute and the Flexibility attribute strongly conflict with Understandability attribute, while with Reusability, functionality and Extendibility attributes, the effect has been reduced, although the values are also relatively large. However, the differentiation between the values of these attributes is due to the non-functional requirements of the customer, so that the software engineer can make appropriate decisions in the light of these values. These features were illustrated graphically in Fig. 10 to give a more comprehensive picture of the relationship among these attribute.



	Total Value	effectiveness	Extendability	Functionality	Understandability	Flexibility	Reuesability
patience	3.944905671	0.680666667	0.368333325	0.92044707	0.379666092	0.571666673	1.024125844
syntaxChecker	5.158801762	1.386395192	0.799321353	0.91110431	-0.898028028	2.359166563	0.600842372
PayRoll	26.56566665	7.519047606	0.797619015	2.070666655	-2.254999934	17.83333331	0.599999994

Fig. 10. Chart for quality attributes in the three tested software projects.

V. CONCLUSION

To sum up this paper, the current software quality models suffer from some ambiguities during the analysis of the quality of the software as there is difficulty for the software engineer during apply quality model because it requires direct intervention for all operations manually and exhaustively, which may affect on the process of software development. Therefore, quality model and support tools are developed in this paper to be used by software engineers to control software product written in Java programming language named Multi Levels Quality Analysis Tool (*MLQA*) as evaluation algorithm for software quality model comprising software product, software metrics, and clean code programming has been proposed based on three levels of abstraction as *package, class, and function*, as it is proved that three levels give more accurate results and recommendations from analysis results of code rather than from one level of abstraction, because software measures that are appropriate for function-level analysis may not be suitable for the class or package level. Moreover, the proposed *MLQA* has been boosted with a graphical analysis of color discrimination to give a quick look to the software engineer about the complexity of the software, as well as to give more ability to update the source code of the software package with viewing the metrics values and recommendations in the single environment regarding the software source code during development. Besides that, it has been concluded that IDs metric should be included in the source code complexity metrics due to its great effect for the clean code, in which IDs is more important than the measure of comments ratio. Proposed *MLQA* is based on ***Bansiya*** model, which is succeeded in finding a mathematical relationship between quality attributes and software metrics in terms of design characteristics, however, it is weak in the normalization process in packages compare, therefore, the modification of the proposed work is to change normalization by using threshold values instead of using one of the packages as the base of normalization process so as to make this model more effective to track the development of software from one version to another. The aforementioned results have come out from the experiment conducted on three Java codes, which have been tested in this experiment to demonstrate the validity of the proposed tool results and their practicality in software analysis. For the future work, there is a possibility of applying artificial intelligence techniques in the field of building tools of computer aided software engineering in the field of quality models, especially the use of fuzzy logic in the field of selecting the appropriate threshold for the software metrics. Moreover, the possibility of incorporating the idea of adding a quality model as part of a development environment of programming languages will increase their adoption. Finally, extend the scope of the proposed quality tool to include more languages that support object oriented programming such as C ++, C #.

REFERENCE

[1] Bansiya, Jagdish.,"*A Hierarchical Model for Object-Oriented Design Quality Assessment*", IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 28, NO. 1. pp. 4-17, 2002.

[2] Mrinal, R. Singh. And Mittal , Arpita.,"*Survey on Impact of Software Metrics on Software Quality*", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 3, No. 1 ,pp. 137-141,2012.

[3] Lee, Alice T.," *Software Analysis Handbook: Software Complexity Analysis and Software Reliability Estimation and Prediction* ", Information Systems Journal, NASA Center for AeroSpace Information, *Available from ston.jsc.nasa.gov,1994.*

[4] Dromey, R. Geoff.,"*SOFTWARE PRODUCT QUALITY: Theory, Model, and Practice*". Software Quality Institute ,Griffith University, AUSTRALIA, 1998.

[5] L. Smith, *et al.*, "Guidelines for Successful Acquisition and Management of Software-Intensive Systems: Weapon Systems Command and Control Systems, Management Information Systems," *Department of the Air Force-Software Technology Support Center,(February),* 2003.

[6] R. Martin, "OO design quality metrics," *An analysis of dependencies,* vol. 12, pp. 151-170, 1994.

[7] Honglei, Tu., Wei, Sun., Yanan, Zhang., " *The Research on Software Metrics and Software Complexity Metrics* ", IEEE journals, International Forum on Computer Science-Technology and Applications, ISBN: 978076953930 Vol: 1 Pages: 131-136 , 2009.

[8] Martin, Robert C., " Clean Code - A Handbook of Agile Software Craftsmanship ", Published by Pearson Education, Inc, ISBN 0-13-235088-2, 2008.

[9] Martin, Robert C.," *The clean coder : a code of conduct for professional programmers* ", Pearson Education, Inc, ISBN-13: 978-0-13-708107-3, ISBN-10: 0-13-708107-3, 2011.

[10] Jetter, Andreas.,"*Assessing Software Quality Attributes with Source Code Metrics*", Diploma Thesis, Department of Informatics, University of Zurich, 2006.

[11] Panovski , Gregor., " *Product Software Quality* ", MASTER'S THESIS, Department of Mathematics and Computing Science, TECHNISCHE UNIVERSITEIT EINDHOVEN, 2008.

[12] Borchert, Thomas.," *Code Profiling: Static Code Analysis* ", MASTER'S THESIS, Department of Computer Science, Karlstad University, Sweden, 2008.

[13] Bhatti, Hassan R., "*Automatic Measurement of Source Code Complexity* ", MASTER'S THESIS, Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology,2010.

[14] Widheden, Göran. , " *Software Complexity: Measures and Measuring for Dependable Computer Systems* ", MASTER'S THESIS, Department of Computer Science and Engineering, University of Gothenburg, Goteborg, Sweden,2010.

[15] Chidamber, S. R., & Kemerer, C. F., " A metrics suite for object oriented design", *IEEE Transactions on software engineering*, *20*(6), 476-493,1994.

[16] Chandra, E., Linda, P. Edith," *Class Break Point Determination Using CK Metrics Thresholds* ", Global Journal of Computer Science and Technology, Vol.10 ,Issue.14, Pp:73-77,2010.

[17] Jamwal, Ranbireshwar S.,Jamwal, Deepshikha. & Padha, Devanand.," *Comparative Analysis of Different Software Quality Models*", Proceedings of the 3rd National Conference; INDIACom,Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi,2009.

[18] KITCHENHAM, BARBARA.,"*SOFTWARE QUALITY: THE ELUSIVE TARGET*", IEEE SOFTWARE, Vol. 13, No. 1: JANUARY 1996, pp. 12-21,1996.

[19] Wang, Su-Hua., Samadhiya, Durgesh. & Chen, Dengjie., "*Software Quality: Role and Value of Quality Models*", International Journal of Advancements in Computing Technology Volume 3, Number 6, , pp. 65-74,2011.

[20] Jamwal. Dr. Deepshikha.," *Analysis of Software Quality Models for Organizations*", International Journal of Latest Trends in Computing ,E-ISSN: 2045-5364,Volume 1, Issue 2, pp. 19-23,2010.

[21] Abdul, Jabbar. P, S.Sarala," *Advanced Program Complexity Metrics and Measurement* ", International Journal of Computer Applications (0975 – 8887), Val 23, No: 2, Pages: 29-33, 2011.

[22] Laird, Linda M.," *Software measurement and estimation: a practical approach* ", Published by John Wiley & Sons and IEEE Computer Society, ISBN 0-471-67622-5,2006.

[23] Herbold, S. ," *Calculation and Optimization of Thresholds for Sets of Software Metrics* ", Technical Report, Institute of Computer Science, University of Goettingen, Germany, No. IFI-TB-2010-01, ISSN 1611-104, 2011.

[24] Sastry, J.S.V.R.S.," *MEASURING OBJECT-ORIENTED SYSTEMS BASED ON THE EXPERIMENTAL ANALYSIS OF THE COMPLEXITY METRICS* ", International Journal of Engineering Science and Technology (IJEST), ISSN : 0975-5462 ,Vol. 3, No. 5 , Pages: 3726-3731, 2011.

[25] El-lateef, T. Abd., Yousef, A. H., "*Object Oriented Design Metrics Framework based on Code Extraction*", Computers and Systems Department, Faculty of Engineering, Ain Shams University,2006.

[26] Pressman, Roger S.," *SOFTWARE ENGINEERING: A PRACTITIONER'S APPROACH*", 8th EDITION, McGraw-Hill, ISBN 978-0-07-802212-8,2014.

[27] S. Misra, et al., "A suite of object oriented cognitive complexity metrics," IEEE Access, vol. 6, pp. 8782-8796, 2018.

[28] Bruntink, M., van Deursen, A., "*Predicting Class Testability using Object-Oriented Metrics*", Proceedings of the Fourth IEEE International Workshop on Source Code Analysis and Manipulation (SCAM'04). Pages: 136- 145, ISBN: 0-7695-2144-4,2004.

[29] Abran, A., "*SOFTWARE METRICS AND SOFTWARE METROLOGY*", IEEE Computer Society, Published by John Wiley & Sons, Inc., ISBN: 978-0-470-59720-0, 2010.

Using an Integrated Framework for Conceptual Modeling

Lindita Nebiu Hyseni, PhD
Faculty of Computer Sciences
Kadri Zeka University
Gjilan, Kosovo

Zamir Dika, PhD
Faculty of Contemporary Sciences and Technologies
South East European University
Tetovo, Macedonia

***Abstract*—The Integrated Framework for Conceptual Modeling (IFCMod) is created to contribute to the quality of the information system through the integration of the functional and non-functional requirements. This paper attempts to explore the outcomes of the IFCMod usage through the Mixed Method Case Studies at the Higher Education Institution and the Central Bank. The case study at the South East European University (SEEU) was the analysis and design of the improvement of the e-Schedule system, while the case study at the Central Bank of the Republic of Kosovo (CBK) was the analysis and design of the Data Collection System for Enterprise Surveys (DCSES). Based on the institutional perspective of the community participation during the semi-structure interviews, at the end of the Joint Approval Requirements (JAR) meetings, in both cases, the outcomes showed that IFCMod usage increases the quality of the information system by increasing quality in the system requirements.**

Keywords*—Integrated framework for conceptual modeling (IFCMod); joint approval requirements (JAR); system requirements; information system; mixed method case studies

I. INTRODUCTION

The conceptual modeling (CM) has a long tradition in the information system (IS) research [1]. The CM supports understanding and communication of the system requirements during the IS development [2], [28]. It has a significant role in the success of the IS [3-5] because of the understandability of the requirements by all project participants. Therefore, was created the Integrated Framework for Conceptual Modeling (IFCMod) in order to integrate the functional (FRs) and non-functional requirements (NFRs) [6]. After its creation, the research questions followed; was it possible to increase quality in the Information System by increasing quality in the System Requirements via IFCMod usage? Based on this research question, there were explored the outcomes of IFCMod usage during the system analysis and design stage through the following Mixed Methods Case Studies: analysis and designing of the improvement of the e-Schedule at the South East European University (SEEU), and analysis and design of the Data Collection System for Enterprise Surveys (DCSES) at the Central Bank of the Republic of Kosovo (CBK).

The most convenient research method to explore the outcomes of this complexity study was the Mixed Methods Case Study [16] because it allows collecting, analyzing, and integrating the qualitative and quantitative aspects of the study to answer the research question through interpretation of the significant results.

The aim of this paper was the interpretation of the significant results received from the institutional perspectives where the IFCMod was used to gather and to document requirements by the contractor, understand, review and approve the requirements by client in meeting with the contractor using the Joint Approval Requirements (JAR) method.

This paper was organized as follows: in the section 2 is a state of the art; in the section 3 are presented the materials and methods; in the section 4 are presented the outcomes from the IFCMod usage at the South East European University (SEEU) and at the Central Bank of the Republic of Kosovo (CBK); in the section 5 is discussion; in the section 6 is the research limitations; in the section 7 is the conclusion and future work; at the end, in the last section are presented the references used in this paper.

II. RELEATED WORK

Through the conducted research was identified the small number of scholars who had been working for the requirements integration, even if they have started to reinitiate this issue in the year 2016, they did not present the final solution per integration of the functional and non-functional requirements in one conceptual model. A solution that would help all project participants to understand the system requirements during the system analysis and design stage [2], [7].

Most of the scholars have been divided the nature of the system requirements into the functional (FRs) and non-functional requirements (NFRs) by representing them separately into the CM and requirements documents [8-12]. Cysneiros et al. (2001) [13], have proposed a framework for integrating the non-functional requirements into conceptual models, but their strategy was not transferable to real world situations because of missing adequate tools and methods for system development. Also, Cysneiros and Leite (2004) [14] has given another solution to the integration of the FRs and NFRs but the strategy that they used had several problems in the functional models and had no significantly impact on the overall development process. Moreover, the authors Eckhardt, Vogelsang, and Fernández (2016) [15] hint the possibility of the integration of the functional requirements (FRs) and non-functional requirements (NFRs) by handling most of the NFRs similar to the FRs. This hint was the main motivation for creation of the Integrated Framework for Conceptual Modeling (IFCMod) which consisted guides to deal with the requirements gathering, the creation of the functional and non-functional requirements document and the conceptual model of the information system. The IFCMod consist also the Joint Approval Requirements (JAR) methods which allow clients reviewing and approving of the requirements document and conceptual model which are created based on guides.

Based on conducted research, there were two techniques per requirements definition, documentation, validation and approval, such as Joint Requirements Planning (JRP) [26] and Joint Application Development (JAD) [27]. In comparison with the JAR method. These techniques differ in a few aspects that are presented in the following Table I.

TABLE I. COMPARASION OF THE JAR METHOD AND TECHNIQUES: JRP AND JAD

Aspects of the JAR method and techniques				
Joint Approval Requirements (JAR)	Functional and Non-Functional Requirements Document was created by contractor before meetings with the client. It is created based on gathering requirements from relevant materials of the institution.	Conceptual Model of Functional and Non-Functional Requirements was designed by contractor before meetings with the client. It was created based on gathering requirements from relevant materials of the institution.	Review & Approve of the prepared Requirements Document by client during the meetings with the contractor	Review & Approve of the prepared Conceptual Model by client during the meetings with the contractor
Joint Requirements Planning (JRP)	The formal written document was published immediately following the JRP session. The content of the formal document is depended on the objective of the JRP session.	It is created Prototypes by contractor during the JRP session only per Functional Requirements if have	Confirm Requirements by client during the JRP session	Approve Prototypes by client during the JRP session only per Functional Requirements in case they had so.
Joint Application Development (JAD)	It was required to develop a set of questions before starting the JAD sessions. The post-session report was created 2-3 weeks after JAD session	Not applied	All issues were discussed by client and contractor in the JAD Session and the needed information was collected	Not applied

The aspects of the techniques presented in the Table I, were also a motivation to add the Joint Approval Requirements (JAR) method as a component of the Integrated Framework for Conceptual Modeling which had the missing aspects in the JRP and JAD technique. Moreover, these missing aspects in the techniques have pushed the Higher Education Institution and the Central Bank using the JAR method and guides of the Integrated Framework for Conceptual Modeling during the system analysis and design. Before the JAR method, the Higher Education Institution has used the JRP technique, while the Central Bank has used the JAD technique during system analysis and design.

III. MATERIALS AND METHODS

In 2017, at the Higher Education Institution and at the Central Bank [16] was applied the Mixed Methods Case Study to explore the outcomes of the Integrated Framework for Conceptual Modeling (IFCMod) usage. While, the Sequential Exploratory Design (QUAL→ quan) [17-19], [28] was used for exploration because it gave priority to the qualitative aspects of the study. This design initiated the qualitative data collection and analysis phase than the qualitative components followed by the quantitative data collection and analysis phase with the aim of findings the generalization increasement.

The qualitative data were collected from semi-structured interviews held at the end of the JAR [19] meetings at the South East European University (SEEU) and at the Central Bank of the Republic of Kosovo (CBK). In this research study, the community participation from the SEEU were the Top Management, Managers, Employees, and Students, while from the CBK were the Top Management, Mangers, Employees and the representative of enterprises with foreign owned capital in Kosovo who reports to CBK. Moreover, in both institution these Joint Approval Requirements meetings were organized to review and approve the functional and non-functional requirements document (FRs & NFRs DOC) and the integrated conceptual model (ICM) which was created using the guides of the IFCMod.

The criteria for inclusion the above-mentioned community of the SEEU was that they should have been invited in the Quarantine of the year 2014 which was held for analyzing and designing of the e-Schedule system by using JRP technique or they should have used or developed that e-Schedule system in order to be invited in the JAR meetings of the year of 2017 for improvement of the e-Schedule system [19]. From this community who participate in the JAR meetings in the year 2017 was collected 15 qualitative semi-structured interviews. Of the 15 participants interviewed, 2 were from Top Management, 2 from IT Department, 6 from Faculties, 2 from the Human Resources Department, 1 from Student Services and 2 from Academic Planning.

Also, the CBK community was included in the research study based on criteria which they should have been involved in the process of data collection from the enterprises survey and should have experience in CBK (priority was the experience in JAD techniques), while the criteria for inclusion the representative of enterprises with foreign owned capital in Kosovo was that they should have been involved in the process of data reporting to CBK. From this community who participate in the JAR meetings in the year 2017 was collected 25 qualitative semi-structured interviews. Of the 25 participants interviewed, 2 were from Top Management, one of them from CBK and one from Enterprises; 7 from Manager positions, one of them from CBK and six from Enterprises; 16 from Employees, three of them from IT Departments.

In both institutions, the interview lasted per half hour and interview questions were focused on the effect of the IFCMod in the quality of the system requirements and the information system. During interviews, the ethical principle ‘Doing no harm’ was respected, where were developed and maintained the relationships of trust with participants in the Mixed Method Case Study. The transcribed interviews were transformed into categorical data which were classified into two binary classifications because it will unequivocally highlight the effect of the IFCMod based on participants` responses [20]. The categorical data were analyzed using SPSS in order to discover if there was a relationship between categorical variables. In this case, were used the Fisher`s Exact test because of small sample size [22-25]. The results of this test are presented in the section 4. During this research study was also treated the dilemma about the small number of sampling. According to the authors Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013) data saturation had occurred by 12 interviews while, theoretical saturation generally occurs between 10 and 30 interviews. Based on the authors, the single case study should generally contain 15 to 30 interviews [23]. Whilst, the authors Malterud, K., Siersma, V. D., & Guassora, A. D. (2016) highlighted that the information power indicates the lower number of participants in the interviews. It depends on the aim of the study, sample specificity, use of established theory, quality of dialogue, and analysis strategy [24]. Furthermore, the authors Ghazi, A. N., Petersen, K., Reddy, S. S. V. R., & Nekkanti, H. (2017) based on literature review highlighted that this is a problem in the field of software engineering [25]. In the following sections, the outcomes of the Mixed Method Case Studies should be presented.

IV. Results of using IFCMod

The research question; was it possible to increase quality in the Information System by increasing quality in the System Requirements via IFCMod usage? was addressed through the Mixed Methods Case Study at the South East European University (SEEU) and at the Central Bank of the Republic of Kosovo. In the following subsections are presented the interpretations of the significant results received from the above-mentioned institutions perspectives where the IFCMod is used [16], [18], [28].

A. *Improvement of the E-Schedule System*

The outcomes of using Integrated Framework for Conceptual Modeling (IFCMod) at the South East European University (SEEU) during the analysis and design of the improvement of the e-Schedule system were explored through Mixed Methods Case Study to gain a better understanding of IFCMod effect based on the perspectives of the community participation during the semi-structure interviews, at the end of the JAR meetings [21].

Considering the focus of the semi-structure interview, which was the effect of the IFCMod in the quality of the system requirements (SRs) and the information system (IS), variable about this effect was analyzed for its association with the other variables, this is shown in the Table II.

TABLE II. Results of using IFCMod at the SEEU

Fisher's Exact Test		
Variable association	***The effect of the IFCMod in the quality of the System Requirements (SRs) and the Information System (IS)***	***The understandability of the requirement document from all participants in the JAR meetings was used as a control variable to assess the relationship between variables***
1) Understandability of the Integrated Conceptual Model (ICM)	P value: **.011**	P value: **.011**
2) FRs & NFRs DOC and ICM help the system developer during the system development	P value: **.011**	P value: **.011**
3) Facilitation of the communication through Functional Requirements and Non-Functional Requirements Document (FRs & NFRs DOC) and Integrated Conceptual Model (ICM)	P value: **.033**	P value: **.033**
4) Review and Approve of FRs & NFRs DOC and ICM from all participants in the JAR meetings	P value: **.033**	P value: **.033**

Based on the outcomes from the perspective of the SEE University, the increasing of the quality in system requirements and information system depended on variables listed on the Table II. The interpretation of the significant results received from the institutional perspectives per variables are presented in the following.

1) Understandability of the Integrated Conceptual Model (ICM)

Based on the results presented in the Table II, this study found that the effect of the IFCMod in the quality of the System Requirements (SRs) and Information System (IS) was significantly and statistically associated with the understandability of the Integrated Conceptual Model. This was shown through the result of p value which is p=0.011. Also, the understandability of the requirement document from all participants in the JAR meetings was used as a *control variable* to assess the relationship between above variables. Based on the Fisher`s Exact test, this variable controlled the relationship of the other variables based on the statistically significant results p=0.011.

2) Facilitation of the communication through Functional Requirements and Non-Functional Requirements Document (FRs & NFRs DOC) and Integrated Conceptual Model (ICM)

In the Table II are presented the results of this study which found that the effect of the IFCMod in the quality of the SRs and IS, was significantly and statistically associated with facilitation of the communication through Functional Requirements and Non-Functional Requirements Document (FRs & NFRs DOC) and Integrated Conceptual Model (ICM). This was shown through the result of p value which is p=0.033. Also, the understandability of the requirement document from all participants in the JAR meetings was used as a *control variable* to assess the relationship between above variables. Based on the Fisher`s Exact test, this variable controlled the relationship of the other variables based on the statistically significant results p=0.033.

3) FRs & NFRs DOC and ICM help the system developer during the system development

The results of this study presented in the Table II found that the effect of the IFCMod in the quality of the SRs and IS, was significantly and statistically associated with how much FRs & NFRs DOC and ICM help the system developer during the system development. This was shown through the result of p value which is p=0.011. Also, the understandability of the requirement document from all participants in the JAR meetings were used as a *control variable* to assess the relationship between above variables. Based on the Fisher`s Exact test, this variable controlled the relationship of the other variables based on the statistically significant results p=0.011.

4) Review and Approve of FRs & NFRs DOC and ICM from all participants in the JAR meetings

As it is shown in the Table II, this study found that the effect of the IFCMod in the quality of the SRs and IS, was significantly and statistically associated with Review and Approve of FRs & NFRs DOC and ICM from all participants in the JAR meetings. This was shown through the result of p value which is p=0.033. Also, the understandability of the requirement document from all participants in the JAR meetings was used as a *control variable* to assess the relationship between above variables. Based on the Fisher`s Exact test, this variable controlled the relationship of the other variables based on the statistically significant results p=0.033.

B. Analysis and Design of the Data Collection System for Enterprise Surveys (DCSES)

Throughout the research study at the Central Bank of the Republic of Kosovo (CBK), was used the Integrated Framework for Conceptual Modeling (IFCMod) to analyze and design the Data Collection System for Enterprise Surveys (DCSES). The intention of this study was to explore the outcomes of IFCMod usage through Mixed Methods Case Study in order to gain a better understanding of its effect based on the perspectives of the community participation during the semi-structure interviews at the end of the JAR meetings [21].

Considering the focus of the semi-structure interview, which was the effect of the IFCMod in the quality of the system requirements (SRs) and the information system (IS), variable about this effect was analyzed for its association with the other variables, this is shown in the Table III.

TABLE III. RESULTS OF USING IFCMOD AT THE CBK

Fisher's Exact Test		
Variable association	***The effect of the IFCMod in the quality of the System Requirements (SRs) and the Information System (IS)***	***The understandability of the requirement document from all participants in the JAR meetings was used as a control variable to assess the relationship between variables***
Understandability of the Integrated Conceptual Model (ICM)	P value: **.022**	P value: **.022**
FRs & NFRs DOC and ICM help the system developer during the system development	P value: **.018**	P value: **.022**
Review and Approve of FRs & NFRs DOC and ICM from all participants in the JAR meetings	P value: **.018**	P value: **.010**

In Table III, are listed the variables in which are depended on the increase of the quality in system requirements and information system based on the outcomes from the perspective of the CBK. The interpretation of the significant results received from the institutional perspectives per variables is presented in the following.

1) Understandability of the Integrated Conceptual Model (ICM)

Based on the results presented in the Table III, this study found that the effect of the IFCMod in the quality of the SRs and IS was significantly and statistically associated with the understandability of the Integrated Conceptual Model. This was shown through the result of p value which is p=0.022. Also, the understandability of the requirement document from all participants in the JAR meetings was used as a *control variable* to assess the relationship between above variables. Based on the Fisher`s Exact test, this variable controlled the relationship of the other variables based on the statistically significant results p=0.022.

2) FRs & NFRs DOC and ICM help the system developer during the system development

The results of this study presented in the Table III found that the effect of the IFCMod in the quality of the SRs and IS, was significantly and statistically associated with the amount of FRs & NFRs DOC and ICM helping the system developer during the system development. This was shown through the result of p value which is p =0.018 which meant there was a statistically significant association between effect of the IFCMod in the quality of the SRs and IS, and the amount of FRs & NFRs DOC and ICM helping the system developer during the development of the DCSES system. Also, the understandability of the requirement document from all participants in the JAR meetings was used as a *control variable* to assess the relationship between above variables. Based on the Fisher`s Exact test, this variable controlled the relationship of the other variables based on the statistically significant results p=0.022.

3) Review and Approve of FRs & NFRs DOC and ICM from all participants in the JAR meetings

Based on the results in the Table III, this study found that the effect of the IFCMod in the quality of the SRs and IS was significantly and statistically associated with Review and Approve of FRs & NFRs DOC and ICM from all participants in the JAR meetings. This was shown through the result of p value which is p=0.018. Also, the understandability of the requirement document from all participants in the JAR meetings was used as a *control variable* to assess the relationship between above variables. Based on the Fisher`s Exact test, this variable controlled the relationship of the other variables based on the statistically significant results p=0.010.

V. Discussion

The IFCMod is used through Mixed Method Case Study at the Higher Education Institution and at the Central Bank in order to gain better understanding of its effect in the quality of the system requirements (SRs) and the information system (IS) considering the perspectives of the community participation during the semi-structure interviews at the end of the JAR meetings in these institutions. Unexpected finding is the weakness of the IFCMod which is that only participants with the experience in system analysis and design stage agreed that the IFCMod facilitated the communication during system development. While, the strength of the IFCMod based on finding is the understandability of the functional and non-functional requirement document and conceptual model by client and contractor side, and the application of the JAR Method to approve and review requirement document and the conceptual model by the client side in the JAR meetings with the contractor.

VI. Research Limitations

The research limitation was the lack of similar work in the integration of the functional and non-functional requirements. Moreover, the sample size from both institutions was dictated from the nature being studied [23-24], [29-31]. The participants in the study are selected based on the criteria listed in Section 3, in order to provide the best data [32] therefore the sample size is acceptable [24].

VII. Conclusions and Future Work

The Integrated Framework for Conceptual Modeling (IFCMod) was created and used at the South East European University (SEEU) during the improvements of the e-Schedule system. Also, the IFCMod was used at the Central Bank of the Republic of Kosovo during the analysis and design of the Data Collection System for Enterprise Surveys (DCSES). Both usages were done through Mixed Methods Case Studies because of the complexity of the research study [16]. Based on the outcomes from the institutional perspectives of the SEEU and CBK, understandability of the Functional and Non-Functional Requirements Document (FRs & NFRs DOC) and the Integrated Conceptual Model (ICM) from all participants in the Joint Approval Requirements meetings increased the quality in the System Requirements (SRs). Also, the review and approve of the FRs & NFRs DOC and the ICM during the JAR meetings of the client and the contractor increase the quality of the SRs. Both institutions are agreed that the IFCMod help system developer during the system development, while only SEEU is agreed that it facilitate the communication between the contractor and the client during the system development. The IFCMod is ready to be used in different industries in order to explore the outcomes of its usage and compare them with the outcomes of this research paper.

The future work would be a creation of the new component of the IFCMod, the Price Model (PM). The PM shall present the way of calculation of total cost per information system development based on Functional and Non-Functional Requirements Document (FRs & NFRs DOC) and the Integrated Conceptual Model (ICM).

References

[1] Borgida, Alexander. "Conceptual modeling of information systems." In On Knowledge Base Management Systems, pp. 461-469. Springer, New York, NY, 1986.

[2] Hyseni, Lindita Nebiu, and Zamir Dika. "Integrated Approach to Conceptual Modeling." International Journal of Advanced Computer Science and Applications 7, no. 12 (2016): 213-219.

[3] Bera, Palash, Anna Krasnoperova, and Yair Wand. "Using ontology languages for conceptual modeling." Journal of Database Management (JDM) 21, no. 1 (2010): 1-28.

[4] Braga, Bernardo FB, João Paulo Andrade Almeida, Giancarlo Guizzardi, and Alessander B. Benevides. "Transforming OntoUML into Alloy: towards conceptual model validation using a lightweight formal method." Innovations in Systems and Software Engineering 6, no. 1-2 (2010): 55-63.

[5] Mehmood, Kashif, Samira Si-Said Cherfi, and Isabelle Comyn-Wattiau. "CM-Quality: a pattern-based method and tool for conceptual modeling evaluation and improvement." In East European Conference on Advances in Databases and Information Systems, pp. 406-420. Springer, Berlin, Heidelberg, 2010.

[6] Hyseni, Lindita Nebiu, and Zamir Dika. "An integrated framework of conceptual modeling for performance improvement of the information systems." In Innovative Computing Technology (INTECH), 2017 Seventh International Conference on, pp. 174-180. IEEE, 2017.

[7] Robinson, Stewart. "Tutorial: Choosing what to model—Conceptual modeling for simulation." In Simulation Conference (WSC), Proceedings of the 2012 Winter, pp. 1-12. IEEE, 2012.

[8] Sommerville, Ian, and Pete Sawyer. Requirements engineering: a good practice guide. John Wiley & Sons, Inc., 1997.

[9] Gorschek, Tony, and Claes Wohlin. "Requirements abstraction model." Requirements Engineering 11, no. 1 (2006): 79-101.

[10] Fatwanto, Agung, and Clive Boughton. "Analysis, specification and modeling of non-functional requirements for translative model-driven development." In Computational Intelligence and Security, 2008. CIS'08. International Conference on, vol. 2, pp. 405-410. IEEE, 2008.

[11] Fatwanto, Agung, and Clive Boughton. "Analysis, Specification and Modeling of Functional Requirements for Translative Model-Driven Development." In Knowledge Acquisition and Modeling, 2008. KAM'08. International Symposium on, pp. 859-863. IEEE, 2008.

[12] Sommerville, I., & Sawyer, P. "Requirements engineering (10th Edition)", Pearson, 2015.

[13] Cysneiros, Luiz Marcio, Julio Cesar Sampaio do Prado Leite, and Jaime de Melo Sabat Neto. "A framework for integrating non-functional requirements into conceptual models." Requirements Engineering 6, no. 2 (2001): 97-115.

[14] Cysneiros, Luiz Marcio, and Julio Cesar Sampaio do Prado Leite. "Nonfunctional requirements: From elicitation to conceptual models." IEEE transactions on Software engineering 30, no. 5 (2004): 328-350.

[15] Eckhardt, Jonas, Andreas Vogelsang, and Daniel Méndez Fernández. "Are" Non-functional" Requirements really Non-functional? An Investigation of Non-functional Requirements in Practice." In Software Engineering (ICSE), 2016 IEEE/ACM 38th International Conference on, pp. 832-842. IEEE, 2016.

[16] Morse, Janice M. Mixed method design: Principles and procedures. Routledge, 2016.

[17] Peng, Guo Chao, J. M. B. Nunes, and Fenio Annansingh. "Investigating information systems with mixed-methods research." In Proceedings of the IADIS International Workshop on Information Systems Research Trends, Approaches and Methodologies. Sheffield, 2011.

[18] Cameron, Roslyn. "A sequential mixed model research design: Design, analytical and display issues." International Journal of Multiple Research Approaches 3, no. 2 (2009): 140-152.

[19] Hyseni, Lindita Nebiu, and Zamir Dika. "Method for System Requirements Approval." INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 8, no. 7 (2017): 331-336.

[20] Hashi, Iraj, and Besnik A. Krasniqi. "Entrepreneurship and SME growth: evidence from advanced and laggard transition economies." International Journal of Entrepreneurial Behavior & Research 17, no. 5 (2011): 456-487.

[21] Myers, Michael D., and Michael Newman. "The qualitative interview in IS research: Examining the craft." Information and organization 17, no. 1 (2007): 2-26.

[22] Freeman, Jenny V., and Steven A. Julious. "The analysis of categorical data." Scope 16, no. 1 (2007): 18-21.

[23] Marshall, Bryan, Peter Cardon, Amit Poddar, and Renee Fontenot. "Does sample size matter in qualitative research?: A review of qualitative interviews in IS research." Journal of Computer Information Systems 54, no. 1 (2013): 11-22.

[24] Malterud, Kirsti, Volkert Dirk Siersma, and Ann Dorrit Guassora. "Sample size in qualitative interview studies: guided by information power." Qualitative health research 26, no. 13 (2016): 1753-1760.

[25] Ghazi, Ahmad Nauman, Kai Petersen, Sri Sai Vijay Raj Reddy, and Harini Nekkanti. "Survey research in software engineering: problems and strategies." arXiv preprint arXiv:1704.01090 (2017).

[26] Bentley, Lonnie D., and Jeffrey L. Whitten. Systems analysis and design methods. Irwin/McGraw Hill, pp. 229-235. 2007.

[27] Dennis, Alan, Barbara Haley Wixom, and Roberta M. Roth. “Systems analysis and design.” John wiley & sons, pp. 119-143. 2012. Aramo-Immonen, Heli. "Mixed methods research design." In World Summit on Knowledge Society, pp. 32-43. Springer, Berlin, Heidelberg, 2011.

[28] Lakhoua, M. N., and F. Khanchel. "Overview of the methods of modeling and analyzing for the medical framework." Scientific Research and Essays 6, no. 19 (2011): 3942-3948.

[29] Morse, Janice M. "Determining sample size." (2000): 3-5.

[30] Sobal, Jeffery. "Sample extensiveness in qualitative nutrition education research." Journal of Nutrition Education 33, no. 4 (2001): 184-192.

[31] Thomson, S. Bruce. "Sample size and grounded theory." (2010).

[32] Strauss, A., and J. Corbin. "Basics of Qualitative Research. Sage Publications." Thousand Oaks, CA (1998).

Development of Mobile Health Application for Cardiovascular Disease Prevention

Vitri Tundjungsari[1], Heri Yugaswara[3], Adhika Trisna Dwi Putra[4]
Faculty of Information Technology
YARSI University
Jl. Letjen Suprapto, Jakarta, Indonesia

Abdul Salam M Sofro[2]
Faculty of Medicine
YARSI University
Jl. Letjen Suprapto, Jakarta, Indonesia

***Abstract*—Cardiovascular diseases are one of major cause of death in the world, as well as in Indonesia. In spite of that fact, cardiovascular diseases (CVDs) could be prevented with healthy behavior and lifestyle, such as: regular health check-up, healthy eating and drinking lifestyle, stress management, sleeping management, regular physical activities. In this paper, we develop mobile health application as a tool to record daily behavior and lifestyle. Mobile health is chosen because nowadays mobile devices are the most popular communication used among people. Thus, we believe that mobile Health (mHealth) is a promising tool to promote healthy lifestyle and behavior. The method we used for developing the application is called as Human-centered design (HCD). The application is evaluated iteratively from the first prototype (low-fidelity prototype) to the final prototype (high-fidelity prototype). Based on the feedback using User Experience Questionnaire (UEQ) shows that the application has above average scores for all of the components, i.e.: attractiveness, clarity, efficiency, accuracy and dependability, stimulation, and novelty. The best score is for Stimulation (Excellent), while the worst score is accuracy and dependability (above average). This shows that mHealth is a potential tool to stimulate users for having healthy lifestyle, however it still required further validation of use from health experts to ensure the accuracy's result of the application.**

***Keywords*—Mobile health; cardiovascular; disease; human-centered design; standard; user-centered design**

I. Introduction

Cardiovascular diseases (CVDs) are one of chronic diseases which cause highest mortality in the world (around 31% people died from CVDs which 85% of it died due to heart attack and stroke) [1]. WHO sheet fact [1] also mentions that over three quarters of CVD deaths are occurred in low and middle-income countries, including Indonesia. According to Indonesian basic health research report [2], cardiovascular diseases are the leading cause of death in Indonesia (31%), which top two including hypertension and stroke.

WHO estimates that the number will increase until 2030, which estimated 23.6 million people will die from heart and blood vessel disease [1]. Cardiovascular diseases defined as a group of disorders of the heart and blood vessels, which include [1]:

- Coronary heart disease: disease caused by the blood vessels to provide blood for the heart muscle
- Cerebrovascular disease: disease caused by the blood vessels to provide blood for the brain
- Peripheral arterial disease: disease caused by the blood vessels to distribute blood supply in the arms and legs
- Rheumatic heart disease: disease caused by damage to the heart muscle and heart valves from streptococcal bacteria
- Congenital heart disease: disease caused by malformations of heart structure at birth
- Deep vein thrombosis and pulmonary embolism: disease caused by blood clots in the leg veins that able to dislocate and travel to the heart and lungs

Cardiovascular diseases can be prevented by having healthy behaviour and life style, such as: no smoking, avoid obesity, increase physical activities. Therefore it is important to foster such initiatives through mobile health (mHealth) application. Mobile application is a very promising tool to promote sustained and successful healthy behavioural lifestyle [3].

MHealth is also efficient and affordable tool to support and deliver information and education among people since almost individual in this world are having mobile communication, such as: smartphone and tablet [4, 5].

Based on the background above, we develop a mobile application as a tool to reduce and prevent cardiovascular risk by promoting healthy life style through mobile healthcare application. We believe that this research can provide an alternate tool to decrease cardiovascular disease prevalence in Indonesia, as well as to prevent it.

This paper consists of five main sections, i.e.: (1) introduction, (2) research related to mobile health application for CVDs management, (3) human-centered design method, (4) results and discussion, and (5) conclusion.

II. Mobile Health Application for Cardiovascular

WHO has defined mobile health (mHealth) technology as a medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants, and other wireless devices [6]. MHealth is using mobile technology to support health projects and outcomes [7]. Mobile technology has different types, such as: mobile phones/smartphones, tablets, or personal digital assistants (PDA). Keisling [7] mentions several categories of mHealth, i.e.:

- Financial transactions and incentives: this category provides services for any payments and insurances, such as: money transfer, health insurance, health payments, health incentives based on performance, and saving accounts.
- Information systems: this category supports any activities for data collection and reporting, such as: statistics, household surveys, health surveillance, electronic health records.
- Service delivery and support: this category includes any system refer to decision support system, diagnosis of certain diseases, disease management and prevention, communication between providers, referrals, and telemedicine.
- Social and behavior change communication: this category offers service such as reminders, health education, and health promotion.
- Supply management: this category involves various services such as cold chain management, commodity tracking and replenishment, and stock management.
- Workforce development and performance support: this category enables feedback for certain services such as service quality, human resource management.

Honeyman et al. [8] illustrate the potential integration and components of a generic mobile health system (figure 1). The figure shows that mobile can support and deliver healthcare in various forms, such as: voice/video calls, SMS/messages, mobile applications, multimedia, inbuilt sensors, and device connectivity. Figure 2 shows the services that can be delivered by mHealth.

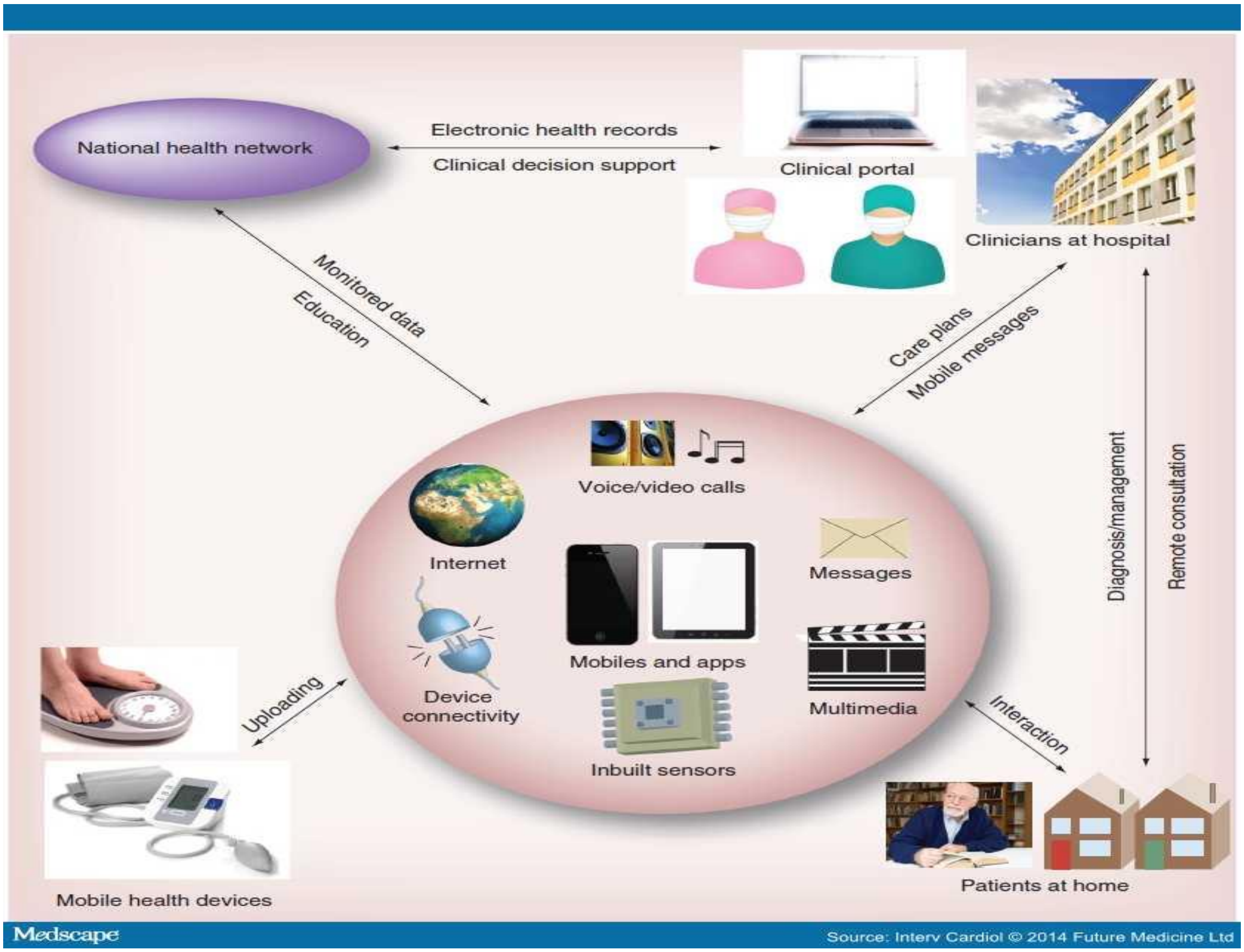


Fig. 1. The potential integration and components of a generic mobile health system [8].

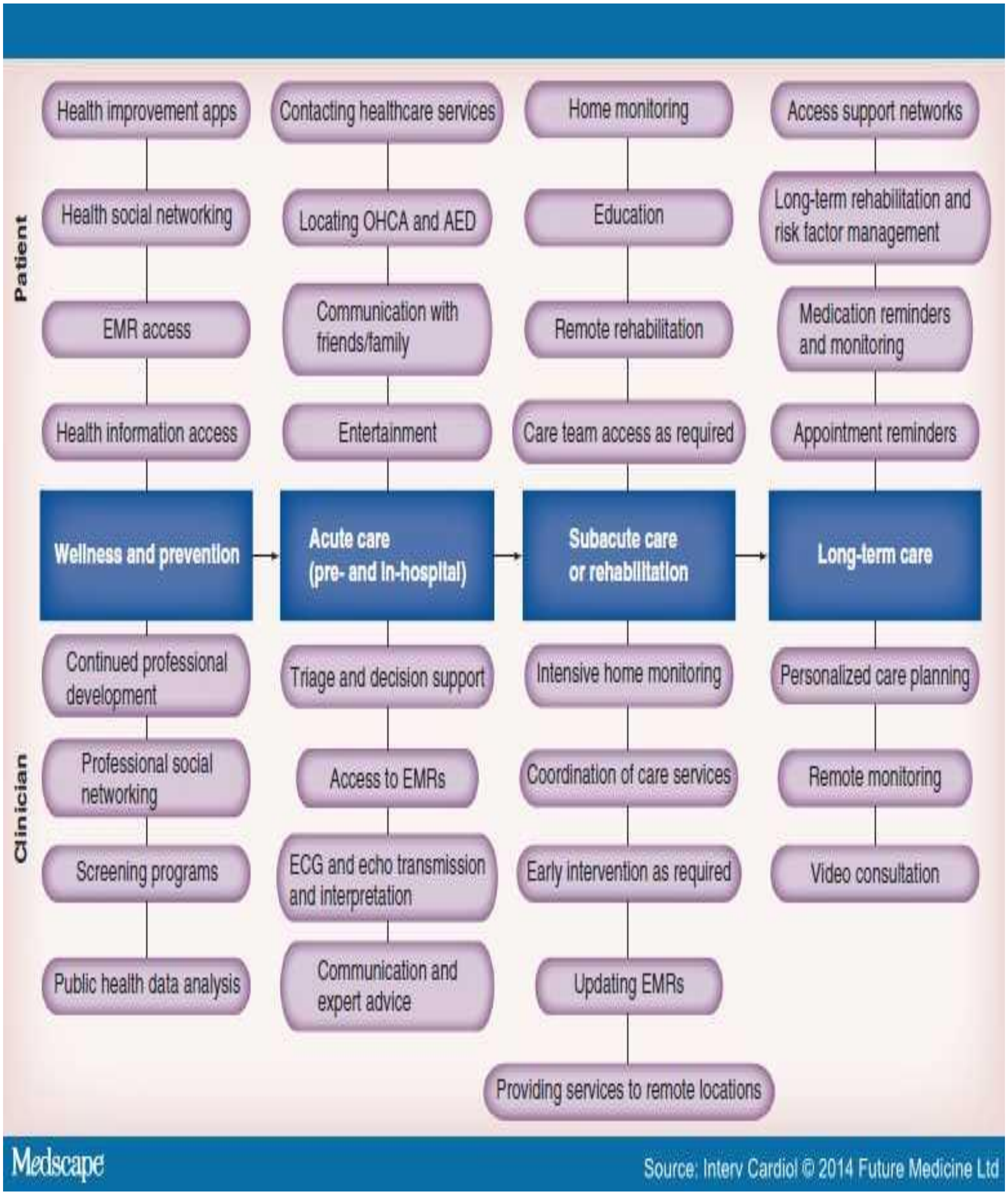


Fig. 2. Potential applications of mobile technologies by patients and clinicians at each stage of the care journey[8].

There are several research have been done to investigate the potential use of mHealth for Cardiovascular disease management [3, 4, 5, 7, 8, 9, 10, 11, 12, 13]. The details explanations for those related research are as follows:

- Authors in [3] investigate how mHealth can control lifestyle behavior as prevention to cardiovascular diseases. In their research, Eapen et al [3] also mention that evolution of mHealth is still in the early stages, thus it requires more thorough roadmap by involving all stakeholders, such as: patients, developers, providers, and payers. By having collaboration of all stakeholders, the developers will be able to comprehend the real problem from different perspectives.
- Chow et al [4] describe the ability of mHealth as a tool for cardiovascular education and prevention, in terms of cardiac rehabilitation. However, it should be performed further investigation to ensure that mHealth is secure, safe, and robust.
- Feinberg et al [9] find out that mHealth has potential use as cardiovascular management, prevention, and

health promotion by conducting case study in Kerala, India. The research has objectives to explore mobile phone usage pattern in rural Kerala (Ernakulam) and to explore acceptability of mHealth delivery of health promotion and cardiovascular prevention.

- Honeyman et al [8] explain detail usage of mHealth for several cardiovascular diseases, such as: cardiac arrest, arrhythmias, myocardial infarction, heart failure, and interventional cardiology. Mhealth offers significant roles to improve cardiac care from both a patient and a clinical perspective. In their research, they highlight some of the key cardiac interventions where mobile health has been applied or studied in both the acute and longer-term care settings.
- Xie et al [10] mention that mHealth can be used for promoting cardiovascular disease self-management. However in their research, they find thath Chinese cardiovascular apps are insufficient to provide comprehensive health information and interactive functions to facilitate cardiovascular self-management. They also mention that mHealth can offer efficient and affordable solution for China, since China has a large population with cardiovascular disease that requires self-management.
- Treskes et al [11] discuss several use of mHealth for remote monitoring cardiovascular patients. New technologies in smartphones have the potential to be used for remote monitoring. However, they find out some weaknesses of mHealth, such as: untrustworthiness of mobile technology, insufficient regulation and poor compensation for mHealt implementation.
- Pfaeffli et al [12] mention mHealth as promising tool as exercise tool for cardiac rehabilitation patients. The exercise is developed using mobile tool with patient input using the following steps: conceptualization, formative research, pre-testing, and pilot testing. The research find out that mHealth developed in the research was effective as an exercise program for cardiac rehabilitation patients.
- Gandapur et al [5] explain that mHealth can be used to improve medication adherence for cardiovascular disease. The aim of their research is to assess mHealth as a tool to provide better medication obedience for cardiovascular disease patients, especially patients with hypertension, coronary artery disease, heart failure, peripheral arterial disease, and stroke. Their research discovers that mHealth is able to improve medication adherence in patients with cardiovascular diseases.
- Tundjungsari et al [13] compare several mobile applications with ability to calculate cardiovascular risk factors. The aim of their research is to investigate the critical success factor of Clinical Decision Support Systems application, mainly for calculating cardiovascular risk. They conduct usability testing toward three different mHealth applications by involving participants from different types of backgrounds (physicians, IT developers, and students). The result of the research indicates that knowing the target user's needs is very critical in the design process.

Table below summarizes those literatures related to mHealth use for cardiovascular disease management.

TABLE I. RELATED RESEARCH OF MHEALTH USAGE FOR CVD MANAGEMENT

Author	Research result
Eapen et al (2016)	mHealth as control lifestyle behaviour for CVD prevention
Chow et al (2016)	mHealth as a tool of CVD prevention, cardiac rehabilitation and education
Feinberg et al (2017)	Suitability and acceptability of mHealth for CVD management, prevention, and health promotion in Kerala
Honeyman et al (2014)	mHealth as a tool of monitoring for Cardiac arrest, arrhythmiac, myocardial infarction, heart failure, interventional cardiology
Xie et al (2017)	mHealth as tool for CVD self-management
Treskes et al (2016)	mHealth as tool for remote monitoring CVD
Pfaeffli et al (2012)	mHealth as tool for cardiac rehabilitation exercise intervention
Gandapur et al (2016)	mHealth as tool for improving medication adherence CVD
Tundjungsari et al (2017)	mHealth as a tool for calculating cardiovascular risk

III. HUMAN CENTERED DESIGN FOR INTERACTIVE SYSTEMS

The implementation of the system was carried out using a principle of Human or User Centered Design (HCD) for interactive systems standard [14]. This standard is used as user-centred interaction design process, also referred as 'the HCD standard'. The standard is based on 'BS EN ISO 9241-210:2010 Ergonomics of human-system interaction. It consists of activities focussing on human-centred design, i.e.:

1) *Understand and specify context of use*
2) *Specify user requirements*
3) *Produce design solutions to meet these requirements*
4) *Evaluate design against requirements*

The above stages also can be summarised as: (1) analysis, (2) specification, (3) design and (4) evaluation. The stages in the HCD standard have to be carried out iteratively, by having evaluation stage. Figure 3 shows how the HCD standard is implemented in a project. The project using HCD standard should involve the user from the beginning of the project until the last stage of development. By involving user from every stages of HCD standard, it will provide clear usability measurement.

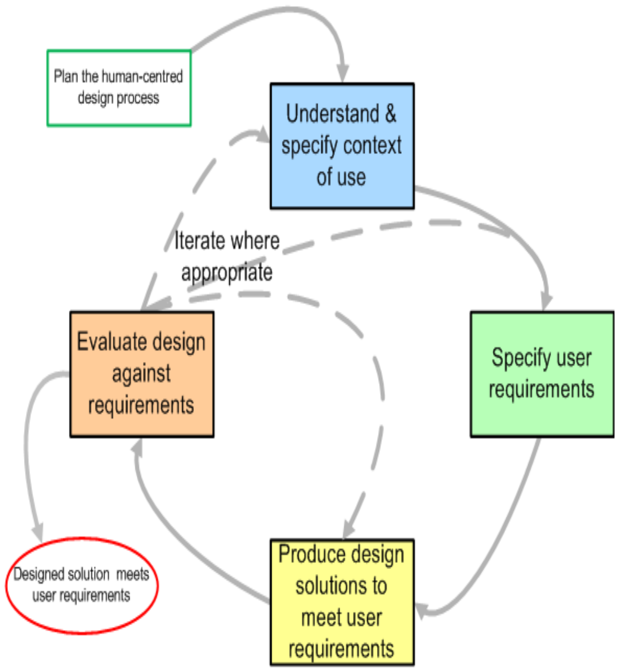


Fig. 3. Human-centered design standard [14].

In this research we perform step by step HCD standards by involving: patients, public, physicians, public health experts, developers, faculty members, and usability experts. The details of HCD process in this project is explained in table below:

TABLE II. PROCESSES AND ACTIVITIES IN HCD STANDARD

Stage	Activities
HCD 1: Understand and specify context of use	• Application concept • Features lists • Task Scenarios lists • Finding reliable source from Minister of Health Republic of Indonesia
HCD 2: Specify user requirements	Requirements details for each task (task, sub-task, significance, and frequency)
HCD 3: Produce design solutions to meet these requirements	Low-fidelity prototype production
HCD 4: Evaluate design against requirements	Evaluation design and high-fidelity prototype production

IV. RESULT AND DISCUSSION

A. Understand and Specify Context of use

Application Concept: This application is a mobile based application with function to prevent cardiovascular disease by promoting healthy behavior and life style. The components of healthy behavior and lifestyle are provided from Ministry of Health Republic of Indonesia. There are six main components which determine the healthy behavior and lifestyle as prevention to cardiovascular diseases, i.e.: (1) encourages the public to carry out regular health checks, (2) eliminate cigarette smoke, (3) diligently engage in physical activity, (4) a healthy and balanced diet, (5) adequate rest and manage stress.

Based on those concepts, we propose several features to be developed in this application, i.e.:

1) Recording and calculating daily sleeping time

2) Recording daily eating time (breakfast, lunch, dinner)

3) Recording and calculating daily physical activities/exercise time

4) Recording and calculating daily amount of water consumption

5) Recording and calculating stress level

6) Recording and calculating daily number of cigarettes consumption (for smoker)

B. Specify User Requirements

Feature 1: Recording and calculating daily sleeping time

Task group	Recording sleeping time
Sub tasks	Save sleeping time Save wake up time
Significance	Calculate duration of sleeping time (sleeping time to wake up time)
Frequency of use	Daily

Feature 2: Recording and calculating daily eating time (breakfast, lunch, dinner)

Task group	Recording eating time (breakfast, lunch, dinner)
Sub tasks	Save breakfast time Save lunch time Save dinner time Save fruit and vegetable consumption
Significance	Calculate average daily eating habit (time and ingredients)
Frequency of use	Daily

Feature 3: Recording and calculating daily physical activities/exercise time

Task group	Recording physical activity/exercise time
Sub tasks	Calculating daily steps done Calculate daily distance steps done
Significance	Calculate average daily physical activity (steps and distance)
Frequency of use	Daily

Feature 4: Recording and calculating daily amount of water consumption

Task group	Recording amount of water consumption
Sub tasks	Save amount of water (glasses) Save caffeine consumption (if any) Save alcohol consumption (if any)
Significance	Calculate average daily drinking habit (amount and ingredients)
Frequency of use	Daily

Feature 5: Recording and calculating stress level

Task group	Recording stress level
Sub tasks	Save stress incidence (if any) Save stress level (0-100)
Significance	Calculate average daily stress level
Frequency of use	Daily

Feature 6: Recording and calculating daily number of cigarettes consumption (for smoker)

Task group	Recording smoking behavior
Sub tasks	Save smoking active and passive behavior (if any)
Significance	Calculate average daily smoking behavior
Frequency of use	Daily

C. Produce design solutions to Meet these Requirements

The first prototype of the system was designed and produced as low-fidelity prototype, as it can be seen in figure 4. The first prototype is used to show the navigation flow and overall design based on the user requirement (stage 2 of HCD).

D. Evaluate Design Against Requirements

The final prototype was designed based on the feedback gathered in the previous prototypes. This final prototype is a fully functional mobile application. Some of the final prototypes are shown in figure 5, 6, 7, and 8.

To evaluate the prototype we use User Experience Questionnaire (UEQ). UEQ is used to measure: (1) attractiveness (overall impression of the application. Do users like or dislike it?); (2) clarity (does the context of application give clarity? Does the interface element give clarity?); (3) efficiency (can users solve their tasks without unnecessary effort? Does it react fast?); (4) accuracy and dependability? (does the application provide accurate result? Does the user feel in control of the interaction? Is it secure and predictable?); (5) stimulation (is it exciting and motivating to use the product? Is it fun to use?); (6) novelty (is the design of the product creative? Does it catch the interest of users?).

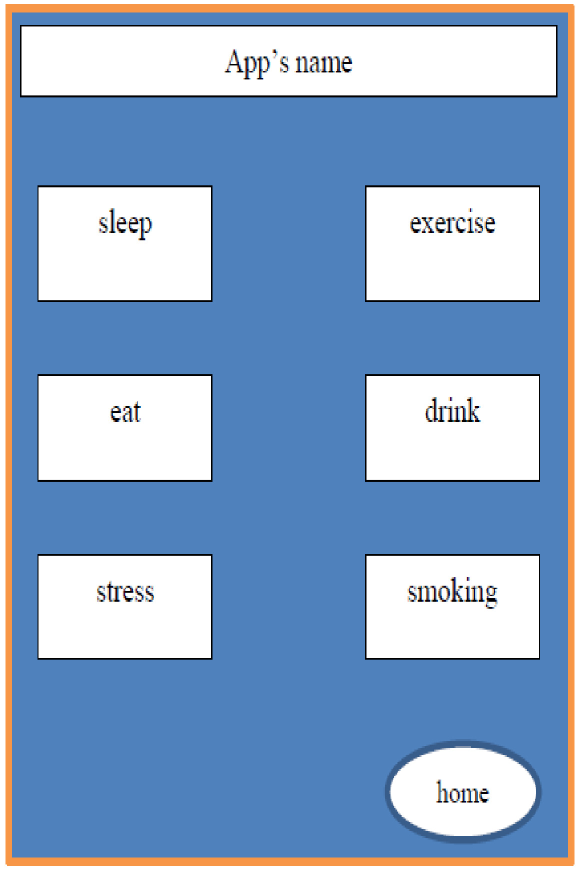


Fig. 4. Low-Fidelity Prototype.

Fig. 5. Interface of Final Prototype (1).

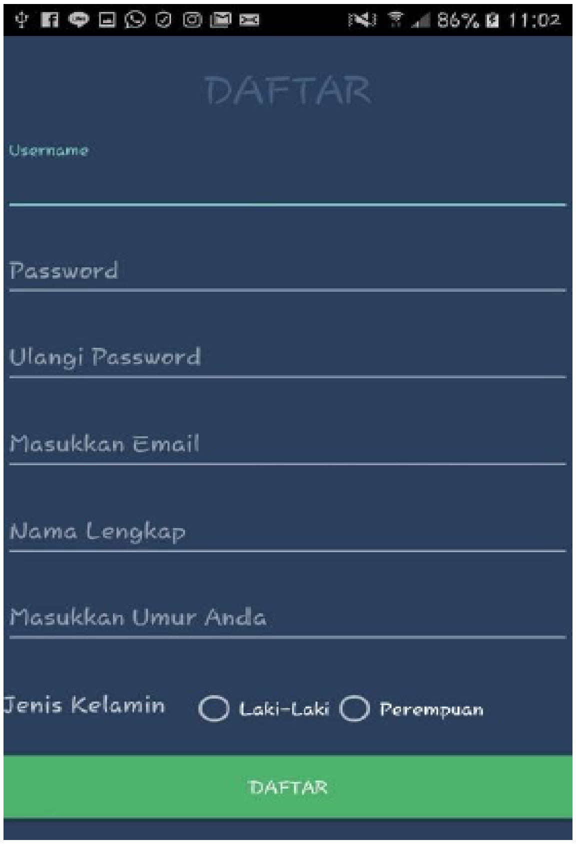


Fig. 6. Interface of final prototype (2).

Fig. 7. Interface of final prototype (3).

Fig. 8. Interface of final prototype (4).

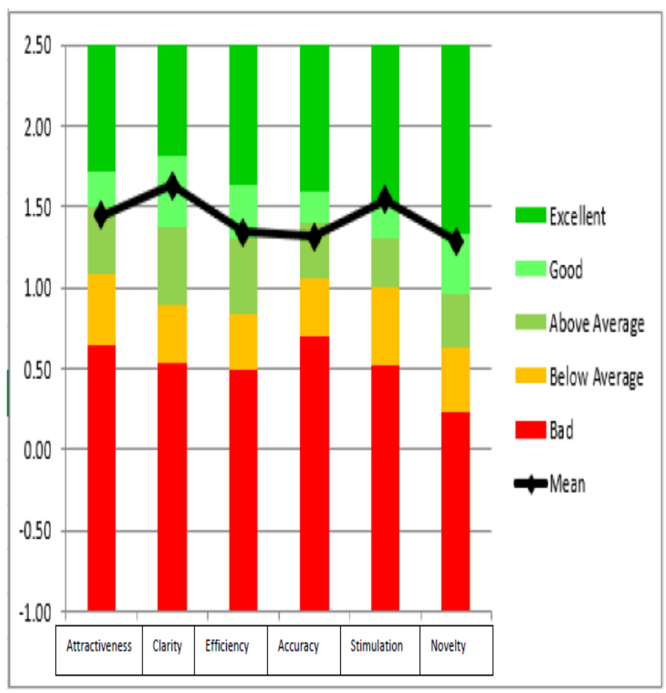


Fig. 9. Application evaluation result using UEQ.

We evaluate the final prototype using User Experience Questionnaire (UEQ) as an instrument. There are 30 respondents with various ages (15-64 yearsold) involved in the assessment. They have to experience the mHealth application and fill in the UEQ based on their experience of using the mHealth application. Figure 9 shows the UEQ test result. The best result is achieved by component **stimulation** with 'excellent' score, followed by **clarity** with 'good' score. However the highest average score is performed by **clarity** component, followed by **stimulation**. On the other hand, the worst result is **accuracy** having score of 'above average'. Overall UEQ result shows that all of the components have score 'above average' which indicate that the application perform 'good' experience to users.

V. Conclusion

In this research, a development of mHealth application as a tool to promote health behaviour and life style is performed. The application is developed using stages determined by HCD for interactive systems standard. The evaluation result obtained shows that the application able to stimulate the respondent as users to use the application. This proves that mHealth is a potential tool to promote healthy behaviour and life style, and in turn as an enabler tool for preventing cardiovascular diseases. However, the accuracy result is not as high as other components. This indicates that there are still many doubts occurred from the respondents regarding the accuracy of the application. It means that we still need more approval from health and medical expert to validate the accuracy of the application.

In the future, we have to evaluate the application in real environment and involving collaboration between health care professionals and app developers, in order to achieve better end users' preferences and more accurate result.

Acknowledgment

This work was funded by the Ministry of Research Technology and Higher Education Republic of Indonesia, under PDUPT research schema. We would like to thank also to Universitas YARSI for its support during the research study.

References

[1] https://www.who.int/cardiovascular_diseases/en/

[2] http://www.depkes.go.id/resources/download/general/Hasil%20Riskesdas%202013.pdf

[3] Z.J. Eapen, M.P. Turakhia, M.V. Mcconnell, G. Graham, P. Dunn, C. Tiner, C. Rich, R.A. Harrington, E.D. Peterson, P. Wayte, 2016,"Defining A Mobile Health Roadmap For Cardiovascular Health And Disease", Journal Am Heart Assoc, 2016 Jul, 5 (7).

[4] C.K. Chow, N. Ariyarathna, S.M. SharifulIslam, A. Thiagalingam, J. Redfern, 2016, "mHealth in Cardiovascular Health Care", Heart, Lung, and Circulation, Volume 25, Issue 8, August 2016.

[5] Y. Gandapur, S. Kianoush, H.M. Kelli, S. Misra, B. Urrea, M.J Blaha, G. Graham, F.A. Marvel, S.S. Martin, 2016, "The Role of mHealth for Improving Medication Adherence in Patients with Cardiovascular Disease: A Systematic Review", European Heart Journal – Quality of Care and Clinical Outcomes (2016) 2, 237–244.

[6] WHO Library Cataloguing in Publication Data, 2011, mHealth: New horizons for health through mobile technologies: second global survey on eHealth.

[7] K. Keisling, 2014, Introduction to mHealth: How to Approach mHealth.mHelp org August 2014.

[8] E. Honeyman, H. Ding, M. Varnfield, M. Karunanithi, 2014, "Mobile Health Applications in Cardiac Care", Interv Cardiol 6 (2).

[9] L. Feinberg, J. Menon, R. Smith, J.G. Rajeev, R.K. Kumar, A. Banerjee, 2017, "Potential for Mobile Health (mHealth) Prevention of Cardiovascular Diseases in Kerala: A Population-based Survey", Indian Heart Journal 69 (2017) 182–199.

[10] B. Xie, Z. Su, W. Zhang, R. Cai, "Chinese Cardiovascular Disease Mobile App' Information Types, Information Quality, and Interactive Functions for Self-Management: Systematic Review", JMIR MHealth and UHealth.

[11] R.W. Treskes, E.T. Van der Velde, R. Barendse, N. Bruining, 2016, "Mobile Health in Cardiology: A Review of Currently Available Medical Apps and Equipment for Remote Monitoring", Expert Review of Medical Devices.

[12] L. Pfaeffli, R. Maddison, R. Whitaker, R. Stewart, A. Kerr, Y. Jiang, G. Kira, K. Carter, L.Dalleck, 2012, "A MHealth Cardiac Rehabilitation Exercise Intervention: Findings from Content Development Studies", BMC Cardiovascular Disorders 2012 12:36.

[13] V. Tundjungsari, A.S.M. Sofro, A. Sabiq, A. Kardiana, 2017, "Investigating Clinical Decision Support Systems Success Factors with Usability Testing", International Journal of Advanced Computer Science and Applications (IJACSA), Volume 8 Issue 11, 2017.

[14] R. Harte, L. Glynn, A. Rodríguez-Molinero, P.M. Baker, T.Scharf, L.R. Quinlan, G. ÓLaighin, 2017, "A Human-Centered Design Methodology to Enhance the Usability, Human Factors, and User Experience of Connected Health Systems: A Three-Phase Methodology", JMIR human factors, Volume 4, No 1.

Image Processing based Task Allocation for Autonomous Multi Rotor Unmanned Aerial Vehicles

Akif Durdu*, Mehmet Celalettin Ergene, Onur Demircan, Hasan Uguz, Mustafa Mahmutoglu, Ender Kurnaz
Robotics, Automation and Control Laboratory (RAC-LAB),
Department of Electrical and Electronics Engineering, Konya Technical University
Konya, Turkey

***Abstract*—Nowadays studies based on unmanned aerial vehicles draws attention. Especially image processing based tasks are quite important. In this study, several tasks were performed based on the autonomous flight, image processing and load drop capabilities of the Unmanned Aerial Vehicle (UAV). Two main tasks were tested with an autonomous UAV, and the performance of the whole system was measured according to the duration and the methods of the image processing. In the first mission, the UAV flew over a 4x4 sized color matrix. 16 tiles of the matrix had three main colors, and the pattern was changed three times. The UAV was sent to the matrix, recognized 48 colors of the matrix and returned to the launch position autonomously. The second mission was to test load drop and image processing abilities of the UAV. In this mission, the UAV flew over the matrix, read the pattern and went to the parachute drop area. After that, the load was dropped according to the recognized pattern by the UAV and then came back to the launch position.**

Keywords*—UAV; multi rotor; quad rotor; image processing; search and rescue; task allocation

I. Introduction

Nowadays Unmanned aerial vehicle (UAV) technology is becoming very popular. It is finding a place in many sectors and studies. This technology is attracting attention to the defense industry, but the usage of UAVs is rapidly increasing every day in civil-life.

From firefighting to search and rescue, agriculture to photography, in many studies UAVs have been used for a long time. So it can be said that the field of use of UAVs is vast. For example, while Karaca et al. were researching in order to localize and rescue the victims on mountainous areas [1], Zhou et al. were studying on real-time video registration for forest fire monitoring [2]. In [1] and [2] UAVs were used for video capturing and data transmission. However, the technology kept evolving. Some researchers made a significant improvement on the mechanical design of the UAV in order to increase the performance of the search and rescue [3], and some of them changed the camera type with an infrared one and increased the performance of the fire detection [4]. In next years, a group of researchers changed their point of view [5]. Getting closer to recent days, Yuan et al. developed an algorithm to detect and track forest fire [6]. Giving another example of an image processing application for UAVs, Moranduzzo et al. added machine learning to image processing and used support vector machine (SVM) in order to detect cars from UAV images [7].

UAVs are used not only for such applications but also for transportation. In their study, Goodchild et al. wrote a paper estimating the carbon dioxide emission of the vehicles while delivering goods and reducing this emission rate by using UAVs for the delivery [8]. Another interesting paper is written by Zheng et al. simulating and evaluating the safety of a long distance search and rescue mission executed by a UAV [9].

Considering this amount of applications and the importance of image based tasks on UAVs, in this paper, we have allocated tasks to a UAV in order to measure the performance of its autonomous flight, image processing ability and autonomously load drop capability. These features are considered as the essential features for a smart UAV. There were two missions to evaluate the performance of the UAV. In the first mission, the UAV flew over a 4×4 sized color matrix. 16 tiles of the matrix had three primary colors, and the pattern was changed three times. The UAV recognized 48 colors of the matrix and returned to the launch position autonomously. The second mission was to test load drop and image processing abilities of the UAV. In this mission UAV flew over the matrix read the pattern and went to the parachute drop area. Three loads with parachutes were dropped according to the recognized pattern of the matrix, and then the UAV returned to the launch position.

More information about the missions and the detailed design of the UAV is given in the methodology part. The output of the missions and the performance of the UAV are given in conclusion.

II. Methodology

A. Mission I

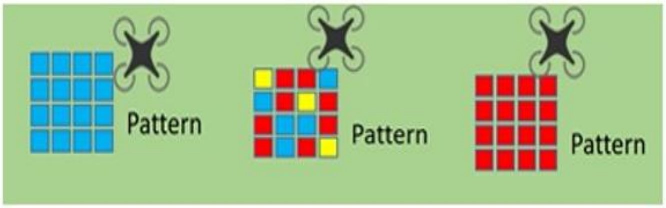
In this mission, autonomous flight and image processing capabilities of the UAV is tested. UAV has a system to fly over a 4×4 color matrix, which contains 1m×1m sized color panels, detect and record the matrix colors during the flight. In this mission, color detection and recording will be done three times, and every time the matrix color changes colors will be given randomly to the matrix. Panel colors will be stable for 10 seconds, and there will be 5 seconds waiting time between color changes. Fig. 1 shows the color changes in the first mission.

B. Mission II

In this mission, autonomous flight, image processing and parachute dropping in the correct order capabilities of the UAV

* Corresponding author

are tested. UAV flew over a 4×4 color matrix contains 1m×1m sized color panels. As it can be seen in Fig. 2, there are three different colored rows among which every row's panels have the same color. The last row will have random colors. According to the first three rows' colors, the parachutes are dropped with 5 seconds gaps. The mission area is shown below.



(a) (b) (c)

Fig. 1. Demonstration of the Color Matrix.

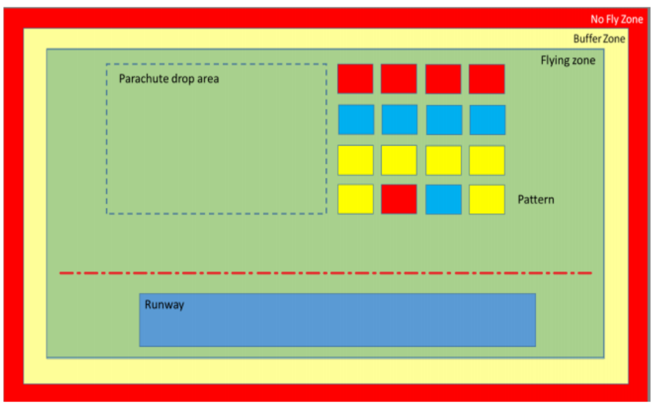


Fig. 2. The Mission Area.

C. Autonomous System

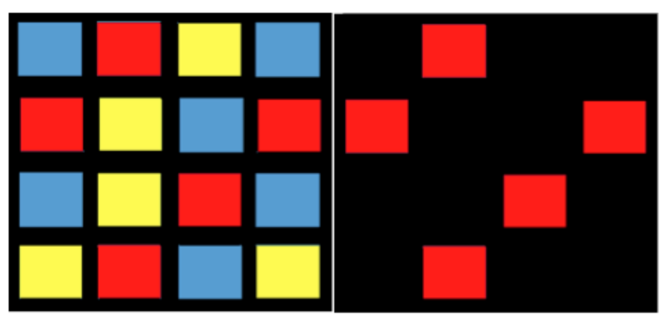
A flight controller board (Pixhawk 2.1) was chosen in order to accomplish the mission and to control its process, Mavlink Protocol is used. MAVLink (Micro Air Vehicle Link) is a protocol for communicating with small unmanned aerial vehicles. It is designed as a header-only message sequence library [10].

The software to control the autonomous flight was written in Linux based using Drone-kit API in Raspberry Pi 3. The algorithm of the software is shown in Fig. 3.

D. Image Processing Algorithm

Several image processing algorithms are used in order to perform both missions. In the first mission which is to recognize the given patterns, the algorithm waits for the UAV to fly on the pattern. After the UAV arrives at the given position on the pattern, image processing algorithms start. After capturing the first frame of the pattern preprocessing is applied, the image is converted to the HSV color space [11]. The main reason for the usage of HSV color-space is to detect color under any circumstances independently of light intensity and brightness. The median filter is used to get rid of the noise in the image. H (hue), S (saturation) and V (value) value range for every color are determined so that the background will be subtracted and every color will be detected separately (Fig. 4.).

Fig. 3. The Algorithm of Autonomous Flight Software.



(a) (b)

Fig. 4. (a) Background Subtraction from the Pattern (b) Detecting Red Color.

After detecting every color midpoints of every square, coordinates are written in an array, and that array is sent via RX-TX ports into the SD card. The output of the system can be seen in Fig. 5.

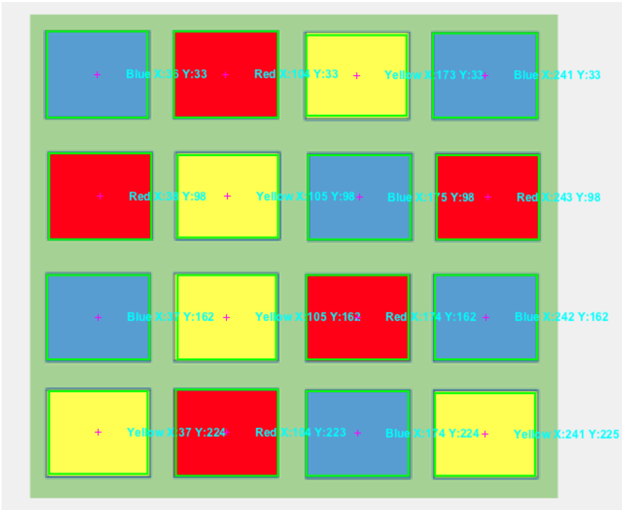


Fig. 5. The Output of the Color Detection Algorithm.

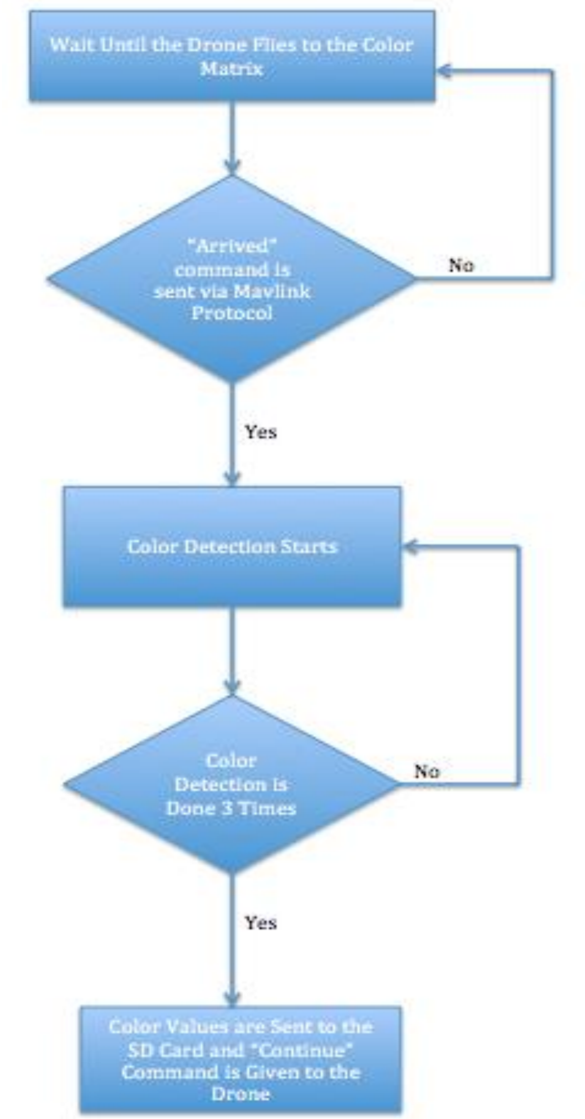


Fig. 6. The Algorithm of the First Mission.

The same algorithm for the second mission is also used. After getting the colors of the pattern first three rows' colors are read, according to those colors the related command is sent to the processor to drop the parachutes in the correct order.

The first mission is planned to be completed in totally 50.7 seconds. When necessary calculations are made the UAV speed is estimated about 58km/h, which makes 16.11m/s, and the flight height is planned about 10-18m. With this speed and the height, the way going to the color matrix and coming back would take 10.7 seconds. For every color change on the matrix, the process speed of the processor to detect the colors is estimated about a maximum of 10 seconds. The algorithm for the first mission is given in Fig. 6.

The second mission, when the flight speed and the height are considered, is estimated about 23.67 seconds. For color detection, the minimum process time is aimed, so the color detection and the parachute-drop order planning time is estimated about minimum 5 seconds. Due to weather conditions, all of these calculations can be changed. That is why estimation is used when planning the mission performance. With the best estimation, the best performance is aimed. The algorithm for the second mission is given in Fig. 7.

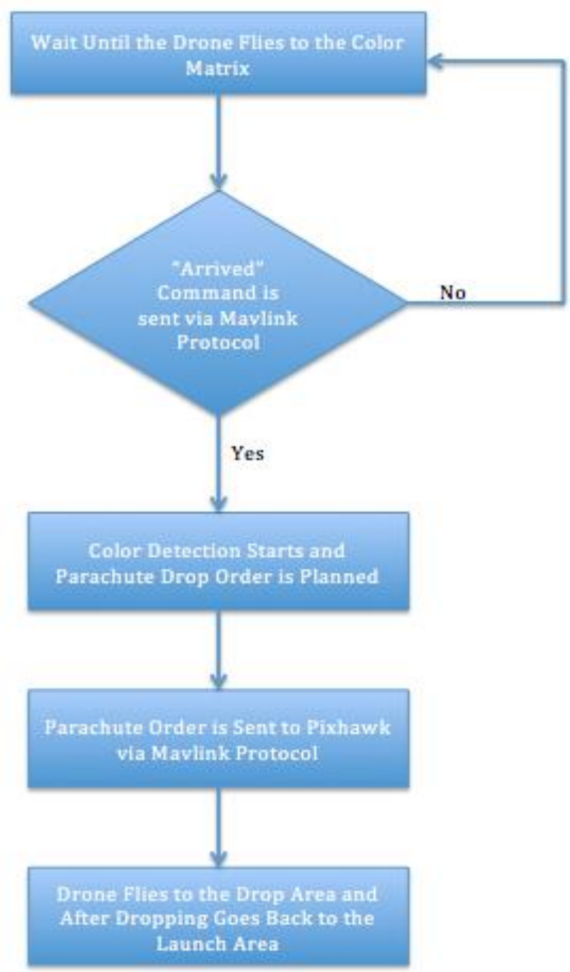


Fig. 7. The Algorithm of the Second Mission.

III. Experimental Setup

The lightest and the fastest UAV design for the missions is chosen. The designed UAV’s measurements are 231×231×88.40 mm. Considering motor and propeller features, the cross motor span is designed 280 mm. The gaps between vertical and horizontal motors are 200 mm. GPS is not included in these measurements. Technical drawing of the UAV is given in Fig. 8. The final design of the UAV is given in Fig. 9.

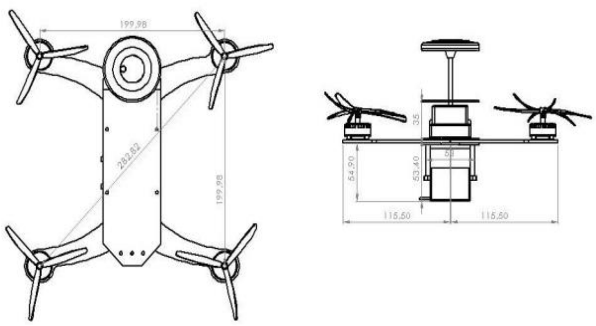


Fig. 8. The Technical Drawing of the UAV.

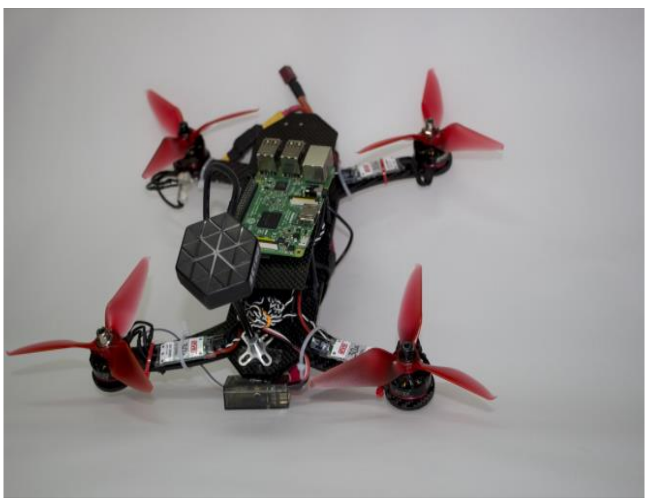


Fig. 9. Final Design of the UAV.

As mentioned before for image processing tasks, Raspberry Pi 3 is used and for obtaining images Pi Camera which is a compatible camera with Raspberry Pi cards. For autonomous flight, an autonomous flight card called The Cube Pixhawk 2.1 is used. Pixhawk 2.1 includes 32 bit STM32F427, 256 KB RAM, 32 bit STM32F103, two inertial measurement units (IMU), barometer, 3 axis gyro, accelerometer and magnetometer.

IV. Experimental Results

On the first mission, the flight time was planned to be 50,3 seconds. Real-time performance of the mission was 41 seconds. Image processing performance is shown below. The recognition of the first pattern is shown in Fig. 10. Only one mistake in 16 squares was made.

Fig. 10. Recognition of the First Pattern.

In Fig. 11, the second pattern is shown, and the recognition rate of the pattern is %100.

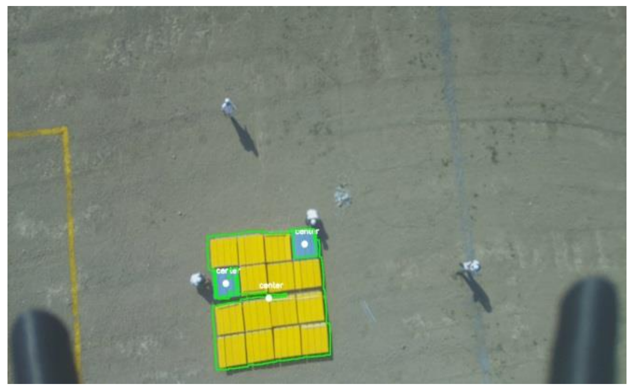


Fig. 11. Recognition of the Second Pattern.

In Fig. 12, recognition of the third pattern was done with 100% success rate.

Fig. 12. Recognition of the Third Pattern.

In the second mission, the flight time was estimated about 23.67 seconds. The real-time flight time is calculated for 26 seconds. During the flight, a small GPS problem occurred, and the UAV could not hover precisely on the pattern. However, the recognition performance of the pattern was 100% successful. In Fig. 13, the pattern is shown, and the first three rows were recognized correctly. After the recognition, the parachutes were dropped in the correct order (Fig. 14.).

Fig. 13. Pattern Recognition of the Second Mission.

Fig. 14. Load Drop in the Second Mission.

V. CONCLUSIONS

During flights, although some problems occurred, the missions were completed successfully. In the first pattern of the first mission, one square could not be recognized due to the angle of the camera and the light. In order to prevent this situation, some machine learning algorithms can be used. For example, the pixels of the image can be clustered with K-Means algorithm into few clusters so that the colors of the image would have been clustered into a few primary colors [12]. In this way, recognition can be done in many different environments. In this study, the UAV is designed for the missions, and we believe this study can lead to many different studies such as firefighting, search, and rescue, surveillance applications. With the addition of deep learning and other machine learning algorithms, image processing performance can be increased, and the area of study can also be widened this drone could be beneficial for the defense industry. Authors believe that UAVs will have a massive part in our life and this study shows the advantages of UAVs in a significant way.

ACKNOWLEDGMENT

Authors are thankful to RAC-LAB (www.rac-lab.com) for providing the trial version of their commercial software and to Robosys Mekatronik Sistemleri San. Tic. Ltd. Şti. for providing the necessary equipment for this study.

REFERENCES

[1] Y. Karaca , M. Cicek, O. Tatli, A. Sahin, S. Pasli, M.F. Beser, S. Turedi, “The potential use of unmanned aircraft systems (drones) in mountain search and rescue operations”, The American Journal of Emergency Medicine, vol. 36(4), pp. 583-588, April 2018.

[2] G. Zhou, L. Chaokui, and P. Cheng, “Unmanned aerial vehicle (UAV) real-time video registration for forest fire monitoring”, Proceedings of 2005 IEEE International Geoscience and Remote Sensing Symposium (IGARSS'05), Seoul, South Korea, 29-29 July 2005.

[3] Y. Naidoo, R. Stopforth, and G. Bright, “Development of a UAV for search & rescue applications”, IEEE AFRICON’11, Livingstone, Zambia, 13-15 Sept. 2011.

[4] L. Merino, F. Caballero, J.R. Martinez-de Dios, A. Ollero, “Cooperative fire detection using unmanned aerial vehicles”, Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005), Barcelona, Spain, 18-22 April 2005.

[5] S. A. Çeltek, A. Durdu and E. Kurnaz, “Design and Simulation of the Hierarchical Tree Topology Based Wireless Drone Networks”, 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), IEEE, Malatya, Turkiye, 28-30 Sept. 2018.

[6] C. Yuan, Z. Liu, and Y. Zhang, “UAV-based forest fire detection and tracking using image processing techniques”, 2015 International Conference on Unmanned Aircraft Systems (ICUAS), IEEE, Denver, CO, USA, 9-12 June 2015.

[7] T. Moranduzzo, and F. Melgani, “A SIFT-SVM method for detecting cars in UAV images”, 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22-27 July 2012.

[8] A. Goodchild, and J. Toy, “Delivery by drone: An evaluation of unmanned aerial vehicle technology in reducing CO2 emissions in the delivery service industry”, Transportation Research Part D: Transport and Environment, vol. 61, Part A, pp. 58-67, June 2018.

[9] L. Zheng, J. Hu, and S. Xu, "Marine Search and Rescue of UAV in Long-Distance Security Modeling Simulation", Polish Maritime Research, vol. 24, pp. 192-199, 2017.

[10] J. A. Marty, “Vulnerability analysis of the mavlink protocol for command and control of unmanned aircraft”, Thesis, Air Force Institute of Technology, Air University, 2014.

[11] M. C. Ergene, A. Durdu, and H. Cetin, “Imitation and learning of human hand gesture tasks of the 3D printed robotic hand by using artificial neural networks”, 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), IEEE, Ploiesti, Romania, 30 June-2 July 2016.

[12] M. C. Ergene, A. Durdu, “Robotic hand grasping of objects classified by using support vector machine and bag of visual words”, 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), IEEE, Malatya, Turkey, 16-17 Sept. 2017.

Towards Adaptive user Interfaces for Mobile-Phone in Smart World

Muhammad Waseem Iqbal[1], Nadeem Ahmad[2], Syed Khuram Shahzad[3], Irum Feroz[4], Natash Ali Mian[5]
[1,2,3,4]Department of Computer Science & IT
The Superior College (University Campus)
Lahore, Pakistan
[5]Beaconhouse National University
Lahore, Pakistan

***Abstract*—All applications are developed for context adaptation and provide communication with users through their interfaces. These applications offer new opportunities for developers as well as users by collecting context data and adapting systems behavior accordingly. Particularly, in mobile devices, these mechanisms provide usability increment tremendously. Rigid and non-adaptive interface blocks the features of context awareness. In this paper, we study methods, technologies and criteria which have been proposed specifically for adaptive interfaces. Based on these guidelines, we elaborate the intelligence of adaptivity and usage of context according to user mental model. Further, we have proposed a model to develop user context ontology (UCO) and adaptive interface ontology (AIO) to optimize the use of adaptive mobile interfaces in the context of user preferences. These ontologies organize the perceptions and thoughts of user. The philosophy of User Centered Design (UCD) is proposed to analyze the usability and validity of mobile device interfaces according to user contexts.**

***Keywords*—Adaptive features; smart-phone; usability experience, user interface; user context; usability engineering; UCD**

I. Introduction

The mobile-phone interfaces provide contented and dependable communication medium between the device and the user. Currently, mobile phones present interfaces having various styles and modes of interaction. These interfaces have different usability requirements and measures, that make the process of interaction design more complex [1]. The competitive environment has used the burst of technical options in interaction design [2] in which interface developers and end users face more difficulties in understanding of hardware and software technical details. Computer interfaces are emerging to be more sovereign in functionality, software systems are getting more complex, and online information spaces are growing rigorously in size. With tremendous growth in technical support, the variant usage patterns are also emerging [3]. There is now immense number of new users, who are not technology expert, but rather naïve to computing devices such as non-technical professionals, elderly people, and children [4]. These users have not only diverse computer skills, but they vary in many other aspects like their knowledge, skills, intellectual and physical capabilities, mood, motivation and most importantly the target tasks and usage of technology.

Numerous high-quality applications introduced but miss-out from the market [5] due to their complex, unattractive, inefficient and confusing user interfaces. The impact of non-adaptive user interface for mobile devices creates frustration and directly effects on performance, usability and reception amongst the users. During the designing of mobile phone interfaces [6], the user requirements should be evaluated in terms of usability, learnability, understandability, effectiveness, efficiency and objectivity. Latest mobile operating systems like Android or iOS provide kids, guest, driving and night modes for the accessibility of applications according to the user's task. These interaction modes are provided for specialized context with variation in values of common features like daylight, age groups and access styles. These modes provide pre-defined and static profiles [7] with factory settings. The user context study and analysis are still missing in the interface settings for pre-defined interaction modes. User is bound to select the most suitable yet rigid modes in any other context.

Adaptive User Interfaces (AUIs) can provide a lot of benefits to address these usability concerns. User interface adaptation has been identified, quite a while ago, as an imperative subject to address in modern information systems design. There have been three adaption design techniques for three major aspects of adaption. The information adaptation selects information to present. The presentation adaptation states the presentation styles of this information. The last interface adaptation defines the interaction mode and style [8]. With these questions answered at design level, the adaptation of mobile user interfaces provides big support for the user satisfaction.

The heterogeneity in user physical demographic properties and limitations of smart device interaction style brings us a challenge to develop a specialized interface for variety of users [9] [10]. Many users face problems in customization panels. These setting dialogues are very difficult to understand in formation especially for users with disabilities and lesser ICT knowledge and skills. Similarly, the need for customization is a significant requirement demanded by children or elderly people. Hence, customization of any device that targets accessibility must include auto-adaptation and self-learning mechanisms for user's requirements. Bad user interfaces and usability disorientation may cause annoyance and could lead to unsatisfied behavior amongst users [11]. For effective mobile interfaces the intended tasks should be mapped with user's mental model [8] [12]. Currently, the benefits of AUI have been realized to develop the interfaces in user's context. The AUI is a feasible adaptation approach than adaptable, because it provides suitable methodology of adaptation and it can handle the usability issues of user's interaction [13] [14]. The adaptive user interfaces are proposed as solution to cater the problems that enable any mobile application to provide dynamically customized interface [5] for different groups of users having similar properties and needs. The user's context-based interfaces help to make the collaborative, supportive, constructive and communicative activities easy.

The research idea under discussion shows the User Centered Design (UCD) approach to add adaptivity in mobile device interfaces. Therefore, the implementation of mobile adaptive user interfaces is very necessary for user's learnability, efficiency and satisfaction. The adaptive interfaces, designed according to the user requirements will help to enhance the user interaction with mobile devices.

II. Adaptivity and Adaptability in Mobile-Phone User Interface

The two approaches such as adaptivity and adaptability are used to personalization for user interfaces of mobile devices. The aim of both interfaces is to provide personalization for the users while these two approaches are different in the adaptation process [15]. Recently, the adaptivity and adaptability have achieved high popularity on the world wide web (WWW) under the notion of personalization [16]. The reason of this popularity is due to the less homogeneity of website users than the users of commercial software's.

In adaptive approach, the interface automatically adjusts and assists the user. The adaptive user interfaces can adapt their activities by monitoring user status, the state of system and the current situation according to adaptation strategy. Usually the intensity of adaptation is measured in the case of effectiveness, efficiency and satisfaction for user interface [4] [17]. The factors such as spatial stability which increases the user satisfaction and high locality, improves discoverability of adaptation in representing interface in its original position. Spatial stability is required to maintain the mental model of application. Moreover, accuracy provides the user perception for algorithm predictability, whereas the user interfaces with higher accuracy provide more predictability and consistency. Further, the interaction frequency and task complexity play an important role in the perception of adaptation [6]. If there is a need of large mechanical interactions with simple tasks [18], the adaptation locality plays important role for complex tasks. The users are able to create mental models for applications to interact frequently with low complexity.

Conversely, the adaptable interfaces provide mechanisms of customization but rely on the adaptation of user mechanisms. In adaptability, the user's preferences and characteristics are known before creating interaction [19]. The information system can be adopted manually by the user or administrator or automatically by the system to fulfill the requirements of users. AVANTI system provided adaptivity and adaptability features within the user interface. This system provides special input and output devices, visual/non-visual interface objects and integrated interaction techniques. The contents, modality and prominence of information, navigation aids, search facility and links to other hypermedia pages are adapted [20] [21]. One of the advantages of adaptable systems is that the users are completely controlled through the interface. On the other side, the behavior of adaptive systems is completely matched with user's mental model. It provides interaction with systems by considering the user performance, ease of system, minimizing request help, removing complexity and avoiding the problems of cognitive overload [14]. Adaptations always support to achieve goals of users for performing actual tasks rather than incorrect predictions [22]. However, considerable amount of work needs to be performed on user's side to adapt the interface. Table 1 elaborates the difference in adaptive and adaptable interfaces.

TABLE I. Comparison Between Adaptive and Adaptable Systems

	Adaptive	**Adaptable**
Definition	It has dynamic adaptation by the system itself to perform the current task.	In adaptable system, the user can change functionality of the system.
Knowledge	It is contained within a system.	It is extended.
Strengths	There is no need of special effort and knowledge by the user.	The system has control of user. Also, user knows the task which is needed to successfully completed.
Weaknesses	There are few success models exists. User has difficulty to develop an intelligible system.	System complexity and incompatibility increased. User needs to learn adaptation component.
Mechanism Required	Models of users, tasks, and dialogs; knowledge base of goals and plans; powerful matching capabilities; incremental update of models.	Layered architecture; domain models and domain-orientation; "back-talk"" from the system; design rationale.
Application Domain	Customization, differential descriptions and information retrieval system is required.	Information retrieval, end-user modifiability, tailor ability, altering, and design is used.

The graphical user interface mechanism to control the customization is usually provided by non-programmatic customizable systems [23] [24]. Generally, this type of interfaces is very helpful for providing direct manipulation of interfaces. The user satisfaction level over the interface interaction can be increased by using this type of customization. It also provides the sense of achievement about completion of tasks especially in situations where responsibility needs to be assigned to the user of a critical system.

III. Usability of Mobile-Phoneinterface

Usability of interfaces for mobile phone devices is really concerned with the satisfaction of users. The basic purpose of usability is to provide guidance for developers to develop user friendly applications. It is not easy to define and evaluate the usability as formal or informal specific environment. Still there is not a precise apparatus to measure the absolute usability of any product [25]. Currently, the term usability is used to enhance interactions among users for their products to complete the tasks according to the environment [26]. In human computer interaction (HCI) the usability [27] is considered as one of the major concepts which has already produced emerging views regarding product usage and customer's satisfaction [28]. The Jakob Nielson [29] is one of the pioneers who tried to objectively evaluate the user experience on digital platforms. Though, it dated back to the 90s when he defined ten important usability heuristic principles (visibility of system status, match between system and real world, user control and freedom, consistency and standards, error prevention, recognition rather than call, flexibility and efficiency of use, aesthetic and minimalist design, help users recognize, diagnose and recover from errors, help and documentation) which are still valid and used in everyday life [30]. Moreover, different attributes of usability (e.g. efficiency, learnability, memorability, rate of errors and satisfaction) are also available for consideration when designing a product or interface. If the interface of any product or mobile phone is difficult to use, even it has excellent functionalities, it doesn't matter, the user will shift towards easy alternative [31]. During the analysis phase of software development, the usability is considered very important for user satisfaction in which user needs to interact with the system. Another aspect is usability architectural impact, which is not only concerned with system's outlook but also required the better user experience (e.g. undo, redo, cancel and selection etc.). By considering the complexity of mobile phone interface usability, there are multiple solutions that have been proposed through user centered methodologies. Amongst the various existing user models, there is not an integral model available which considers the different complex features such as cognitive, motor and psychological [32]. Throughout the usability evaluation, the complexity of major issues is found more multifaceted rather than minor issues. Hence, more effort is required to identify the problems which are violating the heuristics such as missing elements in user interface are difficult to check during interface evaluation. Nielsen [33] evaluated the same product that contained different marked problems. The usability experiments of software application for new users took individually and were also conducted in team. The only one problem was identified as common in four teams while others were quite different.

Molich et al. [34] [35] evaluated the same website of nine organizations to measure the consistency in usability testing. As a result, 310 total problems were reported and 232 of them were unique. Efficiency, effectiveness and satisfaction are the parameters of usability in ISO 9241-11 standard. Although this standard is very compact but provides minor discussion about evaluation of usability. This standard also gives comparatively limited guidelines about how to interpret scores from specific usability metrics.

A. Usability of Mobile-Phone Operating Systems

Recently the usage of mobile phone devices is increasing dramatically throughout the world. There are two major vendors (iOS and Android) prominently available in the global market. iOS is projecting choice of users by providing friendly GUI. It (Apple) also provides a leisure time to customers with most trusted hardware support. On the other hand, Android is also a prominent brand but still struggling for better hardware support. Along with hardware, the usability of mobile phone applications is becoming an increasingly significant part amongst users [36]. Moreover, the operating systems (OS) along with technical possibilities and application compatibilities to enable the functionalities of end device can play vital role in terms of task performing [37]. To measure the usability of smart phone applications, tests were taken on the basis of Ericsson and Simon's work by using think aloud protocol. This protocol encourages the users to think loudly while performing their specific tasks. On the collected information, the user's interface promotes the natural human thinking capabilities to enhance the performance and improvement [38]. The features of interface are developed differently in separate devices according to user desires for better usability. These interfaces may provide large monitors, small screens, enhanced input/output devices (e.g. trackball, keyboard or touch pen) to improve the efficiency, effectiveness and satisfaction [39] [40]. The mobile phone user interfaces should be developed to accommodate the users according to their context.

IV. Mobile usage Context

An interesting point of view about context states that it includes the user's state of emotions, focal point of user's attention, demographics and all the elements present in the user's environment [41]. Even the term context in vocabulary refers to the environment of usage providing details about the user's technological knowledge, user's current state of affairs and application's background, application settings and features of the current usage situation [42] [43]. Thus, it highlights the importance of precise definition of all information that should be considered as context for adaptive application development of mobile phones.

With the above-mentioned aim, we deduced the following major context entities for device, task and usage learning. Usage learning context aims to explore and state the user properties and preference that are quite important instrument for proper adaptations application. In scenario of mobile applications, the user model can be useful in several conducts: the user historical usage pattern can identify and even predict the user needs and select information of use's interest [44]. Such information present principally a significant positive impact on application's interface and its proposed content. The task to model the user with a mobile device and varying environment appears to be much difficult than performing the same for a user in desktop environments [45]. Usability is mostly coupled with interface only while it is the property of overall system. It refers to the quality of use in a specific context [46]. As mentioned above, current methods and techniques for usability prediction are inadequate regarding their accuracy due to consideration of partial context of users, tasks and environment.

There exists a lot of methods to determine the quality of use of any system especially for ICT domain. These evaluations include the effectiveness measures of system by calculating the successful achievement of user's intended goals. The efficiency of the system in measured over the required resources to achieve user's goal. The resource list may include time, money or user's mental effort. The satisfaction is judged over the acceptability of overall system by the user [4]. The overall picture of a system presents users, tasks and technical resources (hardware, software and materials) of that systems surrounded by the physical and organizational infrastructure. All of these elements of the system environment greatly influence the interaction. The promotional campaigns, like "Usability Now!" in the UK, have brought awareness in the buyers of any system about the importance of usability. With the effect of these awareness programs, buyers now give more consideration to ease of use of any (S/W or H/W) systemin their selection. Similarly, the producers and suppliers of the product with high usability get customers attention and a market edge. Microsoft and Amstrad have highlighted the ease of use as a major selling feature in their recent advertising campaigns and gain sales promotions [47]. International usability and the user interface standards are being increasingly referred in public procurement. They also aim to fulfil European Display Screen Equipment Directive.

A. *Context Properties*

Context-aware computing methods and techniques have been used in most of the research for adaptive interface development. The said methodology comprises of sensor technology to collect information about surrounding environment like location, time, daylight, temperature, user identity and action. Adaption with more details has not been generally adopted for such systems like, data input methods based on context. In this research, we tried to identify the user's properties and behavior using domain experience and mobile technology experience. It also aims to congregate all the context elements with scale and value ranges that influence the user's task performance. Consequently, provide design guidelines for the interfaces that can automatically adapt according to the provided context information [48]. The user itself, as a compound entity, is an essential part of the context. There has been very specialized set of behaviors provided by any context-aware application to react against specific context variations. Thus, the software engineer must clearly understand the development goals and classify various context cases in the targeted application.

1) User properties: User's properties are most concerned in many of the context aware applications where context in represented by the user's status like user's age, gender, demographies and emotions. User's demographical properties like location, time, weather or activities are usually recognized through internal/external sensors (e.g.: GPS, accelerometer, Web data camera, microphone), while the emotional state can be mined from user's current responses in contrast with the history of user's actions [49]. The classification and categorization of user's properties is unlike to a data ontology of a person used in social network. User properties can be categorized over the semantics and use of the values in the universe of discourse. Some of the user properties are brought up for user's identification like name any IDs (username, social security number), street, address, city, country, address properties (zip code, country code, telephone number, network code, phone number, home address and email are part of contact details and other business properties may include date of birth) [50]. As humans have diverse lifestyle requirements and these aspects should be modeled through user profile modeling. The profiling of the usercan be used for the desired level of personalized service delivery which attains the capability to adapt itself for a particular user. The user profile is a digital representation of the user and context-aware system places it in the modeling and management layer.A user profile can be characterised as a number of user-related classes. Various dynamic profile aspects have been discussed in five profile classes consisting of CapabilityProfile, InterestProfile, PreferenceProfile, EducationProfile and HealthProfile [51]. Needs of user with specialized requirements have also been addressed in an ontology engineered for user profile to assist people with dementia in mobile usage environment.

2) Device properties: Device context is directly influencial to the usability of any application. Any interface neglecting the device properties cannot achieve ease of use. Incorporation of device elements in interface design process become more severe and complex with the diversity of mobile devices introduced. A mobile application has to be compatible,rather comfortable with several versions. No doubt, currently, mobile devices competencies fulfill many tasks according to user's requirement with multiple hardware options. Since each device claims its own individuality, it is not trivial to design and develop a mobile application. As a mobile application developer, we can refer to the operating system on which this application will be deployed and run while using all hardware facilties (e.g. network connectivity, display, GPS sensors). All these device properties play a vital role mobile applications development. The device context is, in general, considered in the designing and development of mobile application interfaces that works on various operating systems and hardware platforms. The development frameworks, provided by different OS platform association like android or iOS, suggest the functionalities to read the technical information of the device (e.g. memory, display size and resolution, list of sensors and network connectivity options) [52]. The device properties may include some of the followings:

- Font size (Small, medium, large)
- Font color (RGB color, black & white)
- Font format (Times New Roman, Tahoma, etc.)
- Background color (Auto adjust, changing manually)
- Data entry mode (Typing, tapping, voice)
- Display information (Text, sound, LED lights)
- Message delivery (Text, voice, alert, silent, pre-answer)
- Brightness level (Increase/decrease)
- Ring volume (Low, medium, high, alert, vibration)
- Sound level (Mute, regular, loud)

3) Environmental properties: Environemtal properties presents the information about the surrounding environment like spatio-temporal information. The time and space knowledge can be gathered through some sensors with collective knowledge of the cloud. The time can be scaled in various ways like daylight time, office timing, hours/minutes/second, weekend/weekday etc. While space properties may include the location and other properties of surrounding location like weather, humidity, luminosity or noise level. In literature, Module6 of CoDAMoS ontology has been reused for environment modeling in context ontology. Thus, we got a nucleus model for environments and their semantic relationships with the above stated environmental conditions [53] [54]. Different domains of the mobility environment are experts in mobility infrastructure, service development, device connectivity and interface design, etc. [52]. The mobility of the devices tranfering from one to another environment makes it essential to consider the usage environemt in mobile application interface development.

4) Tasks properties: The usability of a mobile application also depends on the task performed and the interface supporting specific tasks. Adaptivity features of a mobile application according to the current activity of user presents another type of context consideration in mobile application designing and development. Unlike the type of context discussed earlier that sense the physical properties, task properties may organise the user interface, with learning the aim of use. The adaptable modalities (e.g. voice, speech synthesis, device vibrations, gesture) are used as a communication way with respect to the tasks between the user and the application [4]. During any adaptive application development, developers need to identify the intent and activities. It is also related to the functionalities provided in specific context by the mobile application. In this way [55], the application is required to read the environment to determine the current activity to be performed by the user. Here, the designer need some state identification to analyze the task and decide the communication options (e.g. sensors, touch screen, network). Another question that arises is to select the proper sensing device or software to determine the intended actions. The option of profile selection (e.g. driving, sport, night) can be given for a specific activity (e.g. texting, calling or listening to news etc.) [56]. Task properties play important role in the development of adaptive features. The context of user activity is necessary to map with device behaviour to fulfill the user desires.

B. Special Needs

Specialized applications and some operating systems interface have been designed for the user with special need or any physical disorder. Assistance for the users with special needs enhance the overall usability of the systems that considers exceptional members of the target user range. These special needs may be provided for people with low and impaired vision due to age factor. Another colour vision problem is colour blindness that refers to a physical deficiency to identify and distinguish among some basic colours (i.e. red, green, blue) in the normal luminosity. These users having colour-blindness may be inadequate to differentiate within specific colour pairs. The applications developed for mobiles may avoid those specific pair of colours in various objects identification in one canvas especially in foreground and background combinations. It is also suggested to provide a colour transformation by the OS [57] to avoid such combination in application running at that platform.

Another study for Deaf' or better to be called naturally challenged users mostly suffer with great hearing loss. Sign Language (SL) can be used as the first language of communication for these naturally challenged users. There can also be options for such users to learn the Sign Language that is based on the combination of movements of hands, arms, body and facial expressions. A study [12] states that there has been nearly sixty thousand people living in Italy who use deaf Italian Sign Language (LIS) as their primary mean of communication.

V. Usability Engineering for Mobile-Phone

Usability of smartphone applications is one of the major concerns in industry today. The trade-offs in user interfaces done by the manufacturers and application developers have resulted in dissatisfaction amongst the users. One of the major reasons is their non-contextual user interfaces [45]. AUIs help us in addressing these problems and increase the usability of applications by providing the functionality that focuses on the user requirements.

Variable regarding user's environment are available in wide range such as demography, cognitive skills, background, education, personality and preferences [46]. The interpretation of different users may not be matched for command names, icons and displays which is one of the major challenges in HCI. In smart world, there is a rapid increase in user interaction with interfaces and direct influence of the context on the user's task in an environment. The context and task define the change that needs to be performed at a specific movement on user's interface. In mobile computing, the context-awareness or physical environment includes surroundings of a user and device (e.g. location and time) [58] [59] [60]. User modeling, in the area of mobile applications can be performed by monitoring users' past behaviors or user profiles. This type of predictive knowledge has an influence on application interfaces and its contents. Context modeling for usage learning has been defined into three categories such as (i) based on domain knowledge (ii) supervised learning approaches and (iii) unsupervised learning approaches. Supervised learning approaches require smaller amount of domain knowledge while unsupervised learning do not require the domain knowledge [52] [61]. It is very difficult to collect relevant information from users because most of the users are not aware which information is important [41]. This issue may create difficulties for application developers to develop adaptive applications according to user's contextual information.

A. Adaptivity and Intelligence in Mobile-Phone Interface

The principles that lay the foundation of adaptive systems consider the situations that make the need of adaptation necessary. Based on these requirements appropriate adaptation plan is decided and the actions are taken accordingly. Fundamental choices that constitute the adaptation process of AUI design are [14]:

a) Establishing the role of the adaptation and by whom the adaptation will be performed.

b) Definition of overall goals of the user and the adaptation process.

c) Definition of adaptation rules which will manage the adaptation.

d) Definition of variables and level of interaction that is required for adaptation.

e) Definition of a complete inference mechanism and methods that will perform the adaptation process. All these methods should be in line with the user preferences.

The mobile phone industry is penetrating in the market and users need specialized interfaces to fulfill their requirements. Although adaptivity is the need of hour and it has some problems and tremendous benefits to create user satisfaction. HAoAok identifies [16] [62] a few problems of adaptive interfaces in one of his articles which summarizes the state of art in the field. First problem refers to the control of user where the user is not provided with the control of the adaptation, this is referred to as *lack of control*. Though the user involvement may increase the satisfaction of the user, but this may result in increasing the problems in developing the user model for adaptation. Second problem refers to the consequences of a user action. As user has no direct control over the adaptation process, hence the user is not sure about the results of some actions, this is classified as *unpredictability*. Third problem identified again refers to the user's understanding of the adaptation process and what the user actually expects from the interface. This makes it difficult for the developers to decide the portions that should be shown to the user at a given time, this problem is named as *transparency*. Fourth problem is termed as *privacy*, this means that user must accept the fact that everything that user is doing is recorded and will be used for adaptation of the user interface. On one side it is compulsory for the system to keep track of every task of user for building a profile, but, on the other side it creates a privacy concern for the user. Fifth problem refers to the *trust* on the adaptation process, according to literature; user's trust is volatile and may decrease further if the adaptation performed by the user does not fulfill his requirement.

AUI gives user flexibility of creation of interface at design level, not only this but it also allows the user to make the required changes during execution as well. This makes the user independent and does not restrict the developers and designers of the system to decide the optimal solution that are specific to a user. By using AUIs the system has the capability to adapt to the user need and also helps in future adaptations by keeping a history of adaptations performed [20] [58]. Adaptation has already been implemented in all types of systems ranging from desktop applications to web applications, smart phone applications, wearable device software and many more platforms.

The basic premise that an adaptive system work on is the user model, this model is built on the user's data, his behavior (which is recorded by keeping a track of actions by the user) and environmental conditions. The system predicts the user's activity and gives relevant information, functionality and suggestions to the user for the next task. By using this mechanism, the content is personalized for each user and is based on the user model developed by system at runtime [63], hence, the application performs differently for each user. This increases the user satisfaction as the feeling for using an application that understands an individual's preference is what every user wants. Another study analyzes the role of context in the telematics devices available in vehicles. Sixty-four students, including 35 male and 29 female students, from Ben-Gurion University having average age of 25.7 participated in the study [58]. It was observed that the overall performance time decreased with the increase in adaptivity, hence the study concluded that adaptivity level directly effects the overall performance time of the user.

B. Criteria to Develop Smart AUIs

The need of user interfaces is rising gradually day by day. These interfaces are running on multiple devices along with various features. Any type of disability in interfaces creates motivation to develop the guidelines of interface generation and description of user interface. There are prominent approaches including MARIA, TERESA with Concur Task Trees, Personal Universal Controller (PUC) as used in Huddle and Uniform, UIML Canonical Abstract Prototypes (CAP) with recent modifications in CAP3. The following parameters may help to develop criteria for mobile phone adaptive user interfaces.

1) Run-Time adaptations and usability: Run time adaptation of user interface used in few systems to provide automatic generation and usability. However, numerous literatures have been written about potential problems of self-adaptive [64] user interfaces. Specially, it lacks in transparency which is described to improve the run time adaptations by animated transitions. Nevertheless, the aim of adaptive user interfaces to accessibility improvements may have a need of different approaches where the system takes a more proactive role [11]. In broader perspective, the goal of intelligent user interfaces can be achieved by relieving of complexity.

2) Direct and indirect adaptation: The direct and indirect adaptation can support to make an approach for a wider variety of scenarios. By providing alternative version of adapted UI or indirect adaptation, the confusion of users can be minimized.Yet the comparison between adaptive and adaptable interfaces is not mature enough, only in some systems (e.g. ubiquitous), it is necessary to adapt UI with direct adaptation. The study is conducted by developing prototype as a frontend to commercial word processor. For example, the participants used MSWord personal interface for evaluation in four weeks, where 14 out of 19 users spent almost less or more than 50% time in their personal interface usage [65]. In daily life, the smart environments which are being used in utilizing the computers as tools.It suppors the users for moving interaction (direct and indirect) with computers from a single system to a complex and distributed environment. The required synchronization [66] of the parts is a problem with distributed user interfaces (e.g. MASP) which adapts the smart home UIs on the basis of changability in environment.

3) Extensibility of adaptive behaviour: Extensibility is an important feature for the development of new user interface.It helps for multiple adaptation (e.g. accessibility, cognition etc.) of UI because it is not restricted to single type of adaptation such as layout optimization. The extensibility of adaptive behaviour makes available to add new adaptive behaviour at run time when desired to provision a diversity of aspects [67]. For example, COMET is modeled for supporting polymorphism that belongs to different technological spaces (e.g. HTML, OpenGL, vocal). The dynamic ability to tailortheuser interfaces with designer and end user to discover design alternatives by substituting COMET presentations at design as well as runtime.

4) Empowering new design participants: The new design participants can be non-developers (end users, IT personels) in the case of adaptive user interfaces. For example, leveraging communities through crowdsourcing could prove useful for applications that require a lot of effort for defining the adaptive behavior.

5) Integrated development environment: The integrated development environment (IDE) is a style of user interface which looks similar to Visual Studio or Eclips [68]. It can provide easiness for organization UI and adaptive behaviour articles of large scale software systems.

6) Supporting multiple levels of abstraction: The CAMELEONsuggeststhe liberty of task modeling, abstraction and technology of concrete UI. Moreover, the different levels might be appropriate for certain type of adaptation [67]. By adapting highest level, the UI features can be reduced and also through a number of levels, the layout could be optimized.

7) Selected modeling approach: In this approach the selected interpreted runtime modeling permits advanced adaptations to be conducted [69]. The old fashioned approach "model driven engineering" (MDE), used in HCI has revived conventional (WIMP) user interfaces [70]. This approach brings some hope by providing theoretical and task level modeling into a unifying and systematic approach to the problem of UI plasticity.

8) Modeling, generation, and synchronization: It provides abstraction at all levels where the tools of model driven UI development must create easiness to the developers. Predictability for the programmers is very important and should not be ignored during the development of tools [71]. However, synchronization is one of the main feature that creates user satisfaction.

9) Supporting multiple data sources: The architectures and procedures of adaptation enable the users interface generation and dynamic adaptivity during run time [11]. The multiple data sources permits adaptations to be carried out in various situations. The models having adaptive behaviour [72] can exemplify data on the basis of different studies, which is the case of adapting UIs to cultural preferences by MOCCA.

10) Preserving designer input on the UI: In some cases, [71] the designers may wish to preserve some characteristics to enhance the predictability of outcome.

11) Reducing solution viscosity: It is achieved if a tool reduces [68] the effort required to iterate on the possible solutions based on the [71] flexibility, express leverage, expressive match, scalability, low threshold, high ceiling andtrade off analysis.

12) User feedback on the adapted UI: The user feedback on AUI [72] provides awareness of automated adaptation decisions and the ability to take priority overthem whenever needed.

C. Development Process Model of AUI

This section discusses the development process model of adaptive user interface.

1) User center design process model: The philosophy of UCD is to analyze the usability and validity of mobile device interfaces according to user contexts. The AIO will provide the designing and scaling of mobile device interfaces according to the context of user. UCD is a framework of processes and methodology which is not restricted to interfaces or technologies [12]. It deals with the product for the understanding of needs, wants, tasks, environments, preferences and limitations in user's context and are given extensive attention at each stage of the design process. It is the process of designing software with interfaces and then solving of multi-stage problems from the perspective of user's understandability. The system can be designed for user's support with their existing beliefs, attitude and behaviors related to their tasks, rather than the users adapt or learn the designed system [73]. UCD approach is used to develop simple models, mock-ups or prototypes on parts or all of the designs such that graphical design, information architecture, interaction design, information visualization. The UCD not only requires designers to analyze and predict that how users use a product but also test the validity with regards to user's behavior. The testing of a product is necessary but it is difficult task for designers to understand the user's experiences. UCD has complete life-cycle to produce the products with high usability and low cost [4]. The major goal of UCD is to offer optimized, efficient and user friendly product which increases the usability and satisfaction of users.

2) Mapping UCD process model in UCO and AIO: All features of UCD will provide real benefits to the user experience. In UCD, the prototypes are very useful to translate the user requirements into contextual experience. It is used to enhance the usability, satisfaction and optimization of adaptive mobile user interfaces. UCO and AIO will be developed to optimize the use of adaptive mobile interfaces in the context of user preferences [74]. These ontologies organize the perceptions and thoughts of user. "Fig. 1" shows the proposed research roadmap for the engineering of UCO and AIO for mobile device interfaces in user's context. It also describes the mapping functions of UCD [50] with UCO to AIO which leads towards the experimentation and evaluation of results for the satisfaction of users.

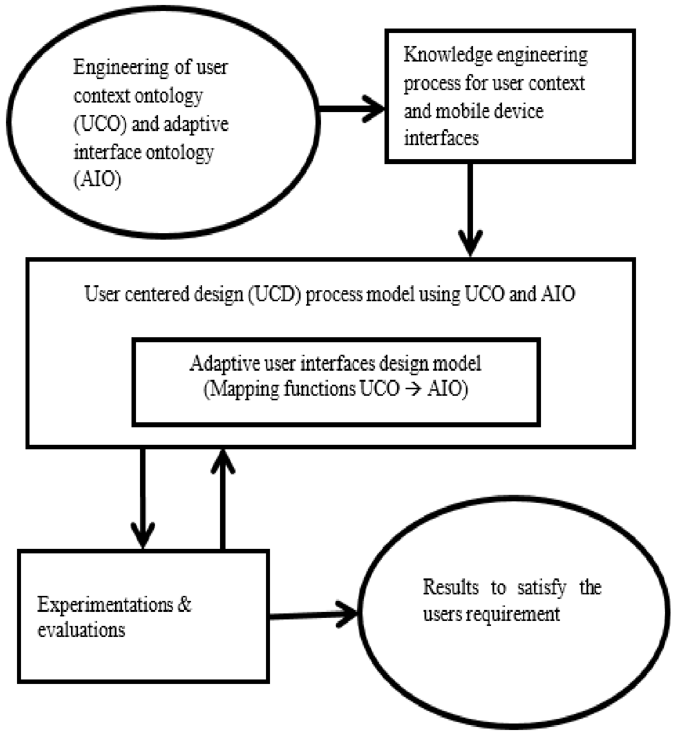


Fig. 1. Development Process Model for Adaptive Mobile User Interface.

VI. Evaluation of Interface

Usability attributes are classified in two types of measures, objective and subjective measures. There are five usability attributes that help us to evaluate an interface in human computer interaction. Objective measures are very useful in evaluation of an interface; however, the collection of data objectively is an expensive, time consuming and challenging task. On the other hand, subjective data can be collected easily, speedily and with comparatively less effort. Attributes like user opinion and preferences can also be measured using subjective approach of evaluation [75]. For instance, many of modern smart phone applications have multiple usability problems which include the navigation, poor support for performing tasks, complicated interfaces, complex interaction styles, limited interaction techniques and confusion due to a lot of options given to user. The magnitude of problem is increased for mobile phones because of their limited processing power, screen size, mobility and multiple network connectivity issues.

There exist a lot of algorithms that can be used to perform multiple types of adaptation and each algorithm has its strength in the evaluation process [76]. According to a study by Lewis in 2006, hidden smart menus of pre-2007 versions of Microsoft Office resulted in multiple usability issues, however, the revised version of MS Office released after 2007 contained predefined adaptive parts which used to show the most recently used tasks increased the user satisfaction and improved the overall usability [62], thus proving to be more beneficial for the user. Menthol project examined the parameters like age and gender on mobile phone usage and the results showed that these parameters have a direct relation to the usability of a smart phone. Due to the popularity and personal aspect, usability of a smartphone is one of the hot research areas in pervasive computing. The project analyzes the above-mentioned factors where the dataset is an output of a longitudinal study. A sample of more than thirty thousand participants was selected, out of which around sixty thousand were male and fourteen thousand five hundred were female. The median age of the participants was 21 years and they were tracked for 28 days. The data of their personal and demographic details was submitted through a predefined questionnaire. The study was conducted in January 2014. It was observed that the average time a female spends on a smart phone is 166.78 minutes, whereas males spend 154.26 minutes, hence, females use their devices for a longer period then males. Another finding of the project was that the younger participants were more inclined towards entertainment and social networking applications through specialized software applications. Young users also spent more time then the users who are older in age [36]. After analysis it was found that the older users used the smart phones for getting general information and preferred to use the smart phone as a conventional phone.

In this paper we have analyzed the usability of the adaptive features that have been provided by the device manufacturers. The features that were analyzed for user satisfaction, efficiency and effectiveness. A user centric design was followed to evaluate the usability of features that are adaptive [4] [64]. It was seen that the users preferred the portrait mode while typing because it gives the user the ease of typing by using single hand, the overall effectiveness of the feature of screen rotation was 25% less effective. An interesting finding was that the voice command adaptive feature, despite of the fact that the feature was 6% more efficient, but surprisingly it was 19% less effective during the user interaction. One of the reasons that were observed in less effectiveness of the voice command feature was the problem in recognizing the accent of the user. The effectiveness of LED notifications adaptive feature was recorded as 88%, not only this but the overall efficiency was also 89%, hence, the LED notification feature was very effective and efficient. For kids the adaptive environment feature was found to be 28% more efficient and effective than using the smartphone in normal mode. It was observed that because of the identical adaptive features provided by multiple vendors, there were a lot of adaptability issues. Another point that creates the problems is that the vendors do not consider the user's ability to perform a task and the context of the task. From the experimental findings, it is inferred that the adaptivity has a positive impact on the usability of a smart phone, when it is applied in an appropriate context [17]. Another example that supports the findings was an adaptive bar which was presented by Debevc et al [77]. In this toolbar, the software system proposed the additions and deletions of tasks to the user; suggestions are given on the basis of history, frequency and context of the user. The results of another study show that the adaptive tool bar was more effective and helped users to build their own tool bar more efficiently, thus, enhancing the user experience.

VII. Conclusion

The fundamental purpose of HCI is to make the systems more usable, useful and to provide better usability according to user experiences. Many researchers have suggested AUI in their own specific perspective. Currently, the smart world has many variations in all aspects of user, task, environment and device. These deviations in usage environment are increasing speedily which cannot be handled by using single interface. Therefore, it is near to impossible to develop specialized interfaces for each context separately. In this survey paper, we studied different principles and mechanisms used in context models. User modelling research is discussed to address these types of issues. Likewise, semantic modelling has also provided a solution for complex scenarios and computational models. These complex contexts can be modelled by using semantics and ontologies. Furthermore, the user interface process model suggested development of AUIs on the basis of guidelines provided in this survey paper.

References

[1] D. Benyon and D. Murray, “Applying user modelling to human-computer interaction design,” AI Rev., 1993.

[2] A. Petrovčič, A. Rogelj, and V. Dolničar, “Smart but not adapted enough: Heuristic evaluation of smartphone launchers with an adapted interface and assistive technologies for older adults,” Computer Human Behaviour, 2018.

[3] R. Hervás and J. Bravo, “Towards the ubiquitous visualization: Adaptive user-interfaces based on the Semantic Web,” Interact. Comput., 2011.

[4] M. W. Iqbal, N. Ahmad, and S. K. Shahzad, “Usability evaluation of adaptive features in smartphones,” in Procedia Computer Science, 2017, vol. 112, pp. 2185–2194.

[5] P. A. Akiki, A. K. Bandara, and Y. Yu, “Adaptive Model-Driven User Interface Development Systems,” ACM Comput. Surv., vol. 47, no. 1, pp. 1–33, 2014.

[6] H. Petrie and N. Bevan, “The evaluation of accessibility, usability and user experience,” Univers. Access Handb., 2009.

[7] Ál. FernáNdez-LóPez, M. J. RodríGuez-FóRtiz, M. L. RodríGuez-Almendros, and M. J. MartíNez-Segura, “Mobile learning technology based on iOS devices to support students with special education needs,” Comput. Educ., 2013.

[8] T. Paymans and J. Lindenberg, “Usability trade-offs for adaptive user interfaces: ease of use and learnability,” Intell. user interfaces, no. c, pp. 1271–1274, 2004.

[9] S. Kane, “Improving Mobile Phone Accessibility with Adaptive User Interfaces,” Ideals.Illinois.Edu, 2010.

[10] A. Ali, M. Alrasheedi, A. Ouda, and L. F. Capretz, “A Study of the Interface Usability Issues of Mobile Learning Applications for Smart Phones from the User’s Perspective,” Int. J. Integr. Technol. Educ., 2014.

[11] M. Peissner, D. Häbe, D. Janssen, and T. Sellner, “MyUI: Generating Accessible User Interfaces from Multimodal Design Patterns,” in Proceedings of the 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, 2012.

[12] S. D. I. Dottorato and N. Ahmad, “Tesi di Dottorato People Centered HMI’s for Deaf and Functionally Illiterate Users,” 2014.

[13] N. U. Bhaskar, S. N. Raju, G. Paladugu, and V. Reddy, “Aspects of Content, Context and Adaptation Modeling in Mobile Learning Application Design,” Int. J. Interact. Mob. Technol., vol. 7, no. 2, pp. 29–37, 2013.

[14] F. Gulla, L. Cavalieri, S. Ceccacci, M. Germani, and R. Bevilacqua, Method to design adaptable and adaptive user interfaces, vol. 528. 2015.

[15] S. Greenberg and I. H. Witten, “Adaptive personalized interfaces: a question of viability,” Behav. Inf. Technol., 1985.

[16] a T. Submitted, I. N. Partial, F. Of, T. H. E. Requirements, F. O. R. The, and D. Of, “Comparing Static, Adaptable, and Adaptive Menus by The University of British Columbia,” no. August, 2004.

[17] E. Tenner, “The Design of Everyday Things by Donald Norman,” Technol. Cult., 2015.

[18] A. Cadenas et al., “Context Management in Mobile Environments: A Semantic Approach,” Proc. 1st Work. Context. Inf. Ontol., p. 2:1--2:8, 2009.

[19] V. Glavinic, S. Ljubic, and M. Kukec, “Transformable Menu Component for Mobile Device Applications: Working with both Adaptive and Adaptable User Interfaces,” Int. J. Interact. Mob. Technol., 2008.

[20] R. Oppermann, Adaptive user support: ergonomic design of manually and automatically adaptable software. 1994.

[21] J. Fink, A. Kobsa, and A. Nill, “Adaptable and Adaptive Information Access for All Users, Including the Disabled and the Elderly BT-User Modeling,” User Model., no. Chapter 19, pp. 171–173, 1997.

[22] L. Findlater and J. McGrenere, “A comparison of static, adaptive, and adaptable menus,” Proc. ACM CHI 2004, vol. 6, no. 1, pp. 89–96, 2004.

[23] B. Shneiderman, “Direct manipulation for comprehensible, predictable and controllable user interfaces,” Proc. 2nd Int. Conf. Intell. user interfaces- IUI ’97, pp. 33–39, 1997.

[24] M. Manca and F. Paternò, “Customizable dynamic user interface distribution,” in Proceedings of the 8th ACM SIGCHI Symposium on Engineering Interactive Computing Systems - EICS ’16, 2016.

[25] G. Fischer, “User modeling in human computer interaction,” User Model. User Adapt. Interact., vol. 11, pp. 65–86, 2001.

[26] J. R. Lewis, “Usability Testing,” Handbook of Human Factors Testing and Evaluation. pp. 1275–1316, 2006.

[27] J. Hewes, “Perspectives on Usability,” Conf. Rec. Crossing Front., 1992.

[28] N. A. Mian and F. Ahmad, “Modeling and Analysis of MAPE-K loop in Self Adaptive Systems using Petri Nets,” vol. 17, no. 12, pp. 158–163, 2017.

[29] J. Nielsen, “Ten Usability Heuristics,” Copyr. © 2005 by Jakob Nielsen. ISSN 1548-5552, pp. 1–2, 2005.

[30] R. Molich and J. S. Dumas, “Comparative usability evaluation (CUE-4),” Behav. Inf. Technol., 2008.

[31] R. Shehzad and N. Ahmad, “Web Usability and User Trust on E-commerce Websites in Pakistan,” vol. 8, no. 12, 2017.

[32] A. M. Figueroa, R. Juárez-Ramírez, S. Inzunza, and R. Valenzuela, “Implementing adaptive interfaces: A user model for the development of usability in interactive systems,” Comput. Syst. Sci. Eng., 2014.

[33] J. Nielsen, “Finding usability problems through heuristic evaluation,” Proc. SIGCHI Conf. Hum. factors Comput. Syst. - CHI ’92, pp. 373–380, 1992.

[34] R. Molich et al., “Comparative evaluation of usability tests,” CHI’99 Ext.vol. 2, no. May, p. 83, 1999.

[35] ISO, “Iso 9241-11,” Ergon. Requir. Off. Work with Vis. Disp. Termin. Part 11 Guid. usability, vol. 1998, p. 22, 1998.

[36] I. Andone, K. Błaszkiewicz, M. Eibes, B. Trendafilov, C. Montag, and A. Markowetz, “How age and gender affect smartphone usage,” in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp ’16, 2016.

[37] G. Jindal and M. Jain, “A Comparative Study of Mobile Phone’s Operating Systems,” Int. J. Comput. Appl. Inf. Technol. I, Issue III, vol. 1, no. 3, pp. 2278–7720, 2012.

[38] N. Ahmad, M. W. Boota, and A. H. Masoom, “Smart Phone Application Evaluation with Usability Testing Approach,” J. Softw. Eng. Appl., vol. 07, no. 12, p. 1045, 2014.

[39] J. Eisenstein, J. Vanderdonckt, and A. Puerta, “Adapting to mobile contexts with user-interface modeling,” Proc. - 3rd IEEE Work. Mob. Comput. Syst. Appl. WMCSA 2000, pp. 83–92, 2000.

[40] I. Sciences, “Transparency in Adaptive Mobile User Interfaces Gaël van Heijst,” no. June, 2016.

[41] A. K. Dey and G. D. Abowd, “Towards a Better Understanding of Context and Context-Awareness,” Comput. Syst., vol. 40, no. 3, pp. 304–307, 1999.

[42] M. Zheng, S. Cheng, and Q. Xu, “Context-Based Mobile User Interface,” no. July, pp. 1–9, 2016.

[43] I. Uddin, A. Rakib, H. M. U. Haque, and P. C. Vinh, "Modeling and Reasoning about Preference-Based Context-Aware Agents over Heterogeneous Knowledge Sources," Mob. Networks Appl., vol. 23, no. 1, pp. 13–26, 2018.

[44] J. Park and S. H. Han, "Integration of Adaptable and Adaptive Approaches for Interface Personalization Through Collaborative Menu," Int. J. Hum. Comput. Interact., vol. 28, no. 9, pp. 613–626, 2012.

[45] R. Mizouni, M. A. Matar, Z. Al Mahmoud, S. Alzahmi, and A. Salah, "A framework for context-aware self-adaptive mobile applications SPL," Expert Syst. Appl., vol. 41, no. 16, pp. 7549–7564, 2014.

[46] S. A. Hoseini-Tabatabaei, A. Gluhak, and R. Tafazolli, "A survey on smartphone-based systems for opportunistic user context recognition," ACM Comput. Surv., vol. 45, no. 3, pp. 1–51, 2013.

[47] N. Bevan and M. Macleod, "Usability measurement in context," Behav. Inf. Technol., vol. 13, no. 1–2, pp. 132–145, 1994.

[48] M. Zheng, O. Ormandjieva, and H. Fan, "Designing Context Sensitive Mobile User Interface," in Proceedings of the International Conference on Software Engineering Research and Practice (SERP), 2015.

[49] A. Schmidt, M. Beigl, and H. W. Gellersen, "There is more to context than location," Comput. Graph., vol. 23, no. 6, pp. 893–901, 1999.

[50] S. K. Shahzad, "Ontology-based User Interface Development: User Experience Elements Pattern," J. Univers. Comput. Sci., vol. 17, no. 7, pp. 1078–1088, 2011.

[51] K. L. Skillen, L. Chen, C. D. Nugent, M. P. Donnelly, and I. Solheim, "A user profile ontology based approach for assisting people with dementia in mobile environments.," Conf. Proc. IEEE Eng. Med. Biol. Soc., vol. 2012, pp. 6390–3, 2012.

[52] J. Hussain et al., "Model-based adaptive user interface based on context and user experience evaluation," J. Multimodal User Interfaces, vol. 12, no. 1, pp. 1–16, 2018.

[53] V. Wiens and S. Lohmann, "Device-Independent Visual Ontology Modeling," 2018, pp. 1–15.

[54] M. Poveda-Villalón, M. C. Suárez-Figueroa, R. García-Castro, and A. Gómez-Pérez, "A context ontology for mobile environments," in CEUR Workshop Proceedings, 2010, vol. 626.

[55] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems," Int. J. Ad Hoc Ubiquitous Comput., vol. 2, no. 4, p. 263, 2007.

[56] G. Meditskos, S. Dasiopoulou, and I. Kompatsiaris, "MetaQ: A knowledge-driven framework for context-aware activity recognition combining SPARQL and OWL 2 activity patterns," Pervasive Mob. Comput., vol. 25, pp. 104–124, 2016.

[57] M. W. Iqbal and N. Ahmad, "Adaptive Interface for Color-Blind People in Mobile-Phones," 2018.

[58] T. Lavie and J. Meyer, "Benefits and costs of adaptive user interfaces," Int. J. Hum. Comput. Stud., vol. 68, no. 8, pp. 508–524, 2010.

[59] A. Holzinger, M. Geier, and P. Germanakos, "On the development of smart adaptive user interfaces for mobile e-business applications: Towards enhancing user experience - Some lessons learned," in 3rd International Conference on Data Communication Networking, DCNET 2012.

[60] I. Riahi and F. Moussa, "A formal approach for modeling context-aware Human-Computer System," Comput. Electr. Eng., vol. 44, pp. 241–261, 2015.

[61] R. Klaassen, R. op den Akker, T. Lavrysen, and S. van Wissen, "User preferences for multi-device context-aware feedback in a digital coaching system," J. Multimodal User Interfaces, vol. 7, no. 3, pp. 247–267, 2013.

[62] M. Hartmann, "Challenges in Developing User-Adaptive Intelligent User Interfaces," Proc. 17th Work. Adapt. User Model. Interact. Syst., pp. 6–11, 2009.

[63] K. Olga, "Adaptive User Interface Patterns for Mobile Applications,"2014.

[64] N. A. Mian and F. Ahmad, "Modeling and Analysis of MAPE-K loop in Self Adaptive Systems using Petri Nets," vol. 17, no. 12, pp. 158–163, 2017.

[65] J. Mcgrenere, R. M. Baecker, and K. S. Booth, "An evaluation of a multiple interface design solution for bloated software\proceedings of the sigchi conference on human factors in computing systems: Changing our world, changing ourselves," Chi '02, no. 4, pp. 164–170, 2002.

[66] M. Blumendorf and S. Feuerstack, "Event-based Synchronization of Model- Based Multimodal User Interfaces," 2006.

[67] A. Demeure, G. Calvary, and K. Coninx, "COMET(s), A Software Architecture Style and an Interactors Toolkit for Plastic User Interfaces," in Lecture Notes in Computer Science, 2008.

[68] D. R. Olsen Jr., "Evaluating user interface systems research," 20th Annu. ACM Symp. User interface Softw. Technol., pp. 251–258, 2007.

[69] R. Soley and OMG Staff Strategy Gropu, "Model driven architecture," OMG white Pap., no. April, pp. 1–12, 2000.

[70] J. Coutaz, "User Interface Plasticity: Model Driven Engineering to the Limit!" Proc. 2Nd ACM SIGCHI Symp. Eng. Interact. Comput. Syst., 2010.

[71] B. Myers, S. E. Hudson, and R. Pausch, "Past, present, and future of user interface software tools," ACM Trans. Comput. Interact., vol. 7, no. 1, pp. 3–28, 2000.

[72] P. A. Akiki, "Engineering adaptive user interfaces for enterprise applications," EICS 2013 - Proc. ACM SIGCHI Symp. Eng. Interact. Comput. Syst., pp. 151–154, 2013.

[73] R. Roth, K. Ross, and A. MacEachren, "User-Centered Design for Interactive Maps: A Case Study in Crime Analysis," ISPRS Int. J. Geo-Information, vol. 4, no. 1, pp. 262–301, 2015.

[74] E. Castillejo, A. Almeida, and D. López-de-Ipiña, "Ontology-Based Model for Supporting Dynamic and Adaptive User Interfaces," Int. J. Hum. Comput. Interact., 2014.

[75] A. Hussain and U. Utara, "Apps vs Devices: Can the Usability of Mobile Apps can be Decoupled from the Device? Apps vs Devices: vol. 9, no. May 2016, pp. 11–16, 2012.

[76] J. L. Wesson et al., "Can Adaptive Interfaces Improve the Usability of Mobile Applications? 2014.

[77] M. Debevc, B. Meyer, D. Donlagic, and R. Svecko, "Design and evaluation of an adaptive icon toolbar," User Modeling and User-Adapted Interaction. 1996.

Electronically Reconfigurable Two-Stage Schiffman Phase Shifter for Ku Band Beam Steering Applications

Rawia Wali[1], Lotfi Osman[2]
University of Tunis El Manar, Faculty of Sciences of Tunis
Microwave Electronics Research Laboratory
2092, Tunis, Tunisia

Tchanguiz Razban[3], Yann Mahé[4]
IETR UMR CNRS 6164
Polytech Nantes, University of Nantes
44306 Nantes, France

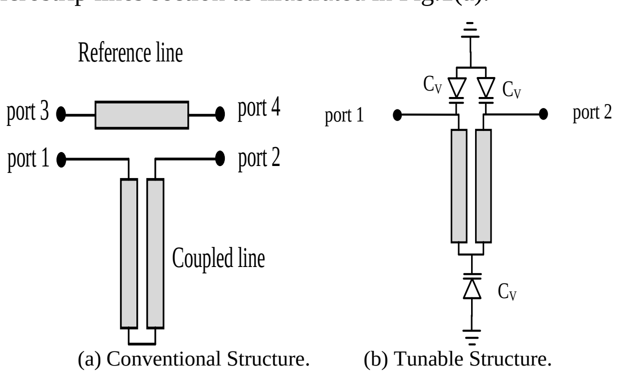
***Abstract*—An electronically reconfigurable phase shifter using two Schiffman sections is performed for beam steering applications in Ku band. The proposed phase shifter consists of only two cascaded coupled-line sections with the reference line removed. This circuit is loaded by varactor diodes that ensure its tunability over a wide bandwidth. By supplying these varactor diodes with suitable bias voltages, a phase shift is continuously adjusted and reached up to 168° at 12.7 GHz with low insertion losses according to the simulations. Thus, the proposed two-stage phase shifter is able to reach a beam steering angle of 28.6° at 12.7 GHz with only one control voltage. The proposed structure exhibits that our phase shifter has a compact size and a large phase shifting range throughout the Ku band. The tunable phase shifter is prototyped and the measurement results are presented.**

Keywords—Schiffman phase shifter; reconfigurable; varactor diode; beam steerability; Ku-band

I. Introduction

Differential phase shifters are frequently exploited in communication systems such as phased array antennas and microwave control devices. This appliance is made of main and reference lines which produce a constant differential phase shift between them [1-3].

One of the most appealing types of differential phase shifters is the Schiffman structure [4-8] due to its simple topology and wideband characteristic. For this configuration, the differential phase shift is provided by a parallel-coupled microstrip lines section as illustrated in Fig.1(a).



(a) Conventional Structure. (b) Tunable Structure.

Fig. 1. Schematic Design of the Single Stage Schiffman Phase Shifter.

Recently, the tunable differential phase shifters have become increasingly beneficial to achieve the beam steering functions using compact structures [9, 10]. In the case of steerable antenna arrays, a large phase shifting range is required to control the direction of the radiation pattern. To that end, a various design of reconfigurable phase shifters suitable for beam steering applications has been mentioned in literature [11-13]. These phase shifters generally contain some kind of active devices offering an adjustable phase shift throughout a large frequency band [12]. One of the prominent tuning components is the varactor diode which has a high phase variation by continuously adjusting the voltages [13].

The Schiffman phase shifter compensated with varactor diodes is a good candidate for beam steering applications because of its potential to achieve a continuously variable phase shift throughout a large frequency band keeping the compactness of the design [14]. This phase shifter model showed in Fig. 1(b), is proposed and developed in [15]. This new approach has some attractive features such as the simplicity of the structure in a planar configuration, the compact size, the wide bandwidth and the large phase shift range with easy adjustment of the phase shift.

Nevertheless, some of the tunable single stage phase shifters produce a restricted range of phase shift variation. Therefore, to resolve this problem, a cascade configuration of phase shifters seems to be an interesting solution to achieve a large phase shift response [16-21]. In this respect, few Ku-band tunable phase shifter designs are referenced in the literature. Therefore, a new structure with both compact size and large phase variation range in Ku band is favored to meet the requirements of beam steering applications.

This letter reports the design, for the first time, of a compact configuration of tunable phase shifter using two Schiffman sections operating especially in Ku band for beam steering applications. To accomplish the continuous tunability of the phase shifter, varactor diodes are introduced to the structure, which includes two pairs of cascaded coupled lines. In the phase shifter structure, it is suggested to cascade the Schiffman sections as a means of increasing the phase shift variation while maintaining the compactness of the phase shifter.

The novelty of this work is proposing a phase shifter that provides an electronically adjustable phase shift through changing the bias voltages of the diodes over the Ku band without increasing the phase shifter size.

The aim of this work is to integrate, as part of a larger ongoing project, our tunable phase shifter in a compact feeding network of the phased arrays antenna, allowing it to continuously steer its beam towards geo-satellites for Ku-band television broadcasting.

II. Phase Shifter Design

Many modified Schiffman phase shifters were proposed for different purposes, such as bandwidth enhancement, and circuit miniaturization [5]. Many researches suggest many structures such as the double-coupled lines [6, 7], cascaded coupled lines, parallel coupled lines [4], and multi-section coupled lines structures [8]. In our proposed design, the original Schiffman phase shifter is amended to develop an accordable phase shifter, having coupled lines charged by varactor diodes without reference line.

So as to enlarge the phase shift variation, two single stage phase shifters are connected in cascade. Each one has a Schiffman structure loaded with three varactor diodes (Cv) to produce a continuously adjustable phase shift by controlling the diode's voltage.

Pursuing simplicity, three identical varactor diodes (MA46H120) are used in the layout of the suggested single stage phase shifter. In fact, two varactor diodes were placed in series, and another one was linked with the ground plane [18]. This diode was selected because of its high speed and its large capacitance variation. According to the instructions provided by the manufacturer, its capacitance value varies from 1.15 pF to 0.15 pF when the diode is biased from 0 to 18 Volts. The measured S-parameters of the varactor diode and its equivalent electrical circuit are reported to utilize them during the optimization step.

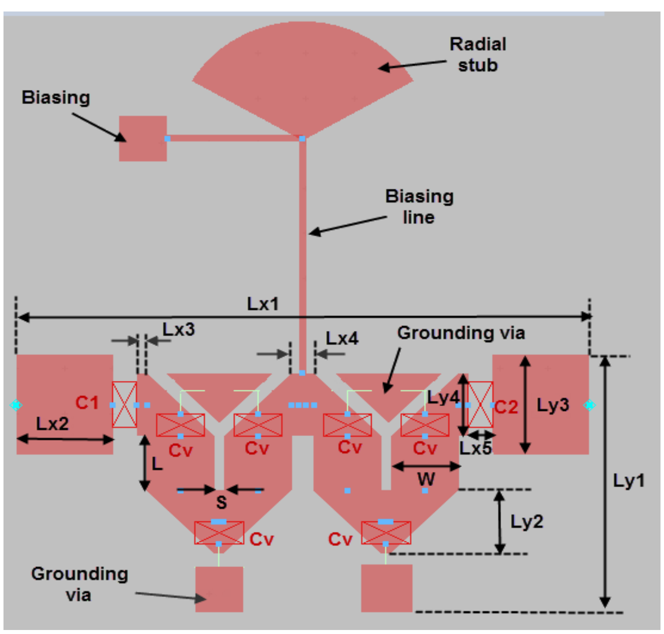


Fig. 2. Design of the Two-Stage Tunable Schiffman Phase Shifter.

TABLE I. Design Parameters and Dimensions Values for the Suggested Phase Shifter

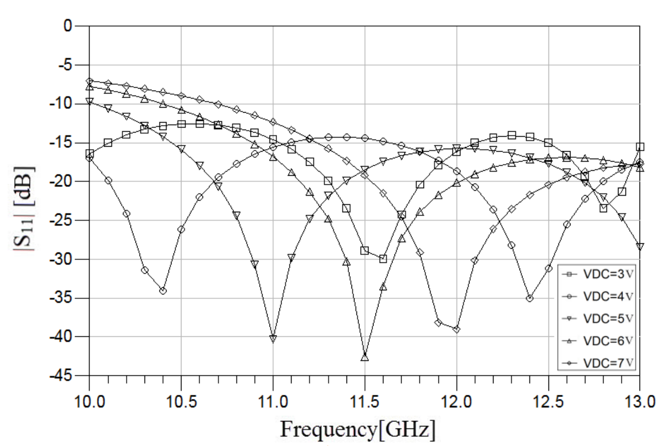
Parameters	Value[mm]	Parameters	Value[mm]
L	1.2	Ly4	1.4
W	1.4	Lx1	11.8
S	0.2	Lx2	2
Ly1	5.7	Lx3	0.2
Ly2	1.4	Lx4	0.45
Ly3	2.23	Lx5	0.5

The placement of the diodes was carefully considered with the aim of getting only to get only one feed line for the whole phase shifter. Thus, the two single stage phase shifters do not need to be individually biased. In this case, one bias line is connected to the phase shifter where the voltages are applied to polarize the six varactor diodes. In addition, the design includes two capacitances C1 and C2 which serve to uncouple the input power in the phase shifter from the supply line.

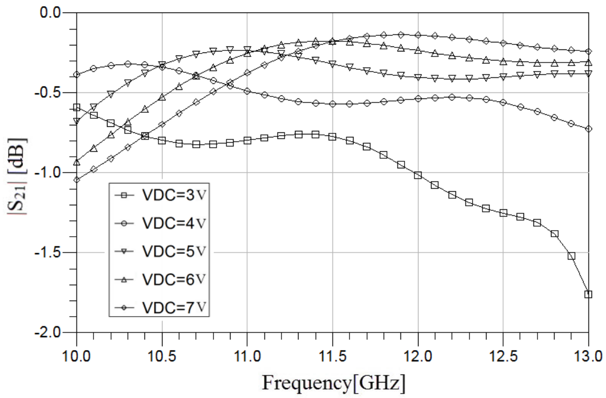
The phase shifter was implemented on a Teflon substrate having a relative permittivity of 2.55 and a thickness of 0.8 mm. It is designed using microstrip technology and simulated through the ADS simulator in the frequency range 10.7-12.7 GHz. The design of the proposed phase shifter is illustrated in Fig. 2 and its dimensions are enumerated in Table I.

III. Results and Discussions

The simulated S-parameters of the two-stage Schiffman phase shifter for different control voltages are presented in Fig. 3. In accordance with this figure, it is noted that the phase shifter is well matched over the entire operating frequency band and the return and insertion losses present good performances at different polarization voltage values. In fact, the return loss remains below -10 dB in the whole operating band when the voltages change from 3 V to 7 V, as illustrated in Fig. 3(a). In addition, the insertion loss reaches the maximum value of -0.57 dB while the voltages exceed 4 V along the Ku band in accordance with Fig. 3(b). Furthermore, as can be seen in Fig. 4, a linear variation of the simulated phase shift is obtained for all frequencies by altering the control voltages from 3 V to 7 V. Our phase shifter achieves a phase shift variation of 112.88° and 168° at 10.7 and 12.7 GHz, respectively.



(a) |S11|



(b) |S21|

Fig. 3. Simulated S Parameters of the Tunable Phase Shifter for Different Control Voltages.

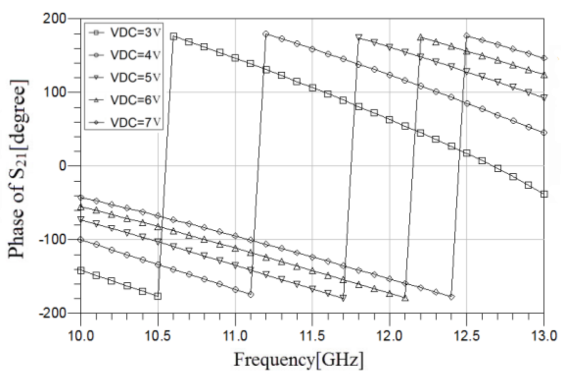
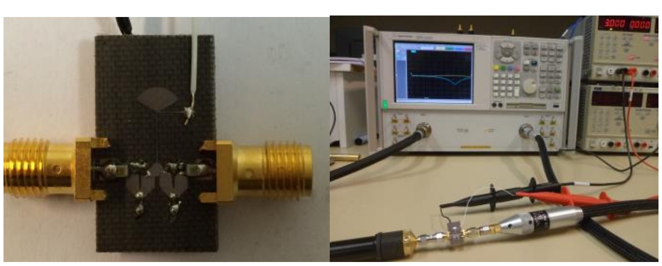


Fig. 4. Simulated Phase Shifts of the Tunable Phase Shifter for Different Control Voltages.

The performances of the tunable phase shifter are qualified regarding its beam steering capability. This latter is characterized by the beam steering angle which depends on the phase shift difference conforming to the following equation:

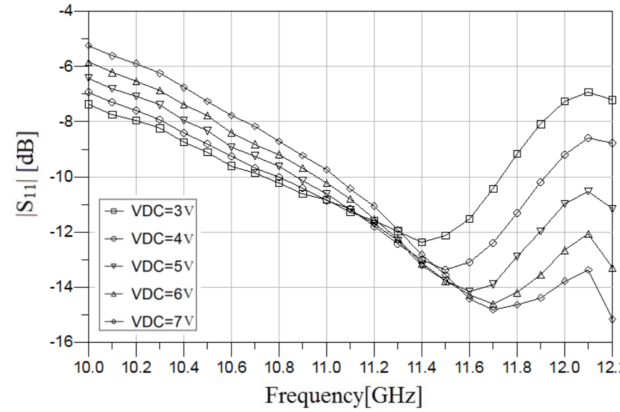
$$\theta = \sin^{-1}\frac{-\Delta\phi\,\lambda}{2\pi d} \quad (1)$$

Where θ is the beam steering angle, d is the distance between two antennas, λ is the wavelength (C/F) and Δϕ is the phase shift difference. When this difference of phase shift is created by the array of two antenna distant by 23 mm, the beam of the antenna array is steered continuously by 22.4° and 28.6° at 10.7 and 12.7 GHz respectively, by adjusting the control voltage throughout the Ku band. To validate the functioning concept, the phase shifter is manufactured and tested. Fig. 5 illustrates the phase shifter prototype.

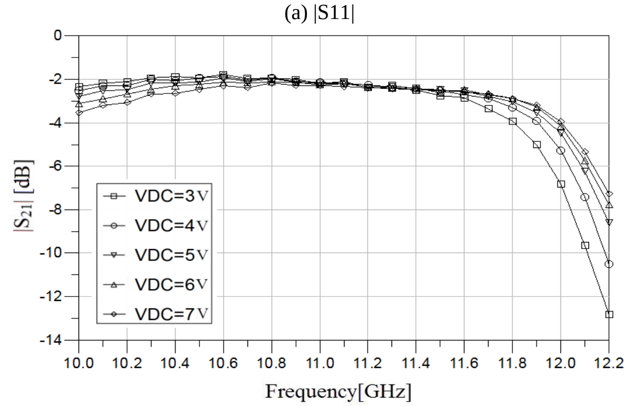


(a) Manufactured Prototype. (b) Prototype Under Test.

Fig. 5. Photograph of the Two-Stage Tunable Phase Shifter.



(a) |S11|



(b) |S21|

Fig. 6. Measured S Parameters of the Tunable Phase Shifter for Different Control Voltages.

Measurements were performed using a network analyzer and the measured S-parameters are provided in Fig. 6 for different control voltages. By adjusting the voltages from 3 V to 7 V, the return loss is less than -10 dB over the frequency band from 11 GHz to 11.7 GHz, as presented in Fig. 6(a). Furthermore, from Fig. 6(b), the maximum insertion loss is around -2 dB over the bandwidth extending from 10.7 GHz to 11.7 GHz for all voltage values.

Fig. 7 displays the measured phase shifts of the proposed phase shifter for different control voltages.

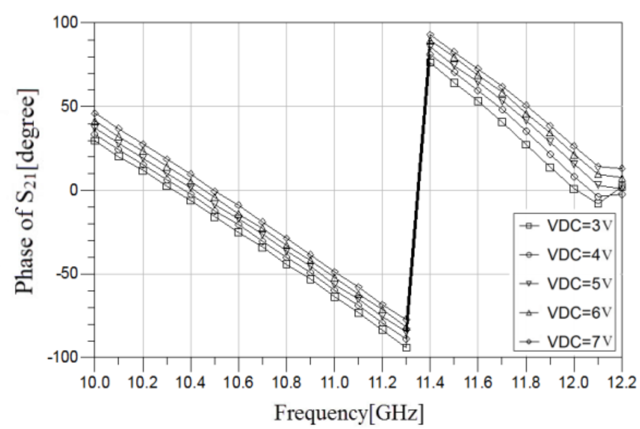


Fig. 7. Measured Phase Shifts of the Tunable Phase Shifter for Different Control Voltages.

From this figure, it can be seen that the phase shift variation of this phase shifter follows the variation of the voltages along the operating bandwidth. Also, when the voltages move from 3 V to 7 V, we can observe that the measured difference phase shifting range is about 15° and 21.13° at 10.7 and 11.7 GHz, respectively.

Inserted in the feed network of two antennas distant by 23 mm, the phase shifter can steer the beam electronically by 3° and 3.75° at 10.7 and 11.7 GHz respectively, by tuning the control voltages. Unfortunately, the measurement results are not close to the simulation results. This degradation is due to the effects of the soldering temperature or exposure time during the diode mounting step.

Despite the limitations of the manufacturing process, the continuous beam steering behavior is approved by adjusting the diode voltage even if the phase shift variation is very low. The future works need to focus on improving the phase variation range of the phase shifter to achieve the full cycle azimuthal scanning when integrated into a large phased antenna array.

The characteristics of the proposed phase shifter were compared to recently issued phase shifters. Only the phase shifters made by the printed circuit board with varactor diodes were taken into consideration. The comparison is outlined in Table II.

TABLE II. PERFORMANCE COMPARISON OF VARACTOR-BASED PHASE SHIFTERS

Ref	Num. diodes	Num. section	Freq (GHz)	Max. IL (dB)	Max. PS (°)	Approach
[17]	6	3	11.5-12.5	2.3	360	RTPS
[18]	6	2	6.7-7.7	4.2	380	All-pass network
[19]	4	2	1.8-2.6	1.5	380	RTPS
This paper	6	2	10.7-12.7	0.57	168	Schiffman PS

Num: number, IL: insertion loss, PS: phase shift, RTPS:Reflection-type phase shifters

Compared to the reported phase shifter performing in the Ku band [17], our phase shifter has the most compact size allowing it to be inserted in a large antenna array for beam steering applications. Indeed, the reflection-type phase shifter have a bulky structure because it consists of a 3 dB/90° coupler combined with reflective circuits [17]. Other structures [18, 19], which are based on coupled lines section charged by varactor diodes, are limited to providing a large phase shift variation across a narrowband at low frequencies.

The above features indicate that our phase shifter offers a compactness and a phase variation throughout 2 GHz of bandwidth in Ku band even with a short phase shift range.

IV. CONCLUSION

In this article, an electronically reconfigurable phase shifter operating in Ku band is designed, fabricated and measured. The proposed design, with the reference line removed, is based on the combination of two Schiffman sections embedded with varactor diodes. The two-stage phase shifter yields a continuously adjustable phase shift over a wide frequency range with just one control voltage. The cascade structure provided in the phase shifter circuit results in the increase of the phase shift variation keeping the compactness of the phase shifter. The measured results clearly validate the tunability of the proposed phase shifter.

ACKNOWLEDGMENT

The deepest thanks are dedicated to the staff of Polytech Nantes of the University of Nantes, France for their support and assistance in the manufacture of the designed phase shifters and for having made the measurements possible.

REFERENCES

[1] P. Sobis, J. Stake, and A. Emrich,“High/low-impedance transmission-line and coupled-line filter networks for differential phase shifters,” IET Microw. Antennas Propag., vol. 5, no. 4, pp. 386-392, 2011.

[2] S. H. Yeung, Z. Mei, T. K. Sarkar, and M. Salazar-Palma, “Design and testing of a single-layer microstrip ultrawideband 90° differential phase shifter,” IEEE Microw. Wirel. Compon. Lett., vol. 23, no.3, pp. 122-124, March 2013.

[3] Y. Wang, M. E. Bialkowski, and A. M. Abbosh,“Double microstrip-slot transitions for broadband ±90° microstrip phase shifters,” IEEE Microw. Wirel. Compon. Lett., vol. 22, no. 2, pp. 58-60, February 2012.

[4] J. L. R. Quirarte, and J. P. Starski, “Novel Schiffman phase shifters,” IEEE Trans. Microw. Theory Tech., vol. 41, no. 1, pp. 9–14, January 1993.

[5] D. Chai, M. Linh, M. Yim, and G. Yoon, “Asymmetric teflon-based Schiffman phase shifter,” Electron. Lett., vol. 39, no. 6, pp. 529-530, March 2003.

[6] Y. -X. Guo, Z. -Y. Zhang, and L. C. Ong,“Improved wideband Schiffman phase shifter,” IEEE Trans. Microw. Theory. Techn., vol. 54, no. 3, pp. 1196–1200, March 2006.

[7] W. Zhang, Y. Liu, Y. Wu, W. Wang, M. Su, and J. Gao,“A modified coupled-line Schiffman phase shifter with short reference line,” Prog. Electromagn. Res. C, vol. 54, pp. 17–27, 2014.

[8] H. Oraizi,and A. Shamsafar,“Optimum design of modified Schiffman multi-section wide band differential phase shifter with impedance matching,” Prog. Electromagn. Res. C, vol. 29, pp. 1–16, 2012.

[9] H. Zhu, and A. M. Abbosh,“A compact tunable directional coupler with continuously tuned differential phase,” IEEE Microw. Wirel. Compon. Lett., vol. 28, no. 1, pp. 19-21, January 2018.

[10] L. Guo, H. Zhu, and A.Abbosh, “Wideband phase shifter with wide phase range using parallel coupled lines and L-shaped networks,”IEEE Microw. Wirel. Compon. Lett., vol. 26, no. 8, pp. 592-594, August 2016.

[11] M. Nikfalazar, C. Kohler, A. Wiens, A. Mehmood, M. Sohrabi, H. Maune, J. R. Binder, and R. Jakoby, “Beam steering phased array antenna with fully printed phase shifters based on low-temperature sintered BST-composite thick films,”IEEE Microw. Wirel. Compon. Lett., vol. 26, no. 1, pp. 3511-3518, January 2016.

[12] M. Sazegar, Y. Zheng, H. Maune, Ch. Damm, X. Zhou, and R. Jakoby,“Compact tunable phase shifters on screen-printed BST for balanced phased arrays,”IEEE Trans. Microw. Theory Techn., vol. 59, no. 12, pp. 3331-3337, December 2011.

[13] E. Öjefors, S. Cheng, K. From, I. Skarin, P. Hallbjörner, and A. Rydberg,“Electrically steerable single-layer microstrip traveling wave antenna with varactor diode based phase shifters,” IEEE Trans. Antennas Propag., vol. 55, no. 9, pp. 2451-2460, September 2007.

[14] R. Ouali, L. Osman, T. Razban, and Y. Mahé, “Compact Tunable Ku band Schiffman Phase Shifter for Beam-Steering Applications,” 5th International Conference on Multimedia Computing and Systems (ICMCS), Marrakech, Morocco, Septembre 2016, pp. 405-409.

[15] R. Wali, L. Osman, T. Razban, and Y. Mahé, "Tunable Schiffman Phase Shifter for Continuous Beam Steering Antenna," International Conference on Internet of Things, Embedded Systems and Communications (IINTEC), Gafsa, Tunisia, October 2017, pp. 171-176.

[16] C. Ding, Y. J. Guo, P. -Y. Qin, and Y. Yang,"A compact microstrip phase shifter employing reconfigurable defected microstrip structure (RDMS) for phased array antennas," IEEE Trans. Antennas Propag., vol. 63, no. 5, pp. 1985-1996, 2015.

[17] P. Padilla, J. F. Valenzuela-Valdés, J. L. Padilla, J. M. Fernández-González, and M. Sierra-Castañer,"Electronically reconfigurable reflective phase shifter for circularly polarized reflectarray systems," IEEE Microw. Wireless Compon. Lett., vol. 26, no. 9, pp. 705-707, September 2016.

[18] K. Khoder, M. Le Roy, and A. Pérennec,"An all-pass topology to design a 0-360 continuous phase shifter with low insertion loss and constant differential phase shift," Ninth European Microwave Integrated Circuits Conf. (EuMIC), Rome, Italy, Oct. 2014, pp. 612-615.

[19] A. M. Abbosh,"Compact tunable reflection phase shifters using short section of coupled lines," IEEE Trans. Microw. Theory Techn., vol. 60, no. 8, pp. 2465-2472, August 2012.

[20] Yue Li, M. F. Iskander, Z. Zhang, and Z. Feng,"A new low-cost leaky wave coplanar waveguide continuous transverse stub antenna array using metamaterial-based phase shifters for beam steering,"IEEE Trans. Antennas Propag., vol. 61, no. 7, pp. 3511-3518, July 2013.

[21] J. Zhou, H. J. Qian, and X. Luo,"Compact wideband phase shifter using microstrip self-coupled line and broadside-coupled microstrip/CPW for multiphase feed-network," IEEE Microw. Wirel. Compon. Lett., vol. 27, no. 9, pp. 791-793, September 2017.

Social Networking Sites Habits and Addiction Among Adolescents in Klang Valley

Yazriwati Yahya[1*], Nor Zairah Ab. Rahim[2], Roslina Ibrahim[3], Nurazean Maarop[4], Haslina Md Sarkan[5], Suriayati Chuprat[6]

Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Jalan Sultan Yahaya Petra, 54100, Kuala Lumpur, Malaysia

***Abstract*—Social networking sites (SNS) is a very popular application in today's world society. SNS, to certain extent has change the way people communicate with each other. This kind of technology has become a trend among the users regardless the impact of the technology to the users either positive or negative. The level of SNS usage among the adolescents has started to raise concern among the parents and also the society. SNS addictions are becoming problematic in certain countries especially in United States and lately this issue has started to spread all over the world. Malaysia is also one of the country affected with SNS addiction. SNS addiction is not an isolated phenomenon as it is started from high engagement on the SNS usage and it originates from habitual behavior. Therefore, it is important to seek and understand habit and addiction of SNS among adolescents in Malaysia. The purpose of this study is to analyze and explore the usage of SNS among the adolescents in Malaysia, specifically in Klang Valley. It examines the SNS usage behavioral, which is habit and addiction. The data was collected from a sample of 60 respondents using an online survey. The data were analyzed using SPSS for descriptive analysis. From the analysis, it was found that most of the adolescents used SNS in daily basis and majority of them use it for more than two hours per day. Patterns on habits and addiction on the SNS usage shows that some adolescents experienced certain habit and addiction behavior.**

***Keywords*—Social networking sites; habit behavior; addiction behavior; SNS usage**

I. Introduction

The Internet World Statistics [1] demonstrates that Malaysia is among the 10 Asia Top Internet Countries with more than 20 million Internet users. As indicated by Malaysian Communications Multimedia Commission (MCMC), the Internet Addiction (IA) and digital security issue are influencing youths in urban territories, as well as understudies in rustic zones, particularly in schools where free portable workstations and Internet access are given. Other than influencing their psychological well-being, the Internet fixation has transformed the physical activity exercises into virtual activity among the youngsters. Fixations or preoccupations with the Internet encourages a hazardous kind of addiction today. According to psychiatric consultant Dr Muhammad Muhsin Ahmad, Deputy Chief Coordinator at Centre of Addiction Sciences, Universiti Malaya, [2] "Malaysia has high chances of developing a segment society that are Internet addicts in the near future".

Recently, SNS usages, which are part of the Internet fixation, have captured the scholar's attention. In Digital SouthEast Asia 2017 [3] report, the total number of active users of social media in Malaysia has reached 22 million. According to Shin & Ismail [4] Malaysian youth are active SNS users spending an average of 19 hours online per week, and 20% of Facebook users of the country are younger than 18. While most social media platforms set a minimum age of 13 to sign up on their sites, Cyber-Security Malaysia's nationwide survey of over 8,000 primary and secondary students found that almost half of the pupils aged between 7 and 9, have social media accounts. According to a local newspaper, Sunday Star in its report on Oct 8, 2017, this percentage went up to 67% for children aged between 10 and 12.

Many people tend to check their SNS from the moment they woke up early in the morning. Information comes straight to us without our hassle to find them. This shows that how dependent our society to the SNS. The engagement in SNS has shown some positive implication such as easy and fast communication with friends and family; able to update people's life and maintain social relationship; and also support positive behaviours such as self-promotion and self-disclosure [5]. Nevertheless, some people spend too much time on the SNS platforms and experience the negative outcome without them realizing it [6]. The positive and negative outcomes of the SNS usage are resulted from the habitual behaviour. A stronger SNS habit may lead to pathological problems such as addiction. Researchers also have suggested that the excessive use of SNS has started to be one of major concern among the adults and also may be particularly problematic to youngsters as well [7].

SNS addiction may contribute negative effect to adolescents. Perceived low self-esteem which then lead to low well-being is one of the effects due to the negative comments from friends about their SNS profile [8]. Lower grades in academic as a result of poor time management for adolescents who spent more time on the SNS compared to those who did not use SNS [9]. Apart from that, other negative behavior such as stealing information, cyber bulling, contacting strangers and also spread false information was also triggered to be the causes the SNS addiction[10][11]. An article in The Star online newspaper dated January 5th 2015 has highlighted that Malaysians engaged for long hours on their SNS rather than meeting their friends' offline. They also tend to share everything about themselves on the SNS rather than having actual conversation with their families and friends. They did not realize that SNS have slowly destroyed their rapport with the families and friends. SNS also affect the academic growth among students since much of their time is spent staring at the screen [12]. Such behaviors may cause negative effects that will lead to further serious problems.

The aim of this paper is to focus on the SNS usage behavioural which is habit and addiction experience by the adolescents in Klang Valley area. Klang Valley is centered in the city of Kuala Lumpur in Malaysia and incorporates its connecting urban areas and towns in the province of Selangor. The paper will discuss on the SNS usage among the adolescents and also compare the SNS usage behavioural between age categories.

II. Related Works

A. Platform of Social Networking Sites

According to statistic portal on social media and networking sites in United States (US) [13], Snapchat, Facebook and Instagram are categorized to be the mostly used SNS platform among the adolescents in 2017 (see Fig. 1).

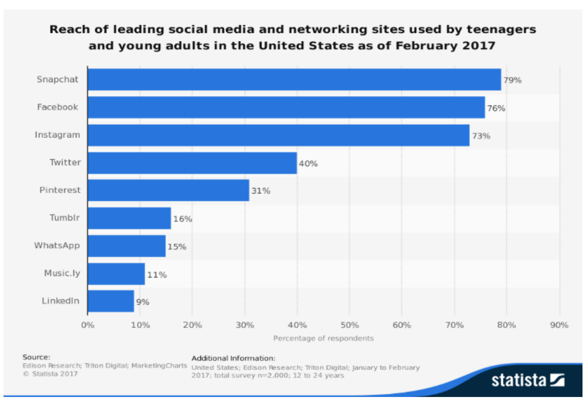


Fig. 1. Social media and networking sites used by adolescents in US [13].

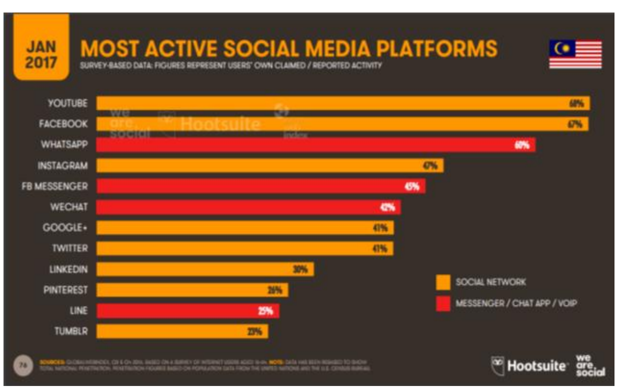


Fig. 2. The most active SNS and messenger application used in Malaysia [3].

In general, YouTube has dominated the SNS platforms used by Malaysians regardless adults or adolescents as depicted in Fig. 2.

Fig. 3 and 4 show the analysis of Facebook usage and Facebook user's profile. Even though the percentage of users' aged 13 – 17 years in Malaysia is not as high as other age, it still contributes to the percentage of SNS use.

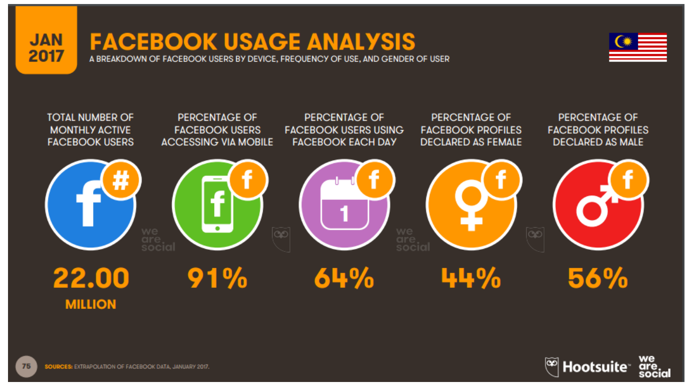


Fig. 3. Facebook usage analysis in Malaysia [3].

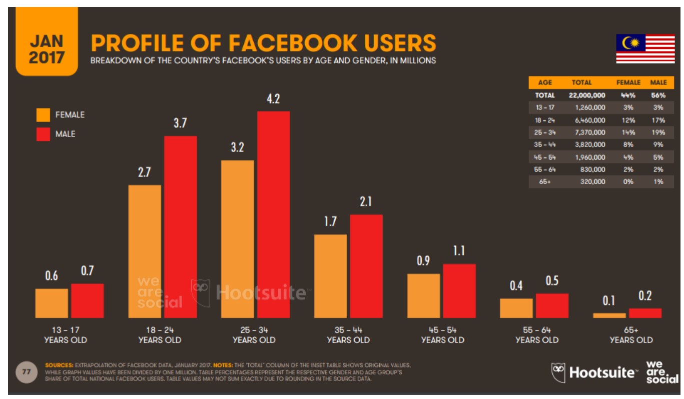


Fig. 4. Profile of Facebook users in Malaysia according to age and gender [3].

Professor Datuk Dr Chiam Heng Keng [14], a leading figure in child and adolescent psychology and early childhood education, said reports about adolescents turning to social media and destructive online games such as the Blue Whale game, which purportedly encouraged adolescents to end their lives as a way to solve their problems, is worrying. According to her, in today's society, many teenagers spend most of their times on the smartphone and Internet. This also has resulted the lack of ability for them to communicate face to face with other people. They probably cannot communicate with their parents, teachers or peers about their stress and fears, and are so desperate to be heard that they resort to the Internet. She said that the teenagers may feel frustrated with their lives, involved in love affairs and facing exams stress, making them believe that the help is available from their online friends.

B. Habit Formation

Based on prior research, the concept of habit is not new [15]. According to [16], habit has more influences when foreseeing repeated behavior rather than variables like intention or attitudes. Habit is a type of automatic action that needs deliberateness, mindfulness, and controllability, despite the fact that its effectiveness is high [17]. However, even though habit used to describe as a behaviour, it is actually not a behaviour in itself as it is a mental state that drives a person to perceive habit-related stimulus cues in order to achieve certain goals [16][18]. Hence, habitual behaviour is the action resulting from the cue. These cues can be times, places, specific situations, moods, goals, etc. [19]. For example, the tendency of a person's behavioural to check their SNS from the mobile phone after waking up. In the literature, habit and past behaviour were used equally. Past behaviour may turn into habit once it is successfully repeated over the time [20].

Habit is the outcome from the automatic process where once a habit is shaped; the individual will have less attention on their behavior performance and most likely to be involved with non-reflective cognitive processing [21][22]. The existing studies of habit are more focused on how IS habit is theorized and measures, plus with its relationship with continuance intention and IS continuance use [23][18]. Overall, IS researcher have come to one conclusion that if individuals are habitually performing a behavior such as using a SNS, the future behavior which is continued using SNS will be mostly controlled by habit instead of reasoned action.

The formation of habit was based on the action done repetitively without consciousness for a specific timeframe and normally for long term of period. It includes characteristic such as unintentional, uncontrollable, lack of awareness and efficiency [24]. Accordingly, people characterize their habits as an intentional sequence of behavior where to a certain period it is controllable, executed with less awareness and efficient [16].

C. SNS Addiction Formation and Symptoms

According to [25], the formations of SNS addiction are based on three theoretical perspectives: (1) the cognitive behavioral model; (2) the social skill model; (3) socio-cognitive model. The cognitive behavioural model suggests that some SNS users can develop nonadaptive apprehension that could be caused by various environmental factors. For example, social isolation or lack of peer support, and lead to the development of nonadaptive obsessive use patterns. The social skill model highlights that people can be lacked of self-presentational skills if they prefer online communication rather than face to face interactions. The socio-cognitive models emphasize that the obsessive behaviors are resulted from expectation of positives results combine with self-usefulness and weak self-control. [26] used this formation to prove that those who encounter transition from normal SNS use to problematic SNS use are usually use SNS to relieve stress, loneliness or depression. According to [27], frequent SNS users normally are not good at interacting face to face. Since SNS provides satisfaction and self-efficacy, these people tend to use SNS as frequent as they can which then causes many other problems such as ignoring relationship with family and friends or problems with study and at work.

Griffiths believes that any behavior that fulfill six components of addiction behavior or also known as addiction symptoms can be operationally defined as addiction [28]. The six symptoms are:

1) Salience: Salience happens when even if we do not use SNS at the moment, we are persistently keep on thinking about using it and anxiously waiting to use it as soon as possible. This reflects when SNS has becoming the most important agenda in our lives that dominates not just our thinking, but feelings and behavior as well.

2) Mood modification: Changing on the mood such as feelings of escape or upset as a result of consequence of social networking and can be seen as coping strategy.

3) Tolerance: In order to accomplish the previous mood modification effect, a huge amount of social networking activities are needed. This basically means that for people engaged in social networking, they gradually build up the amount of the time they spend social networking every day.

4) Withdrawal symptoms: When people fail to disengage with the social network because of certain reason, unpleasing feeling such as an adverse mood and irritations occurs.

5) Conflict: This refers to the conflicts between a person and those around that person (interpersonal conflict), conflicts with other activities (social life, hobbies, and interests), or from within the individual him or herself (e.g., subjective feelings of loss of control) that are concerned with spending too much time social networking.

6) Relapse: This is the tendency for repeated reversions to earlier patterns of excessive social networking to recur and for even the most extreme patterns typical of the height of excessive social networking to be quickly restored after periods of control.

These symptoms were used as benchmark by the psychology and clinical studies to identify the existence of SNS addiction among the SNS users. The addiction symptoms are very important to be recognized because if it is not to be treated will established negative effects among the SNS users.

Despite being a current topic, dependency on SNS, which may lead to excessive use usually, goes unnoticed by family members [29]. Even though SNS addiction studies have started to get scholars attention at the beginning of its introduction, but most of them were focused on adult users. SNS addiction studies on adolescents are still at early stage [10]. Studies on SNS Addiction among adolescents are more focused on the usage patterns [30], factors of SNS addiction [31] and negative consequences for being addicted [10] and all these studies concentrated on psychological and clinical perspective. Whilst, studies on SNS addiction from the perspective of hedonic information systems (IS) have emphasized more on the impact of addiction on SNS continuance intention of use [32][33].

III. METHODOLOGY

The methodology in this study was divided into three sections which are the participants, instrument and data analysis. The result of this particular study is descriptive in nature.

A. Participants

An online survey were created using Google form and link were distributed among parents in Klang Valley. Data were collected for 3 weeks and 60 responded answered were captured. The sample comprised 70% female and 30% male.

B. Instrument

Data were collected by means of an online survey developed by the researchers. The survey consisted of three sections. The first section contains on demographic information of the adolescents. The second section examines the use of SNS such as how long and how often adolescents view and access SNS and also, length of time spent in SNS. The third section was consisted of items related to SNS Usage behavioral such as habit and addiction. The addiction scales were extracted from the [34] study while the habit scales were retrieved from [23] study. The Cronbach's alpha reliability coefficient of this scale was 0.61 for addiction scale and 0.74 for habit.

C. Data Analysis

Data were collected using the developed scale. The data obtained by the survey was analysed using the SPSS program with the percentage, frequency, and statistical analysis techniques.

IV. Findings and Discussions

A. Demographic Analysis

A several demographic questions were used in the preliminary online survey and 60 feedbacks were recorded. Based on the results, a demographic profile data are retrieved. The respondents were categorized into two group of age, which are: Below 19 years and above 19 years. Respondents above 19 years are categorized as late adolescents or young adults. The survey was responded 70% by female and 30 % by male (see Table I). Based on the analysis, 90% of the respondents used the SNS more than 4 years as displayed in the table and they accessed and viewed the SNS more on daily basis (Table II), which means that they apparently are quite familiar in using the SNS platform.

TABLE I. Demographic Analysis

Variables	**N Respondents (N = 60)**	**Percentages (%)**
Gender		
Female	42	70
Male	18	30
Age		
< 19 years	41	32
> 19 years	19	68
Start Using SNS		
< 6 months	2	0
1- 2 years	4	7
> 3 years	54	93

B. Time Spent on SNS

The average time of spending on the SNS is more than two hours (see Table II). This result is in line with study by [4] that Malaysian youth are actually are active SNS users spending an average of 19 hours online per week.

TABLE II. SNS Usage

Variables	**N Respondents (N=60)**	**Percentages (%)**
Access & View		
Daily	53	88
Weekly	6	10
Monthly	1	2
Average Time Spend per day		
< 30 minutes	17	28
1 to 2 hour	20	33
> 2 hours	23	39
The Most Used SNS (Top 4)		
Instagram		30
Facebook		21
Twitter		15
Snapchat		10

The most popular SNS platform used based on this analysis is Instagram (30%), followed by Facebook (21%) and Twitter (15%) as illustrated in Fig. 5. This is in line with Fig. 1 shows the most popular in United States (US) among adolescents are Snapchat, Facebook and Instagram. Even though the usage Snapchat platform in Malaysia is not as popular as in US, however, Facebook and Instagram are still on top of the list among Malaysia which is similar with US.

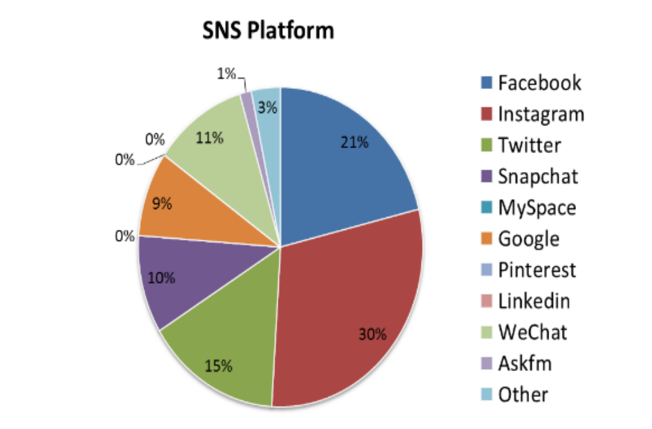


Fig. 5. The most used SNS platform.

C. SNS Habit

Three questions on SNS habit were highlighted in the survey which were adopted from [23]. The questions are:

HAB1. Using this SNS has become automatic to me.

Automatic here means something is done or occurring spontaneously, without thinking. Example given in the survey is “Once I reach home after school, automatically I will access my social networking sites”.

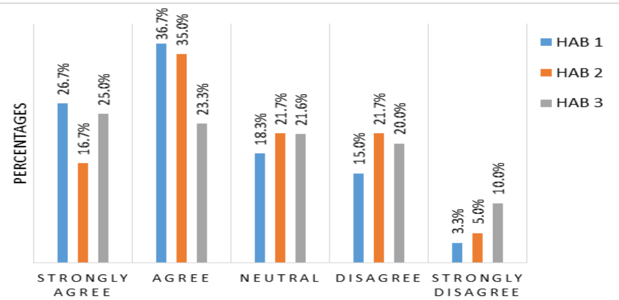
HAB2. Using this SNS is natural habit to me.

When using the SNS, it is naturally done without forcing.

HAB3. *When I want to interact with friends and relatives, using SNS is an obvious choice for me.*

This means that SNS is main medium communication that will be used when it comes to communicate with friends and relatives.

Fig. 6 shows the result of mean, median and mod values together with percentage of habit item for construct habit. Items ranged from Strongly Disagree (1) to Strongly Agree (5). Based on the overall result, 63.4% agreed that using SNS is done automatically without hesitation. 51.7% also agreed that using SNS is natural for them. As for SNS is an obvious choice for main medium communication, 48.3% agreed that it is an obvious choice for medium interaction while 30.0% does not agree and 21.7% sometimes agreed. There are possibilities that they are using other medium communication besides SNS. The results suggested that about 50% to 60% of respondents found SNS is a natural to them and became a habit in their daily life.



	Mean	Median	Mod
HAB1	3.68	4.00	4
HAB2	3.37	4.00	4
HAB3	3.33	3.00	5

Fig. 6. Result of overall descriptive analysis on SNS habit

TABLE III. RESULT OF DESCRIPTIVE ANAYLSIS BETWEEN AGE AND HAB1 (AUTOMATIC USE)

HAB1							
Age		***Strongly Disagree***	***Disagree***	***Neutral***	***Agree***	***Strongly Agree***	***Total***
< 19 years	*f*	0	7	8	16	10	41
	%	0.0	17.1	19.5	39.0	24.4	100.0
> 19 years	*f*	2	2	3	6	6	19
	%	10.5	10.5	15.8	31.6	31.6	100.0
	Total *f*	2	9	11	22	16	60
	Total %	3.3	15.0	18.3	36.7	26.7	100.0

**f* = frequency

Table III shows the cross tabulation between age categories with HAB1 (Automatic use) while Table IV shows the cross tabulation between age categories with HAB2 (Natural habit). From the tables, most of the respondents either below 19 years (63%) or above 19 years (63%) agreed that the SNS use is automatically or spontaneously done and has becoming a natural habit. Only 7 out of 41 respondents disagree that use of SNS is automatic to them while for respondents above 19 years, 4 out of 19 respondents disagree with the same statement. When behaviour becomes habitual, users tend to use the system automatically without going through the cognitive planning process.

TABLE IV. RESULT OF DESCRIPTIVE ANAYLSIS BETWEEN AGE AND HAB2 (NATURAL HABIT)

HAB2							
Age		***Strongly Disagree***	***Disagree***	***Neutral***	***Agree***	***Strongly Agree***	***Total***
< 19 years	*f*	1	8	12	15	5	41
	%	2.4	19.5	29.3	36.6	12.2	100.0
> 19 years	*f*	2	5	1	6	5	19
	%	10.5	26.3	5.3%	31.6	26.3	100.0
	Total *f*	3	13	13	21	10	60
	Total %	5.0	21.7	21.7	35.0	16.7	100.0

TABLE V. RESULT OF DESCRIPTIVE ANAYLSIS BETWEEN AGE AND HAB3 (OBVIOUS CHOICE)

HAB3							
Age		***Strongly Disagree***	***Disagree***	***Neutral***	***Agree***	***Strongly Agree***	***Total***
< 19 years	*f*	4	10	12	5	10	41
	%	9.8	24.4	29.3	12.2	24.4	100.0
> 19 years	*f*	2	2	1	9	5	19
	%	10.5	10.5	5.3	47.4	26.3	100.0
	Total *f*	6	12	13	14	15	60
	Total %	10.0	20.0	21.7	23.3	25.0	100.0

**f* = frequency

As for Table V, the percentages for adolescents below 19 years show an even spread of frequency between agreed (36.6%), not agreed (34.4%) and sometimes (29.3%). Contrasted to the frequency of adolescents below 19 years with above 19 years, the latter apparently seems to agree that SNS is their main medium communication to interact with family, friends and relatives. It could be related to Table VI where a cross tabulation between age with time spend on SNS shows 57.9% late adolescents use SNS more than two hours in daily basis (refer Table II for frequency of daily basis). They spend more time on SNS which may indicate that SNS is a preferred medium for them to interact with family and friends. This is in line with [32] study where they have found that late adolescents adopt SNS platforms in order to (1) communicate with friends who they hardly meet, (2) encouragement from friends who have SNS account, (3) get connected with relatives and families and also (4) planning regular meetings with friends.

TABLE VI. RESULT OF DESCRIPTIVE ANAYLSIS BETWEEN AGE AND TIME SPEND ON SNS

			Time spend on SNS			Total
			> 30 minutes	***1 - 2 hour***	***< 2 hours***	
AGE	< 19 years	*f*	13	16	12	41
		%	31.7	39.1	29.3	100.0
	> 19 years	*f*	4	4	11	19
		%	21.1	21.0	57.9	100.0
		Total *f*	17	20	23	60
		Total %	52.8	60.1	38.3	100.0

*f = frequency

D. SNS Addiction

This section discusses the findings for SNS addiction part. Six questions on SNS addiction were highlighted in the survey which were adopted from [34].

ADT1. I spend a lot of time thinking about social networking sites or planning to use it. (Salience)

ADT2. I urge to use social networking sites more and more. Urge means strong desire to use social networking sites. (Tolerance)

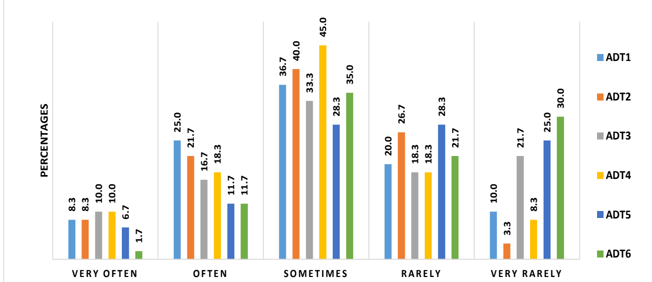
ADT3. I use social networking sites in order to forget about personal problem. (Mood modification)

ADT4. I tried to cut down on the use of social networking sites without success. (Relapse)

ADT5. I become restless or troubled if I have been prohibited from using social networking sites. (Withdrawal)

ADT6. I used social networking sites so much that it has had a negative impact on my studies. (Conflict)

Each question portrays the six core symptoms of addiction such as salience, tolerance, mood modification, relapse, withdrawal, and conflict. Fig. 8 demonstrate the result of overall descriptive analysis on construct SNS addiction with items ranged from Very Rarely (1) to Very Often (5). Based on the table, salience (ADT1), tolerance (ADT2) and relapse (ADT4) show an even spread of percentages between rarely and often. However, mood modification (ADT3), withdrawal (ADT5) and conflict (ADT6) show a slightly different patterns from the former three symptoms. The respondents are actually rarely and only sometimes experienced the symptoms. For withdrawal, 53% of total respondents rarely use SNS to forget personal problems, and only 28.3% sometimes felt that SNS is actually a space of escapism to forget about personal problems. While, 51.7% of respondents rarely thought that SNS actually gave a negative impact on their studies. However, though the often scale of percentages for six symptoms is not that high compared to rarely scale of percentages, majority of the respondents chose sometimes scale for each of the symptoms. Sometimes may lead to often if SNS is used excessively without control.



	Mean	Median	Mod
ADT1	3.02	3.00	3
ADT2	3.05	3.00	3
ADT3	2.75	3.00	3
ADT4	3.03	3.00	3
ADT5	2.47	2.00	2
ADT6	2.33	2.00	3

Fig. 7. Result of Overall Descriptive Analysis on SNS Addiction.

Table VII until Table X further analyze the SNS addiction symptoms according to the levels of age. By differentiate between levels of age (below 19 years and above 19 years), an added possibility may be projected. Table VII shows a cross tabulation result between salience and age. Salience means even when they are not engaging with the SNS, they will continuously keep on thinking about using it [28]. Adolescents below 19 years seem to either rarely (34.1%) or sometimes (43.9%) thinking about using SNS or planning on using it compared to those who are above 19 years where 57.9% of the respondents are actually often thought about using and planning to use the SNS. The same occurred for tolerance as depicted in Table VIII, where adolescents below 19 years rarely (39%) or sometimes (41.5%) have the ability to tolerate in spending huge amount of time using the SNS. This is different than adolescents above 19 years where 52.6% of them often and 36.8% sometimes have strong desire to spend more time using the SNS.

TABLE VII. CROSS TABULATION BETWEEN AGE AND ADT1 (SALIENCE)

ADT1							
Age		***Very Rarely***	***Rarely***	***Sometimes***	***Often***	***Very Often***	***Total***
< 19 years	*f*	3	11	18	6	3	41
	%	7.3	26.8	43.9	14.6	7.3	100.0
>19 years	*f*	3	1	4	9	2	19
	%	15.8	5.3	21.1	47.4	10.5	100.0
	Total *f*	6	12	22	15	5	60
	Total %	10.0	20.0	36.7	25.0	8.3	100.0

*f = frequency

TABLE VIII. CROSS TABULATION BETWEEN AGE AND ADT2 (TOLERANCE)

ADT2							
Age		***Very Rarely***	***Rarely***	***Sometimes***	***Often***	***Very Often***	***Total***
< 19 years	*f*	2	14	17	5	3	41
	%	4.9	34.1	41.5	12.2	7.3	100.0
>19 years	*f*	0	2	7	8	2	19
	%	0.0	10.5	36.8	42.1	10.5	100.0
	Total *f*	2	16	24	13	5	60
	Total %	3.3	26.7	40.0	21.7	8.3	100.0

**f* = frequency

Relapse as can be seen in Table IX shows the percentages of respondents above 19 years where some of them have tried to make several attempts to reduce the usage of SNS but fail to do so (sometimes: 31.6% and often: 31.6%). In contrast, different pattern appears in conflict as shown in Table X. It shows that adolescents from both levels of age do not think that using SNS have negative impact on their studies.

In general, the patterns show that most of the respondents, regardless their age, *sometimes* experience the six symptoms when using the SNS. Comparing the results of respondents age of 19 years old and above with those below 19 years old, it shows that adolescents below 19 years old have experienced all the six core symptoms either rarely or sometimes compared to those above 19 years old where majority of them either sometimes or often experience the core symptoms. Even though most of the respondents are more likely to seldom experience the symptoms while using the SNS, what is more concern is actually on the ones that are frequently experience addiction symptoms. Based on the data value, 5 to 6 respondents are actually encountering the symptoms (they have chosen either often scale or very often scale at each of the addiction behavior) and it is recommended to further discuss before it prevails any negative impact to the adolescents.

TABLE IX. CROSS TABULATION BETWEEN AGE AND ADT4 (RELAPSE)

ADT 4							
Age		***Very Rarely***	***Rarely***	***Sometimes***	***Often***	***Very Often***	***Total***
< 19 years	*f*	4	7	21	5	4	41
	%	9.8	17.1	51.2	12.2	9.8	100.0
>19 years	*f*	1	4	6	6	2	19
	%	5.3	21.1	31.6	31.6	10.5	100.0
	Total *f*	5	11	27	11	6	60
	Total %	8.3	18.3	45.0	18.3	10.0	100.0

**f* = frequency

TABLE X. CROSS TABULATION BETWEEN AGE AND ADT6 (CONFLICT)

ADT6							
Age		***Very Rarely***	***Rarely***	***Sometimes***	***Often***	***Very Often***	***Total***
< 19 years	*f*	12	12	12	4	1	41
	%	29.3	29.3	29.3	9.8	2.4	100.0
> 19 years	*f*	6	1	9	3	0	19
	%	31.6	5.3	47.4	15.8	0.0	100.0
	Total *f*	18	13	21	7	1	60
	Total %	30.0	21.7	35.0	11.7	1.7	100.0

**f* = frequency

These findings might further indicate that there are adolescents who might experience some of the habit or addiction symptom on the SNS usage even though it is only minor. Overall result shows that older adolescents (aged above 19 years) have developed habit on using SNS compared to younger adolescents (aged below 19 years). The similar patterns were seen in addiction as well where older adolescents seem to have more symptoms appeared. It is in line with the higher percentage of SNS as a preferred choice of medium communication and time spend on using SNS, where most of them have used the SNS more than two hours per day. In general, therefore, it seems that stronger SNS engagement were identified among the older adolescents which shows that it is important to understand the habitual and addiction phenomena on SNS usage at early stage. Before the younger adolescents become older and have the same experience of SNS engagement, it is worth to understand factors that influenced the SNS habit and addiction behavior. However, a bigger sample is needed for researcher to see the patterns of the SNS usage in order to further analyze the habit or addiction behavior. It would be also significant if some further insights can be obtained to see what factors that actually triggers the habit and addiction behavior and how does it triggers.

V. CONCLUSION

This study evaluates and examines SNS usage behavioural which is habit and addiction among adolescents in Malaysia, specifically in Klang Valley. Based on the online survey, this study offers some important remarks on how essential it needs to be pursued. Looking at the findings, it can be concluded that adolescents in Malaysia (in this study specifically in Klang Valley) experienced habit and addiction. Although all the core symptom may not appear and mostly were found in older adolescents, this study is hoped to attract attention of the scholars and shed some light on the niche area. A further study with more focus on the factors that influenced the younger adolescents to use SNS are needed because apparently they will get older and might develop the habit and addiction behavior in time. Future studies with bigger sample and advanced statistical analysis is also recommended.

ACKNOWLEDGMENT

This study is funded by research grant under Universiti Teknologi Malaysia (UTM) Research University Grant (RUG) with Vote No. 14J78. Thank you also to the Ministry of Higher Education Malaysia and UTM Research Management Team for the continuous supports.

REFERENCES

[1] Internetworldstats, "Asia Internet Usage Stats Facebook and Population Statistic," 2015.

[2] R.S.Kamini, "Malaysia faces growing Internet addiction threat," The Rakyat Post, 2014, pp. 5–7.

[3] Hootsuite, "Digital in 2017: Southeast Asia." [Online] Available: http://www.slideshare.net/wearesocialsg/digital-in-2017-southeast-asia?ref=https://aseanup.com/southeast-asia-digital-social-mobile. [Accessed: 18-Dec-2017].

[4] W. Shin, and N. Ismail, "Exploring the role of parents and peers in young adolescents' risk taking on social networking sites," Cyberpsychology, Behavior, and Social Networking, vol. 17, no. 9, pp. 578–583, 2014.

[5] M. Mäntymäki and A. K. M. N. Islam, "The Janus face of Facebook: Positive and negative sides of social networking site use," Computers in Human Behavior, vol. 61, pp. 14–26, 2016.

[6] T. K. H. Chan, C. M. K. Cheung and Z. W. Y. Lee Mr, "Transition from Urge to Excessive Use of Social Networking Sites : The Moderating Role of Self- Control and Accessibility," Pacific Asia Conference Information Systems (PACIS), 2015.

[7] M. D. Griffiths, "Social networking addiction: Emerging themes and issues," Journal of Addiction Research & Therapy, vol. 4, no. 5, pp. 4–5, 2013.

[8] P. M. Valkenburg, J. Peter and A. P. Schouten, " Friend Networking Sites and Their Relationship to Adolescents' Well-Being and Social Self-Esteem," CyberPsychology & Behavior, vol. 9, no. 5, pp. 584–590, 2006.

[9] I. Ozer, A. C. Karpinski and P. A. Kirschner, "A Cross-cultural Qualitative Examination of Social-networking Sites and Academic Performance," Procedia - Social and Behavioral Sciences, vol. 112, pp. 873–881, 2014.

[10] S. Lee, "Analyzing Negative Social Networking Sites Behaviors of Elementary and Middle School Students in Korea," Computers in Human Behavior, vol. 43, pp. 15–27, 2015.

[11] K. ISIS, "Survey on the Internet Usage (2012)," 2013.

[12] B. Subramaniam, "On the Social Media Circuit," The Star Online, [Online] Available: https://www.thestar.com.my/news/education/ 2014/ 01/05/on-the-social-media-circuit. [Accessed: 12-Jul-2017].

[13] Statista, "The Statistics Portal." [Online] Available: https://www.statista.com/statistics/199242/social-media-and networking-sites-used-by-us-teenagers. [Accessed: 12-Jul-2017].

[14] S. Pillay, "Suicide on the Rise among Malaysian Youth," The Star Online, [Online] Available: https://www.nst.com.my/news/exclusive/ 2017/05/243354 /suicide-rise-among-malaysian-youth. [Accesssed: 28-May-2018].

[15] H. Aarts and A. Dijksterhuis, "Habits as knowlegde structures: Automicity in Goal Director Behaviour," Journal of Personality and Social Psyschology2, vol. 78, no. 1, pp. 53–63, 2000.

[16] B. Verplanken and H. Aarts, "Habit, Attitude, and Planned Behaviour: Is Habit an Empty Construct or an Interesting Case of Goal-directed Automaticity?," European Review of Social Psychology, vol. 10, no. September, pp. 101–134, 1999.

[17] S. Lee and J. Jahng, "Tense moods make you use habitually facebook: A new framework for habitual IT use based on the mechanisms of mood regulation," 19th Americas Conference on Information Systems, AMCIS 2013 - Hyperconnected World: Anything, Anywhere, Anytime, vol. 4, pp. 3187–3195, 2013.

[18] M. Limayem and S. Hirt, "Force of Habit and Information Systems Usage: Theory and Initial Validation," Journal of the Association for Information Systems, vol. 4, pp. 65–95, 2003.

[19] G. L. Polites and E. Karahanna, "Shackled to the Status Quo: The Inhibiting Effects of Incumbent System Habit, Switching Costs, and Inertia on New System Acceptance," MIS Quarterly, vol. 36, no. 1, pp. 21–42, 2012.

[20] J. Cabinakova, J. Kroenung, A. Eckhardt and S. Bernius, "The importance of culture, trust, and habitual patterns: determinants of cross-cultural egovernment adoption," 21st European Conference on Information Systems (ECIS 2013), p. 13, 2013.

[21] D. L. Ronis, J. F. Yates and J. P. Kirscht, "Attitudes, decisions, and habits as determinants of repeated behavior," Attitude structure and function, pp. 213-239, 1989.

[22] D. T. Neal, W. Wood, J. S. Labrecque and P. Lally, "How do habits guide behavior? Perceived and actual triggers of habits in daily life," Journal of Experimental Social Psychology, vol. 48, no. 2, pp. 492-498, 2012.

[23] M. Limayem, S. G. Hirt and C. M. K. Cheung, "How habit limits the predictive power of intention: The case of IS continuance," MIS Quarterly, vol. 31, no. 4, pp. 705–737, 2007.

[24] W. K. Lee, "The temporal relationships among habit, intention and IS uses," Computers in Human Behavior, vol. 32, pp. 54–60, 2014.

[25] O. Turel and A. Serenko, "The benefits and dangers of enjoyment with social networking websites," European Journal of Information Systems, vol. 21, no. 5, pp. 512–528, 2012.

[26] H. Xu and B. Tan, "Why do I Keep Checking Facebook : Effects of Message Characteristic on the Formation of Social Network Services Addiction," Thirty Third International Conference on Information Systems, pp. 1–12, 2012.

[27] R. LaRose and M. S. Eastin, "A Social Cognitive Theory of Internet Uses and Gratifications : Toward a New Model of Media Attendance A Social Cognitive Theory of Internet Uses and Gratifications : Toward a New Model of Media Attendance," Journal of Broadcasting & Electronic Media, vol. 48, no. 3, pp. 358–377, 2004.

[28] M. D. Griffiths, "A "components" model of addiction within a biopsychosocial framework," Journal of Substance Use, vol. 10, no. 4, pp. 191–197, 2005.

[29] E. Guedes, A. E. Nardi, F. M. C. L. Guimarães, S. Machado, and A. L. S. King, "Social networking, a new online addiction: a review of Facebook and other addiction disorders," Medical Express, vol. 3, no. 1, pp. 1–6, 2016.

[30] P. S. Gamito, D. G. Morais, J. G. Oliveira, R. Brito, P. J. Rosa, M. Gaspar and D. Matos, "Frequency is not enough : Patterns of use associated with risk of Internet addiction in Portuguese adolescents," Computers in Human Behavior, vol. 58, pp. 471–478, 2016.

[31] D. J. Kuss, and M. D. Griffiths, "Online social networking and addiction-A review of the psychological literature," International Journal of Environmental Research and Public Health, vol. 8, no. 9, pp. 3528–3552, 2011.

[32] H. Kefi, A. Mlaiki, and M. Kalika, "Social Networking Continuance : When Habit leads to information overload," ECIS 2015 Proceedings, pp. 1–13, 2015.

[33] X. Lin, M. Featherman and S. Sarker, "Understanding Factors Affecting Users' Social Networking Site Continuance: A Gender Difference Perspective," Information & Management, vol. 54, no. 3, pp. 383-395. 2017.

[34] C. S. Andreassen, T. Torsheim, G. S. Brunborg and S. Pallesen, "Development Of A Facebook Addiction Scale," Psychological Reports, vol. 110, no. 2, pp. 501–517, 2012.

Performance Evaluation of Trivium on Raspberry Pi

Ari Kusyanti[1], Yazid Samanhudi[4]
Department of Information Technology
Universitas Brawijaya
Malang, Indonesia

Syahifudin Shahid[2]
Department of Computer System
Universitas Brawijaya
Malang, Indonesia

Harin Puspa Ayu Catherina[3]
Department of Information System
Universitas Brawijaya
Malang, Indonesia

***Abstract*—High connectivity of billions of IoT devices lead to many security issues. Trivium is designed for IoT to overcome the security challenges of IoT. The objective of this study is to implement a security service to provide confidentiality for the communication of IoT devices. Furthermore, this study aims to analyze Trivium performance in terms of keystream generation time and memory utilization on Raspberry Pi Zero, Raspberry Pi 2B, and Raspberry Pi 3B. The result showed that there was a statistically significant difference between the keystream generation time and memory utilization on Raspberry Pi Zero, Raspberry Pi 2B, and Raspberry Pi 3B based on Kruskal-Wallis H test. Further test of Jonckheere-Terpstra indicates that the fastest keystream generation time was on Raspberry Pi 3B, and the smallest memory utilization was on Raspberry Pi 2B. The implemantation of Trivium on three versions of Raspberry Pi shows promising results with less than 27 MB of memory utilization for cryptography leaves more resources available to applications.**

Keywords—Trivium; Raspberry; Kruskal-Wallis

I. Introduction

The Internet of Things (IoT) being a promising technology that enables people and objects in the physical world also numerous devices and sensors will be communicating with each other. The increased volume of communication with a huge number of nodes has the potential leading to serious security challenges. One node of IoT is an important sources of information and also a sensitive information that need to be protected, while IoT is vulnerable to attacks [1], [2], such as message modification and/or alteration, traffic analysis, Denial of Service (DoS), Distributed DoS, eavesdropping, Sybil attacks, etc. According to a study by HP states that 70% of the devices in IoT are vulnerable to attacks [3].

Various cryptographic algorithms have been researched to overcome the problem. Trivium is a stream cipher designed by Christophe De Canniere and Bart Preneel who participated in the eSTREAM competition and has been selected as part of a portfolio for hardware-oriented ciphers [4]. Based on eSTREAM, NIST recommends Trivium as one of the standard for lightweight cryptography.

Some studies evaluated Trivium and other lightweight cryptographic algorithms performance as in [5], [6] and [7]. The hardware implementation of eSTREAM ciphers have been implemented on FPGA devices [8] or 8-bit AVR microcontrollers [9] and on NodeMCU [10].

The objective of this study is to implement Trivium on three versions of Raspberry Pi as an embedded device concept of IoT to provide security services and compare the results among them. To evaluate the performance, the memory utilization and keystream generation time of Trivium algorithm is observed. To analyze the data, The Kruskal-Wallis H test is used to examine the difference implementation of Trivium on three Raspberry Pi versions.

This paper is structured as follows. Section 2 explains Trivium algorithm and the target devices, i.e Raspberry Pi Zero, Raspberry Pi 2B and Raspberry Pi 3B. Section 3 describes an explanation related to the implementation of Trivium on Raspberry Pi and the result analysis. Finally, conclusions are drawn in Section 4.

II. Trivium Architecture and Raspberry Pi

This section will disscuss the overview of Trivium and the target devices i.e. Raspberry Pi Zero, Raspberry Pi 2B and Raspberry Pi 3B, also the test to analyze the data namely Kruskall-Walis H test.

A. Trivium Architecture

In 2005 Christophe De Canni`ere and Bart Preneel developed a stream cipher called Trivium [11]. Trivium is a synchronous stream cipher that takes a secret key and an Initial Value (IV) as inputs and produces keystream that is used for encryption process. It is designed for constrained devices to generate up to bits of keystream from an 80-bit secret key and an 80-bit IV. The state consists of 288 bits which are denoted as s_0, s_2,....,s_{287} and depicted in Figure 1. There are two phases in Trivium process i.e. phase internal initialization state by using the key and IV.

The initialization process is performed by loading 80 bits of key and 80 bits of IV into the initial state of 288-bit and the rest are padded with 0, except for bit S_{286}, S_{287}, S_{288}. Then it is rotated four times, a process that is done similar to the key generation phase, but without generating a bit key stream. A complete explanation is given by the following pseudo-code:

$(s_1, s_2, \ldots, s_{93}) \leftarrow (K_1, \ldots, K_{80}, 0, \ldots, 0)$
$(s_{94}, s_{95}, \ldots, s_{177}) \leftarrow (IV_1, \ldots, IV_{80}, 0, \ldots, 0)$
$(s_{178}, s_{279}, \ldots, s_{288}) \leftarrow (0, \ldots, 0, 1, 1, 1)$

for i = 1 to 4 · 288 do
$t_1 \leftarrow s_{66} + s_{91} \cdot s_{92} + s_{93} + s_{171}$

$t_2 \leftarrow s_{162} + s_{175} \cdot s_{176} + s_{177} + s_{264}$
$t_3 \leftarrow s_{243} + s_{286} \cdot s_{287} + s_{288} + s_{69}$
$(s_1, s_2, \ldots, s_{93}) \leftarrow (t_3, s_1, \ldots, s_{92})$
$(s_{94}, s_{95}, \ldots, s_{177}) \leftarrow (t_1, s_{94}, \ldots, s_{176})$
$(s_{178}, s_{279}, \ldots, s_{288}) \leftarrow (t_2, s_{178}, \ldots, s_{287}$
end for

Keystream generation consists of an iteration process by extracting values from 15 specific bits and using them to update 3 bits andproducing keystream z_i. Full description according to pseudocode below:

for i = 1 to N do
$t_1 \leftarrow s_{66} + s_{93}$
$t_2 \leftarrow s_{162} + s_{177}$
$t_3 \leftarrow s_{243} + s_{288}$
$z_i \leftarrow t_1 + t_2 + t_3$
$t_1 \leftarrow t_1 + s_{91} \cdot s_{92} + s_{171}$
$t_2 \leftarrow t_2 + s_{175} \cdot s_{176} + s_{264}$
$t_3 \leftarrow t_3 + s_{286} \cdot s_{287} + s_{69}$
$(s_1, s_2, \ldots, s_{93}) \leftarrow (t_3, s_1, \ldots, s_{92})$
$(s_{94}, s_{95}, \ldots, s_{177}) \leftarrow (t_1, s_{94}, \ldots, s_{176})$
$(s_{178}, s_{279}, \ldots, s_{288}) \leftarrow (t_2, s_{178}, \ldots, s_{287}$
end for

B. *Raspberry Pi*

Raspberry Pi is a single-board, low-cost and high-performance developed in the UK by the Raspberry Pi Foundation that can also be used for embedded systems [10]. There are a variety of Raspberry Pi models with different features and improvements in their latest versions that add more components. In this paper, three types of Raspberry are used, i.e. Raspberry Pi Zero, Raspberry 2B, and Raspberry 3B. Table I details the comparison among Raspberry Pi.

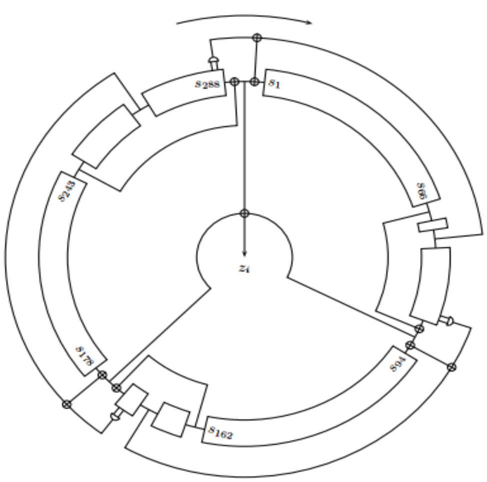


Fig. 1. Trivium Architecture.

TABLE I. RASPBERRY PI SPESIFICATION

	Rasp Pi Zero [12]	Rasp Pi 2B [13]	Rasp Pi 3B[14]
CPU	1GHz ARMI1 Core	An ARM Cortex-A7 Quad-Core 900MHz CPU	An ARM Cortex-A7 Quad-Core 900MHz CPU
RAM	512 MB	1 GB	1 GB

C. *The Kruskal-Wallis H Test*

The Kruskal-Wallis H test is a test that can be utilized to analyze if there are statistically significant differences between two or more categories of an independent variable [15]. Kruskal-Wallis is used to determine the significance (p-value) also the medians or mean ranks of three or more categories. This test is a non-parametric test in which the data within each of the multiple categories do not need to follow a normal distribution.

III. IMPLEMENTATION

Trivium algorithm is tailored for IoT to address the security challenges. The vital process in Trivium, as a symmetric key algorithm, is the generation of keystream. The keystream generation process involves complex mathematical operations. To examine the performance, Trivium is evaluated on the basis of memory utilization and keystream generation time. To measure the memory utilization and the keysteram generation time of the Trivium, each of Raspberry Pi is configured in single mode. The memory utilization and keystream generation time of Trivium algorithm are observed and compared among three types of Raspberry Pi.

The experiment run 40 times for each Raspberry Pi with various combination of Key and IV as inputs which generates keystream. The average of keystream generation time was 3391.275 microsecond for Raspberry Pi Zero, 1426.8 microsecond for Raspberry Pi 2B, and 1051.775 microsecond for Raspberry Pi 3B. In the meantime, memory utilization was 26.75 MB for Raspberry Pi Zero, 20.45 MBfor Raspberry Pi Zero and 24.00 MB for Raspberry Pi Zero. With less than 27 MB of memory utilization indicated that with fewer resources for cryptography leaves more resources available to applications.

The collected data were then analyzed by using Kruskal-Wallis H test to determine whether there is statistically significant differences of memory utilization and keystream generation time among Raspberry Pi Zero, Raspeberry Pi 2B, and Raspberry Pi 3B. Kruskall-Wallis H test was used since the data were not normally distributed. In Kruskall-Wallis H test, the p-value for the chi-square approximation test is reasonably accurate when the number of experiments is greater than or equal to 30 [16].

The result of Kruskal-Wallis H test is presented in Table II to V.

TABLE II. KRUSKAL-WALLIS H TEST RESULT OF KEY GENERATION TIME

Ranks	Raspberry Pi	N	Mean Rank
Keystream Generation Time	Raspberry Pi Zero	40	99.40
	Raspberry Pi 2 B	40	58.38
	Raspberry Pi 3 B	40	23.73
	Total	120	

TABLE III. TEST STATISTICS RESULT OF KEYSTREAM GENERATION TIME

Test Statistics	Keystream Generation Time
Chi-Square (χ2)	94.886
Asymp.Sig.	.000

TABLE IV. KRUSKAL-WALLIS H TEST RESULT OF KEY MEMORY UTILIZATION

Ranks	Raspberry Pi	N	Mean Rank
Memory Utilization	Raspberry Pi Zero	40	100.50
	Raspberry Pi 2 B	40	20.50
	Raspberry Pi 3 B	40	60.50
	Total	120	

TABLE V. TEST STATISTICS RESULT OF MEMORY UTILIZATION

Test Statistics	Memory Utilization
Chi-Square (χ2)	112.848
Asymp.Sig.	.000

Based on Table II and Table III, there was a statistically significant difference between the keystream generation time ($\chi2$= 94.886, p = 0.000), with a mean rank of 99.40 for Raspberry Pi Zero, 58.38 for Raspberry Pi 2B and 23.73 for Raspberry Pi 3B.Whilst for the memory utilization, according to Table IV and Table V, there was a statistically significant difference between memory utilization ($\chi2$= = 112.848, p = 0.000), with a mean rank of 100.50 for Raspberry Pi Zero, 20.50 for Raspberry Pi 2B and 60.50 for Raspberry Pi 3B. For further analysis, if the compared groups are ordered in a certain way, Jonckheere-Terpstra Test is conducted.

Table VI and VII presents the result of the Jonckeere-Terpstra test.

Based on Table VI, Jonckheere-Terpstra test for ordered alternatives showed that there was a statistically significant trend of keystream generation time of Raspberry Pi Zero, Raspberry Pi 2B, and Raspberry Pi 3B T_{JT} = 133.00, $p < 0.0005$. Or by way of the fastest keystream generation time was on Raspberry Pi 3B. As for the memory utilization, Jonckheere-Terpstra indicated that there was a statistically significant trend of memory utilization of Raspberry Pi Zero, Raspberry Pi 2B, and Raspberry Pi 3B T_{JT} = 1600.00, $p < 0.0005$. Or in other word, the largest memory utilization was on Raspberry Pi Zero.

TABLE VI. JONCKHEERE-TERPSTRA TEST RESULT OF KEYSTREAM GENERATION

Jonckheere-Terpstra Test	Key GenerationTime
Number of Levels in Raspberry Pi	3
N	120
Observed J-T Statistic	133.000
Asymp. Sig. (2-tailed)	.000

TABLE VII. JONCKHEERE-TERPSTRA TEST RESULT OF MEMORY UTILIZATION

Jonckheere-Terpstra Test	Memory Utilization
Number of Levels in Raspberry Pi	3
N	120
Observed J-T Statistic	1600.000
Asymp. Sig. (2-tailed)	.000

IV. CONCLUSION

Trivium is designed for IoT tohandle the security issues of IoT. The implemantation of Trivium on Raspberry Pi Zero, Raspberry Pi 2B, and Raspberry Pi 3B presents promising results that it is suitable candidate to be adopted in IoT applications as with less than 27 MB of memory utilization for cryptography leaves more resources available to applications. The results of memory utilization and keystream generation time were then compared among three types of Raspberry Pi. The result showed that there was a statistically significant difference between the keystream generation time on Raspberry Pi Zero, Raspberry Pi 2B, and Raspberry Pi 3B, and the fastest keystream generation time was on Raspberry 3B. While for memory utilization, the result presented that there was a statistically significant difference between memory utilization on Raspberry Pi Zero, Raspberry Pi 2B, and Raspberry Pi 3B, and the smallest memory utilization was on Raspberry Pi 2B.

REFERENCES

[1] M. A. Simplicio Jr, M. V. Silva, R. C. Alves, and T. K. Shibata, “Lightweight and escrow-less authenticated key agreement for the internet of things,” Computer Communications, 2016.

[2] L. Atzori, A. Iera, and G. Morabito, “The internet of things: A survey,” Computer networks, vol. 54, no. 15, pp. 2787–2805, 2010

[3] S. A. Kumar, T. Vealey, and H. Srivastava, “Security in internet of things: Challenges, solutions and future directions,” in 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016, pp. 5772–5781.

[4] C. De Canni`ere & B. Preneel. TRIVIUM A Stream Cipher Construction Inspired by Block Cipher Design Principles. In Workshop on Stream Ciphers Revisited (SASC2006), 2006.

[5] T. Good and M. Benaissa, “ASIC hardware performance,” New Stream Cipher Designs: The eSTREAM Finalists, LNCS 4986, pp. 267–293, 2008.

[6] Robshaw, M., Billet, O. (eds.): New Stream Cipher Designs. The eSTREAM Finalists. Springer (2008)

[7] Mora-Gutiérrez J.M., Jiménez-Fernández C.J., Valencia-Barrero M. (2013) Low Power Implementation of Trivium Stream Cipher. In: Ayala J.L., Shang D., Yakovlev A. (eds) Integrated Circuit and System Design.

Power and Timing Modeling, Optimization and Simulation. PATMOS 2012. Lecture Notes in Computer Science, vol 7606. Springer, Berlin, Heidelberg

[8] Levent Ertaul, Arnold Woodall. "IoT Security: Performance Evaluation of Grain, MICKEY, and Trivium - Lightweight Stream Ciphers" on Int'l Conf. Security and Manaement, SAM 2017.

[9] A. Jafarpour, A. Mahdlo, A. Akbari and K. Kianfar, "Grain and Trivium Ciphers Implementation Algorithm in FPGA Chip andAVR Micro Controller," in ICCAIE, 2011.

[10] G. Meiser, T. Eisenbarth, K. Lemke-Rust and C. Paar, "Efficient Implementation of eSTREAM Ciphers on 8-bit AVR Microcontrollers," Industrial Embedded Systems, 2008.

[11] C. De Canni`ere & B. Preneel. TRIVIUM Specifications. eSTREAM, ECRYPT Stream Cipher Project, Report 2005/001, 2005. http://www.ecrypt.eu.org/stream.

[12] Raspberry. Raspberry Pi 1 B+. raspberry.org.

[13] Raspberry. Raspberry Pi 2 B. raspberry.org.

[14] Raspberry. Raspberry Pi 3 B. raspberry.org

[15] Kathleen F. Weaver Vanessa Morales Sarah L. Dunn Kanya Godde Pablo F. Weaver. An Introduction to Statistical Analysis in Research: With Applications in the Biological and Life Sciences. Wiley Book, 2017.

[16] Green, S. B., & Salkind, N. J. Using SPSS for Window and Macintosh: Analyzing and Understanding data (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall. 2008.

Improving K-Means Algorithm by Grid-Density Clustering for Distributed WSN Data Stream

Yassmeen Alghamdi[1], Manal Abdullah[2]
Faculty of Computing and Information Technology, Department of Computer Science
King Abdul-Aziz University, KAU
Jeddah, Saudi Arabia

***Abstract*—At recent years, Wireless Sensor Networks (WSNs) had a widespread range of applications in many fields related to military surveillance, monitoring health, observing habitat and so on. WSNs contain individual nodes that interact with the environment by sensing and processing physical parameters. Sometimes, sensor nodes generate a big amount of sequential tuple-oriented and small data that is called Data Streams. Data streams usually are huge data that arrive online, flowing rapidly in a very high speed, unlimited and can't be controlled orderly during arrival. Due to WSN limitations, some challenges are faced and need to be solved. Extending network lifetime and reducing energy consumption are main challenges that could be solved by Data Mining techniques. Clustering is a common data mining technique that effectively organizes WSNs structure. It has proven its efficiency on network performance by extending network lifetime and saving energy of sensor nodes. This paper develops a grid-density clustering algorithm that enhances clustering in WSNs by combining grid and density techniques. The algorithm helps to face limitations found in WSNs that carry data streams. Grid-density algorithm is proposed based on the well-Known K-Means clustering algorithm to enhance it. By using Matlab, the grid-density clustering algorithm is compared with K-Means algorithm. The simulation results prove that the grid-density algorithm outperforms K-Means by 15% in network lifetime and by 13% in energy consumption.**

***Keywords*—WSNs; data mining; clustering; data stream; grid density**

I. Introduction

In recent years, a widespread use of WSNs are found in various applications. A WSN is a specific kind of ad-hoc networks that is able to sense and process information. They can be used in many areas such as environmental, industrial, military, and agriculture fields. WSNs consist of built-in, independent and tiny equipment's called sensor nodes. Sensor nodes contain four main components: energy source, processing unit, sensing [1] unit and transducer. Sensor nodes are mainly used in processing data and report parameters continuously. The reports are transferred by sensor nodes and collected by special controllers called Base Stations (BSs). A WSN has many resource constraints including high computational power and limited energy source. WSNs depend on their nodes that consumes battery energy. Unfortunately, the WSNs nature makes it difficult to recharge sensor nodes batteries. Therefore, energy efficiency is an important design objective in WSNs [2] and their algorithms should be precisely designed based on energy saving.

In some WSN applications, data that WSNs process usually contain a large amount of datasets that flow rapidly in a very high speed and arrive online. Data are considered to be unlimited and the arriving order of elements being processed is out of control. Such data are called Data Streams [1, 3].

The widespread deployment of WSNs and the need for aggregating data streams requires an efficient organization of network topology to reach load balancing and extension of network lifetime. This is performed by using mining techniques. Clustering is a data mining technique that is considered to be an efficient tool in WSNs to solve the problem of network lifetime, energy consumption, data aggregation, load balancing, scalability [4], delay and delivering data stream packets. It organizes WSNs into a connected hierarchy. In general, two categories of network structure are found in WSNs, flat and clustered (i.e. hierarchical) [5]. At any case, clustering plays an important part in network organization, and also affects network performance. To reach many advantages, clustering is preferred in mining WSNs data. In a clustered WSN, the network is divided into groups called clusters, each cluster has a leader elected from sensor nodes called Cluster Head (CH). Data streams are aggregated from nodes by their CH inside a cluster. Then it is transmitted from CHs to the BS. Transmitting streaming data in the wireless environment by a multi-hop communication to reach the BS resumes nodes energy leading to shorten the lifetime of a network.

As mentioned previously, a WSN suffers from power consumption during data stream transmission. Sensor nodes should be energy efficient. Energy efficiency affects the entire WSN lifetime. Therefore, to gain WSN's operational long-lasting, consuming energy is considered during designing WSNs algorithms. Furthermore, since sensor nodes are in difficult-to-reach locations, replacing batteries is unpractical. A WSN can achieve energy saving from clustering algorithms. However, to achieve better energy conservation, data stream mining [6] must be formed in a distributed manner, due to their resource constraints.

Clustering algorithms are designed to obtain load-distribution between CHs, high connectivity, saving energy and fault tolerance. In WSNs, using resources and reducing energy is provided in clustering technique by decreasing number of nodes that transmit data streams through long distance transmission. Clustered WSN algorithms running streaming data are usually partitioned in two main steps, cluster formation step and data transmission step [7]. But specifically, the cluster-based operation of clustered WSN algorithms consist of rounds. Rounds involve cluster creation, CH-Election, and data transmission.

Grid-based clustered WSNs, are type of networks [4] where a sensed area is partitioned into a number of equally sized small cells called grids. Grid-based clustering scheme has proven to have a fast processing time compared to other types of clustering algorithm schemes due to computational operations are performed on grid cells instead of the whole dataset stream.

This paper develops a distributed clustering algorithm for WSNs based on the well-known K-Means to enhance it. The algorithm is based on combining a grid technique and a density technique. Besides clustering advantages, density technique can find arbitrary shaped clusters with noise, while grid technique is used to avoid clustering quality problems by discarding the boundary nodes of grids. This combination of techniques decreases algorithm computational time, reduce energy consumption and thus increasing network lifetime resulting desirable simulation results. To reach this paper aims, the algorithm must converge the limited dataset streams as fast as possible, to ensure that a processor can take on next set of streams. The paper provides an evaluation of our grid-density clustering algorithm to prove its efficiency by comparing its final results with K-Means results. The remaining of this paper is ordered as follows. Literature review on clustering algorithms is provided in section II. An overview on K-Means clustering algorithm is given in section III. The proposed grid-density clustering algorithm is explained in section IV. Simulation analysis and results are discussed in section V. Concluding is given in section VI.

II. Literature Review

Some clustering algorithms are found in WSNs process traditional sensed data. Based on network structure, algorithms found in WSNs can be divided into two classes: algorithms for either flat networks or hierarchical networks. In a flat network structure, all nodes have the same tasks and perform exact functionalities. Data transmission is done in a hop-by-hop manner using flooding. Some clustering algorithms in flat WSNs include Energy-Aware Routing (EAR), Gradient-Based Routing (GBR), Sequential Assignment Routing (SAR) and many more. These clustering protocols are efficient in networks with a small-scale. However, flat WSN algorithms are unfavorable in networks with large-scale due to resource limitation, but nodes such networks preform more data processing [5]. In a hierarchical structure, nodes have different functions and are organized into groups according to specific requirements or metrics. Generally, each cluster has a specific CH and other sensor nodes. CHs have the highest energy inside clusters to perform processing and transferring data, while other nodes with low energy perform sensing [5]. Some clustering algorithms in a hierarchical WSN topology include Low-energy Adaptive Clustering Hierarchy (LEACH), Energy-Efficient Uneven Clustering (EEUC) algorithm, Algorithm for Cluster Establishment (ACE). Clustering technique is an important scheme in hierarchical WSNs due to many advantages, such as data aggregation [5], scalability, less load, low energy consumption and more robustness.

Other clustering algorithms stream data stream in other environments rather than WSNs. In 2006, Feng Cao proposed the DenStream algorithm for clustering dynamic data stream [8]. It is an effective method that can discover clusters with arbitrary format in data streams, but it is insensitive to noise [9]. Heng Zhu Wei proposed a density and space clustering algorithm called CluStream [8]. CluStream is a clustering data stream algorithm based on K-Means that is inefficient to get clusters of arbitrary formats and cannot process outliers. Further, they have to predetermine a parameter K (i.e. number of clusters) [10].

K-Means is used in an offline phase of some algorithms such as Clustream. It is a divide and conquer schemes that partition data streams into segments and discover clusters in data streams. K-Means has a number of limitations. Firstly, it doesn’t reveal clusters with arbitrary formats and usually identifying spherical clusters. Secondly, it is unable to discover outliers and noise. Thirdly, K-Means requires multiple scans of data, making it impractical for huge data stream [8, 10]. STREAM and CluStream are data stream clustering algorithms that are extensions of K-Means [11].

LOCALSEARCH, STREAM, DenStream and CluStream are clustering algorithms involving data streams. They disregard grid border problems. Data streams come with a large number in chronological order, and makes original grids no longer adapt to new data mapping, so a large number of data is likely to fall on grids borders. But if the data is simply discarded, it affects the clustering quality. If grids are updated in time, cost is greatly increased and the clustering efficiency is affected greatly [8].

D-Stream is a real-time density-grid stream data clustering algorithm where nodes are assigned to grids and grids are gathered to form clusters based on their density. D-Stream clustering quality depends on the lowest grid structure level. This may reduce the clusters accuracy despite the technique processing time speed [11]. D-Stream assigns input data into grids by using an online component. It also computes density of each grid and performs clusters based on their density by using an offline component [10]. MR-Stream is an algorithm that can cluster data streams at various resolutions. It divides a given data space to cells and a data structure tree that keeps the space dividing. MR-Stream increases the clustering performance by determining the exact time to generate clusters [11]. FlockStream is a density clustering algorithm that is based on the concept of bio-inspired model. It uses the flocking model, where independent micro-cluster agents form clusters together. FlockStream combines online and offline components where agents form clusters once required. It can get clustering results without performing offline clustering. DenStream, MR-Stream, D-Stream and FlockStream are density-based clustering algorithms carrying data streams. They can affectively reveal clusters with arbitrary shapes and handle noise, but their quality decrease when using clusters with variant densities. LOCALSEARCH algorithm [8] uses dividing and conquering to partition data streams into segments, and discovers clustering of data streams in finite space, by using the K-Means algorithm.

A framework to dynamically cluster multiple evolving data streams called Clustering on Demand (COD) was proposed [12]. It produces a summary hierarchy of data statistics in the online phase, whereas clustering is performed in the offline phase [12]. It summarizes data streams using the Discrete Fourier Transform (DFT) technique. Then it applies a K-Means algorithm to cluster the summarized data streams [12]. An Online Divisive-Agglomerative Clustering (ODAC) algorithm was also proposed to incrementally construct a tree-like hierarchy of clusters using a top-down strategy. The previous techniques assume that all data streams are gathered at a centralized site before they are processed [12]. Many density-based clustering algorithms for multi density datasets are not suitable for data stream environments. First, they need two-pass of data and this condition is impossible for data streams where they arrive continuously and need a single scan to be performed. GMDBSCAN and ISDBSCAN use a two-pass data. Second, some multi-density clustering algorithms require using the whole data. Third, other algorithms have a high execution time which makes them unsuitable when applying data streams. DSCLU [11] is considered to be a density clustering algorithm run streaming data in multi-density environments.

Another class of clustering algorithm is when applying data streams on WSNs. It is divided into two subsections [13]: algorithms based on Fuzzy clustering in WSNs and algorithms based on multimedia streaming data streams found in Multimedia Wireless Sensor Networks (MWSNs). Our proposed density-grid clustering algorithm is similar to research study scope of algorithms belongs under this class. Fuzzy C-Means or Fuzzy Clustering Means (FCM), is a widely used data stream mining clustering algorithm in WSNs. Most clustering algorithms are descendant from FCM when solving data stream problems in WSNs. FCM requires prior information about how many needed clusters C to divide the data space. The clusters number C is unknown previously.

An algorithm based on FCM, a distributed WSN data stream clustering algorithm called SUBFCM (Subtractive Fuzzy Cluster Means) is proposed to decrease nodes energy consumption and extend network lifetime in WSNs involving data streams. The SUBFCM focuses on the clustering problem on data streams. Simulations show that the energy efficient algorithm SUBFCM can obtain clustering with less energy than the FCM. SUBFCM reduces the overall data transmission needed without affecting vital information in data streams [13]. SUBFCM is a result of blending a subtractive clustering algorithm with the FCM.

For algorithms based on multimedia streaming data, in WMSNs, multimedia clustering protocols use the quality of service (QoS) parameters [14]. The requirements of QoS differ based on to the multimedia applications type. QoS has several metrics such as jitter, delay, bandwidth, reliability [15] and packet loss [14]. A lot of multimedia applications are time critical, they require to be managed with a limited time. Sensors in multimedia are able to grab image, audio, video, and so on. Then send \ multimedia content by sensors [15]. FoVs is a wireless multimedia sensor network clustering algorithm proposed based on Overlapped Field of View (FoV) areas. FoVs prolongs network lifetime and saves energy [16].

III. K-Means Algorithm

The simplest algorithm that solve a well-known clustering problem is called the K-Means clustering algorithm. K-Means has an efficient CH selection method to maximize energy efficiency of a WSN. K-Means is based on finding a CH that minimizes the sum of Euclidean distances between CH and nodes [17, 18]. It reduces communication overhead, energy consumption and extends network lifetime. It is used to partition a sensed area into K clusters. The procedure follows a simple way to classify a given dataset through a certain number of clusters fixed a priori [18]. In K-means, there is a distance threshold called R for calculating distance between CH and BS. If their distance is less than R, they use a single-hop transmission, otherwise, they use a multiple-hops transmission [7]. There is also an energy threshold called E for all CHs. If CH energy is less than E, then CH broadcasts a quit message to all nodes inside the cluster. Hence, other nodes which have higher residual energy are elected to become CHs [7]. Nodes near boundary region in K-Means are affected since, the degree of belongingness is described in terms of either zero or one. For this reason, K-Means clustering is called hard clustering. Edge nodes may have the same degree of belongingness to more than one clusters. In K-Means, there is an optimal cluster formation. Nodes are assigned to a cluster based on the degree of belongingness when network area deployment. Degree of belongingness needs to be computed in each round for every node inside a cluster. Obviously, the major limitation of K-Means clustering algorithm is predetermining parameter K [6]. The traditional K-Mean algorithm in figure 1.

K − Means Algorithm

Input: K, n

1- *Set* (k) *as number of clusters.*

2- *Set* (n) *as total number of nodes.*

3- *Initialize value* (k) *to estimate each cluster centroid.*

4- *Assign each* (n) *to its cluster whose centroid has nearest distance.*

5- *Centroid recalculation* (k) *after assigning each node to its cluster.*

6- *Repeat* (3) *and* (4) *until*

   6.1 *no node changes its cluster assignment.*

   6.2 *or until* (k) *no longer moves.*

Fig. 1. K-Means Algorithm Pseudo Code.

IV. Grid-Density based Clustering Algorithm

Clustering WSN algorithms could be considered under specific schemes as shown in figure 2, such as, hierarchical scheme, grid scheme, heuristic scheme, weighted scheme, PSO-Based scheme. The developed grid-density algorithm is a grid scheme. Figure 2 summarizes clustering schemes in WSNs with an example of each clustering scheme. The grid-density clustering algorithm is a clustering algorithm that forms clusters based on density of each grid in a gridded WSN. Grid-density algorithm is proposed based on K-Means to enhance it.

It solves the same problems that K-Means clustering algorithms solve. But grid-density clustering algorithm forms clusters in a different manner, where cluster formation is not based on the Euclidian distance calculation. Cluster formation is based on finding the density of each grid. Additionally, it doesn't require a predetermination of clusters number as required in K-Means.

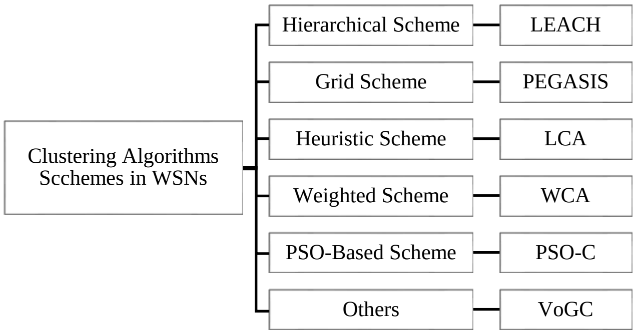


Fig. 2. Clustering Algorithms Schemes in WSNs.

To form network clusters, the developed scheme is done by dividing a sensor network area into equal size of grids. The area is divided by a value called grid size g where $\boldsymbol{g} \in \boldsymbol{X}$, and $\boldsymbol{g} \in \boldsymbol{Y}$. Grids then are classified based on their densities by using a specific value called threshold $\boldsymbol{\sigma}$. By using the suggested algorithm and both values $\boldsymbol{g}$ and $\boldsymbol{\sigma}$, grids close to each other are combined after finding their density to form arbitrary shaped clusters. Empty grids are used as delimiters to reduce the algorithm execution time. Cluster formation process in the grid-density algorithm is based on $\boldsymbol{g}$ and $\boldsymbol{\sigma}$ to find number of clusters $\boldsymbol{C}$. After forming clusters, the grid-density selects a CH for each cluster based on nearest distance to the BS. Figure 3 (a) shows a sensed area that is already divided into grids assuming that $\boldsymbol{g} = \mathbf{155}$ by using the grid-density algorithm. In contrast, figure 3 (b) presents the same sensed area in K-Means that is not gridded.

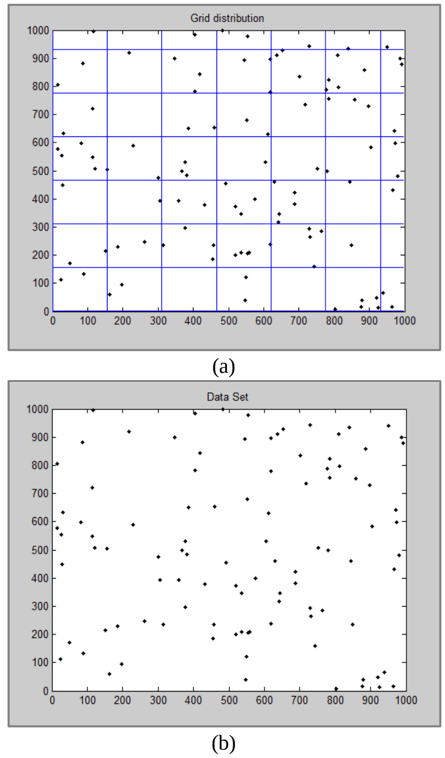


Fig. 3. (a) Gridded WSN in Grid-Density Clustering Algorithm, (b) Sensed Area in K-Means.

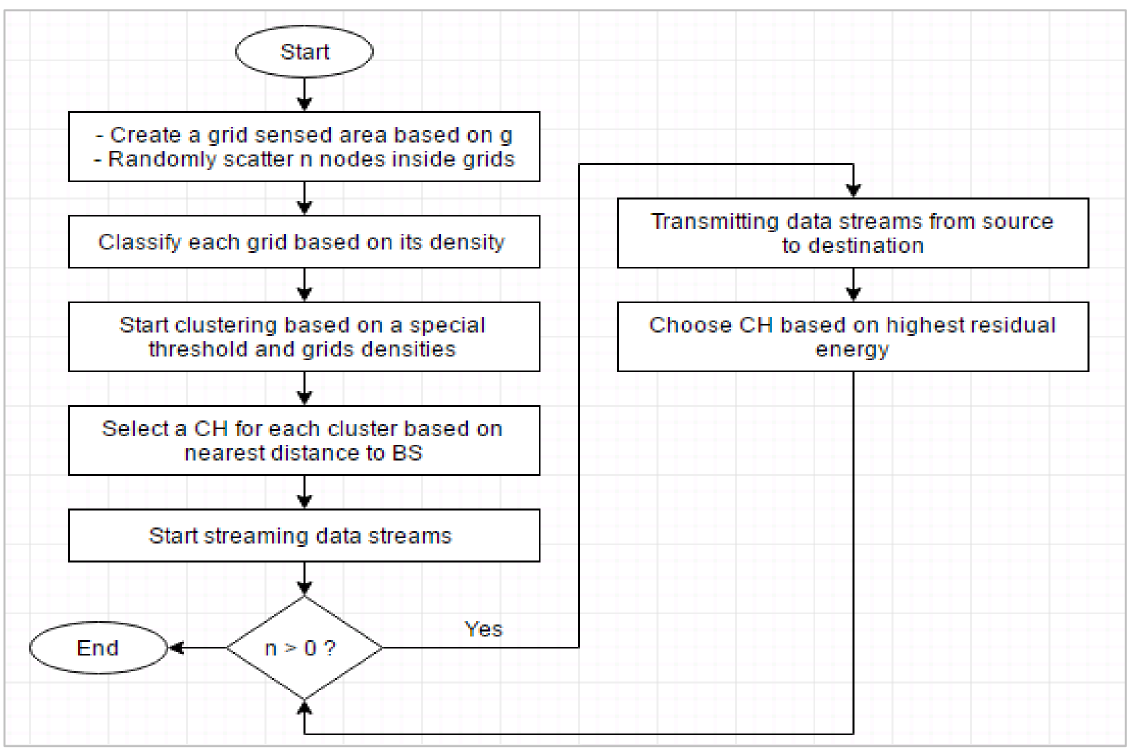
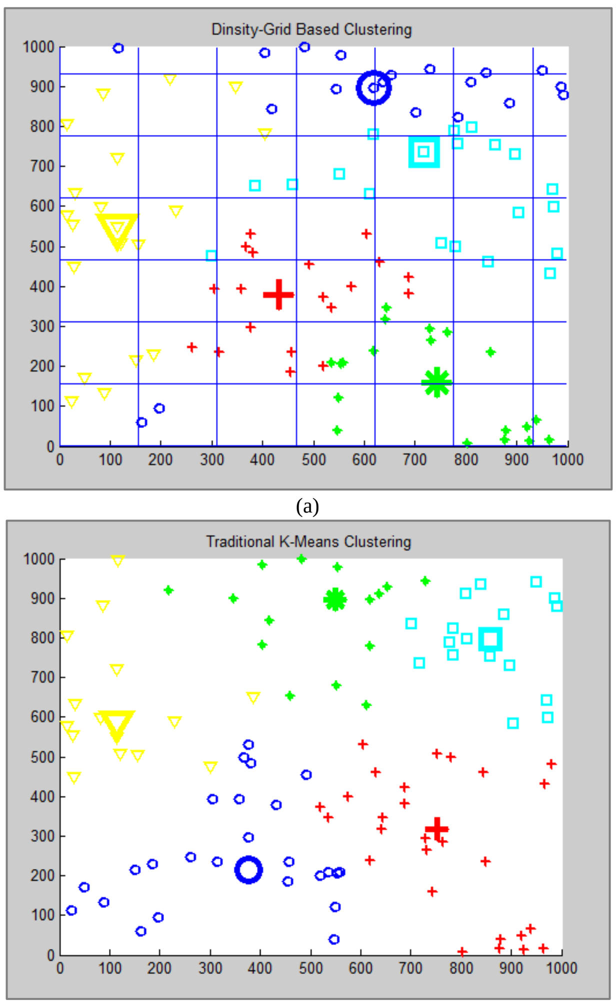


Fig. 4. Grid-Density Clustering Algorithm Flowchar.

After forming network clusters and choosing CHs initially, the network is ready to stream data. To stream data, the algorithm goes through several rounds until the end of network lifetime. Each round consists of two steps, transmitting data and choosing CHs. First step is responsible for transmitting sensed streaming data from source nodes to the final destination at the BS through CHs. The second step is to rotate the role between CHs based on nodes highest residual energy. The procedure used to obtain the final experimental results for both competitors is by running the grid-density algorithm first to form its clusters, then gain number of clusters $\boldsymbol{C}$. After that, the traditional K-Means is run individually using the predetermined value $\boldsymbol{C}$ gained from the grid-density algorithm. Comparisons between competitors is done based in network lifetime and energy consumption for the entire network. Figure 4 represents grid-density algorithm flowchart.

V. Simulation Analysis and Results

To implement and evaluate the grid-density algorithm, Matlab version R2008b was used. In our research experiments, Matlab is used in a machine with Windows 7 Service Pack 1 with 1TB disk space, 64-bit operating system, Intel ® Core ™ i7 processor and 8GB RAM.



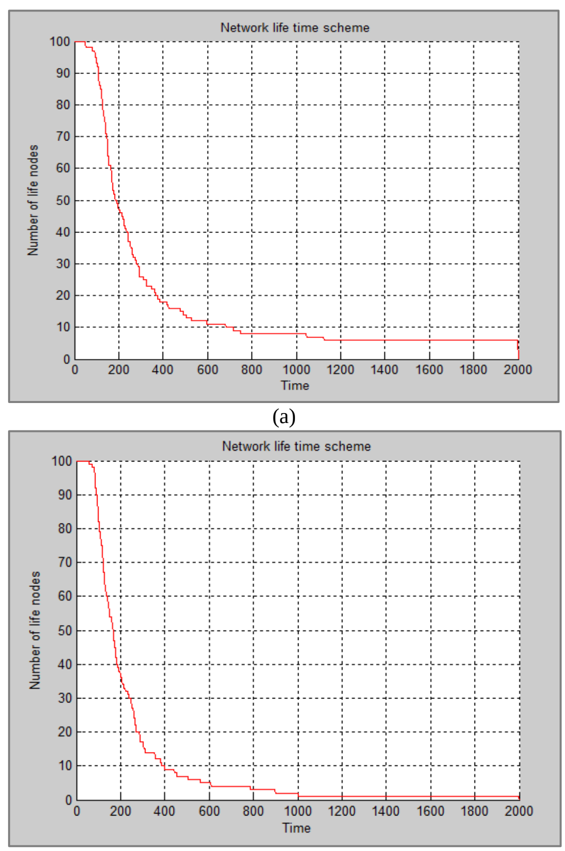
(a)

(b)

Fig. 5. (a) Cluster Formation in Grid-Density Clustering Algorithm, (b) cLUSTER FORMATIOn in K-Means.

After several simulation experiment, the following experiment has the best results and is chosen to compare between the two competitors' final performance metrics results in terms of network lifetime and energy consumption. Assuming that grid size $\boldsymbol{g} = 155$ and threshold $\boldsymbol{\sigma} = 5$. Both algorithms in their experiment are streaming the same dataset stream packet with size 126 byte/message in a (1000×1000) m^2 sensed area and BS located at the center, with $\boldsymbol{n} = 100$ node scattered randomly each with an initial energy equal to $1Joule$. At cluster formation process in the grid-density algorithm, clusters number obtain from this experiment is $\boldsymbol{C} = 5$. Figure 5 (a) represents five individual clusters formed in the grid-density algorithm, each with a clear CH. By using number of clusters $\boldsymbol{C} = 5$, gained from the grid-density algorithm, then applying $\boldsymbol{C}$ as an input for K-Means on the given sensed area, figure 5 (b) shows the cluster formation result in K-Means each with a clear CH. It is clear from figure 5 that the sensed area is gridded in the grid-density algorithm while not gridded in K-Means.

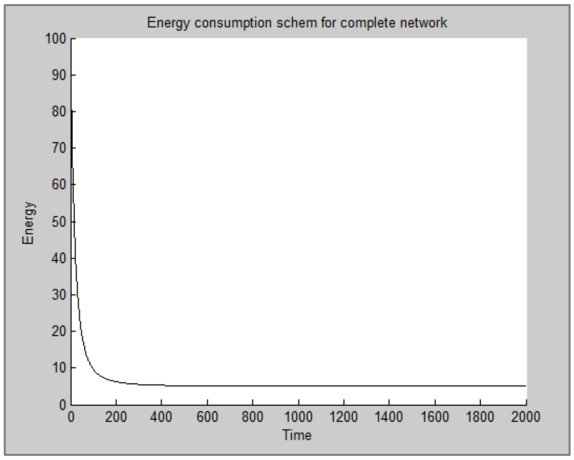
In figure 6 (a) and (b), both graphs present network lifetime for grid-density algorithm and K-Means consequently. X-axis presents the running time in seconds (sec.) and Y-axis presents number of live nodes. It is clearly shown that the proposed algorithm extends network lifetime more than K-Means, where K-Means nodes stats to die at 1000 sec before the proposed algorithm nodes. From experimental results, the grid-density algorithm extends network lifetime by about 15% more than K-Means.



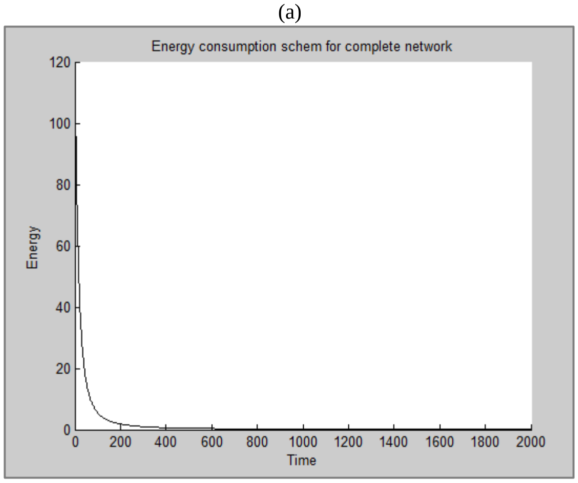
(a)

(b)

Fig. 6. (a) Network Lifetime in Grid-Density Clustering Algorithm, (b) Network Lifetime in K-Means.



(a)



(b)

Fig. 7. (a) Energy Consumption in Grid-Density Clustering Algorithm, (b) Energy Consumption in K-Means.

Figure 7 (a) and (b) present graphs of grid-density algorithm and K-Means consequently for their energy consumption. X-axis presents the energy in percentage in Joule while Y-axis presents the running time in Second (sec.). It is found that the proposed algorithm reduces energy consumption by about 13% less than K-Means.

The grid-density algorithm processes small grids, were all operations are performed on grid cells rather than processing the whole sensed area space and exhaustion the network as found in K-Means. Gridding reduces network exhaustion, thus reduces energy consumption that in turn extends network lifetime.

VI. Conclusion

Recent years witnessed a wide use of WSNs in several applications and it has become an interesting research area of study in data mining field. This paper provided an overview on clustering algorithms. It proposed a WSN clustering algorithm involving data streams based on the well-known K-Means. The proposed grid-density clustering algorithm is based on the concept of finding the density of each grid to form clusters in a WSN. The grid-density clustering algorithm results are regarding some performance metrics that are compared with K-Means algorithm results. Simulation results prove that the grid-density algorithm outperforms K-Means by 15% in network lifetime and by 13% in energy consumption performance metrics.

References

[1] A. L. de Aquino, C. M. S. Figueiredo, E. F. Nakamura, L. S. Buriol, A. Loureiro, A. O. Fernandes, et al., "A sampling data stream algorithm for wireless sensor networks," in Communications, 2007. ICC'07. IEEE International Conference on, 2007, pp. 3207-3212.

[2] O. Younis, M. Krunz, and S. Ramasubramanian, "Node clustering in wireless sensor networks: recent developments and deployment challenges," Network, IEEE, vol. 20, pp. 20-25, 2006.

[3] A. L. de Aquino, C. M. Figueiredo, and E. F. Nakamura, "Data Stream Algorithms For Processing of Wireless Sensor Network Application Data."

[4] M. Abdullah, H. N. Eldin, T. Al-Moshadak, R. Alshaik, and I. Al-Anesi, "Density Grid-Based Clustering for Wireless Sensors Networks," Procedia Computer Science, vol. 65, pp. 35-47, 2015.

[5] X. Liu, "A survey on clustering routing protocols in wireless sensor networks," sensors, vol. 12, pp. 11113-11153, 2012.

[6] H. Sabit and A. Al-Anbuky, "Multivariate spatial condition mapping using subtractive fuzzy cluster means," Sensors, vol. 14, pp. 18960-18981, 2014.

[7] S. Zhong, G. Wang, X. Leng, X. Wang, L. Xue, and Y. Gu, "A Low Energy Consumption Clustering Routing Protocol Based on K-Means," Journal of Software Engineering and Applications, vol. 5, p. 1013, 2012.

[8] C. Jia, C. Tan, and A. Yong, "A grid and density-based clustering algorithm for processing data stream," in Genetic and Evolutionary Computing, 2008. WGEC'08. Second International Conference on, 2008, pp. 517-521.

[9] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-Based Clustering over an Evolving Data Stream with Noise," in SDM, 2006, pp. 328-339.

[10] Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 133-142.

[11] A. Amini, H. Saboohi, and T. Y. Wah, "A multi density-based clustering algorithm for data stream with noise," in Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on, 2013, pp. 1105-1112.

[12] J. Yin and M. M. Gaber, "Clustering distributed time series in sensor networks," in Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, 2008, pp. 678-687.

[13] H. Sabit, A. Al-Anbuky, and H. Gholam-Hosseini, "Distributed WSN data stream mining based on fuzzy clustering," in Ubiquitous, Autonomic and Trusted Computing, 2009. UIC-ATC'09. Symposia and Workshops on, 2009, pp. 395-400.

[14] J. R. Diaz, J. Lloret, J. M. Jimenez, and J. J. Rodrigues, "A QoS-based wireless multimedia sensor cluster protocol," International Journal of Distributed Sensor Networks, vol. 2014, 2014.

[15] M. Abazeed, N. Faisal, S. Zubair, and A. Ali, "Routing protocols for wireless multimedia sensor network: a survey," Journal of Sensors, vol. 2013, 2013.

[16] V. Kumar, S. Jain, and S. Tiwari, "Energy efficient clustering algorithms in wireless sensor networks: A survey," 2011.

[17] D. A. S. Juned M. Khan, "PERFORMANCE COMPARISON OF FCM AND K-MEAN CLUSTERING TECHNIQUE FOR WIRELESS SENSOR NETWORK IN TERMS OF COMMUNICATION OVERHEAD," Global Journal of Advanced Engineering Technologies and Science, pp. 26 - 29, May, 2014.

[18] A. Thangavelu and A. Pathak, "Clustering Techniques to Analyze Communication Overhead in Wireless Sensor Network," Editorial.

Conditional Text Paraphrasing: A Survey and Taxonomy

Ahmed H. Al-Ghidani, Aly A. Fahmy
Computer Science Department
Faculty of Computers and Information,
Cairo University
Cairo, Egypt

***Abstract*—This work introduces a survey for the *Text Paraphrasing* task. The survey covers the different types of tasks around text paraphrasing and mentions the techniques and models that are regularly used when approaching towards it, alongside the datasets that are used while training and evaluating the models. Text paraphrasing has an effective impact when it is used in other applications, so, the paper mentions some text paraphrasing applications. Also, this work proposes a new taxonomy that it is called *Conditional Text Paraphrasing*. To the best of our knowledge, this is the first work that shows varieties and sub-problems of the original text paraphrasing task. The target of this taxonomy is to expand the definition of the text paraphrasing by adding some conditional constraints as features that either control the paraphrase generation or discrimination. This expanded definition opens in mind a new domain for research in Natural Language Processing (NLP) and Machine Learning. Finally, some useful applications for the conditional text paraphrasing are represented.**

***Keywords*—Natural Language Processing; Text Paraphrasing; Conditional Text Paraphrasing**

I. Introduction

A. Problem Definition

Text Paraphrasing is a core and challenging problem in Natural Language Processing. The problem refers to texts that convey the same meaning but with different expressions. It can be considered as a transformation for a given text while keeping its semantic meaning. These transformations may be at the level of the texts linguistics and structure. Paraphrasing differs from Entailment in the type of relationship between instances. *Entailment* occurs when one may draw necessary conclusions from an input instance. For example, *Rocky is a Dog*. *Rocky is an animal*. Entailment has a form of ***"If A then B"***, while Paraphrasing has a form of ***"A is B"***. Entailment is a one-way relationship. ***"If A then B"*** is not ***"If B then A"***. For instance: *Rocky is a Dog* means *Rocky is an animal*, while *Rocky is an animal* does not necessarily mean that *Rocky is a Dog* as it is maybe another type of animals. Unlike Entailment, Paraphrasing is a two-way relationship. For instance, *What is the distance between Earth and Sun?* is the same of *How many miles between Earth and Sun?*

This work demonstrates the defined tasks around the text paraphrasing problem (section 1.2) and formulates discrimination and generation tasks, alongside mentioning some existed researches and efforts that are done on both directions (sections 2.1 and 2.2). After that, some evaluation metrics are represented that are regularly used when evaluating the model (section 3) and the datasets used (section 4). We talk about several applications on which text paraphrasing is used, either for data augmentation or as a module in large systems (section 5). Finally, a proposed definition for conditional text paraphrasing is introduced and its taxonomy (section 6), alongside showing some important applications on it (section 7).

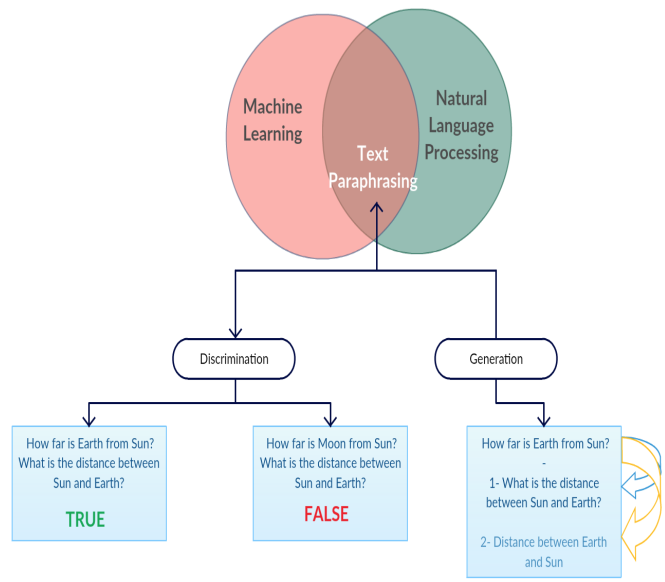


Fig. 1. Text Paraphrasing Tasks

B. Defined Tasks

As shown in Figure 1, text paraphrasing is a type of problems at which natural language processing and machine learning could co-operate to solve it. Text paraphrasing involves two different tasks, **Discrimination** and **Generation**. The target of the discrimination is to check if the two given texts are paraphrased texts or not. In that case, the task is considered to be a discriminative problem. In the generation, the target is to generate text(s) given a reference text, in that case, the task is a generative problem.

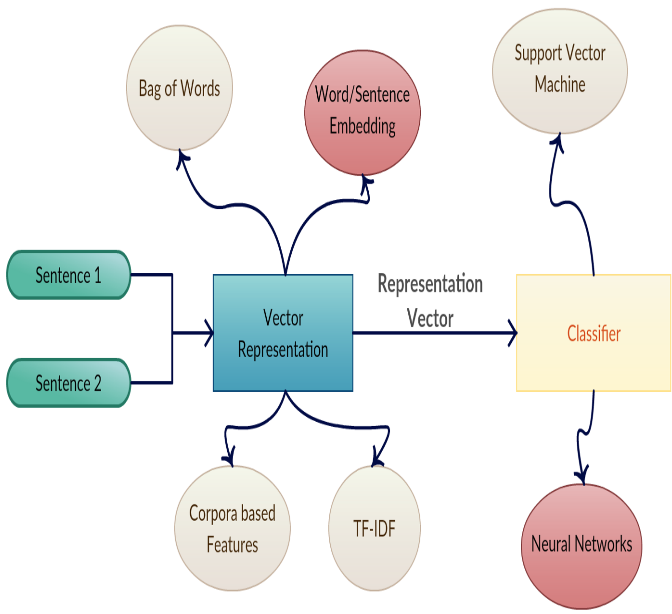


Fig. 2. Discrimination Model Pipeline

II. Problem Formulation

A. Discrimination Task

Given two sentences (S_1, S_2), where $S_1 = \{w_1, w_2, .., w_n\}$ and $S_2 = \{w_1, w_2, .., w_m\}$, the target is to check if they are paraphrases or not.

Some researches look to this problem as a Semantic Text Similarity problem on which some distances metrics, such as Euclidean and Cosine distances are used [1], [2], alongside either using binary vectors as feature representation for sentences extracted from lexical-based features or TF-IDF representation [3]. However, the great successes of distributed words and sentences representation [4], [5], [6] altered the basic representations for texts to be used in distances measure [7].

Other researches look to the problem as a supervised learning problem. The problem is often formalized as a binary classification problem $y = \{0, 1\}$. Like in [8] on which Support Vector Machine (SVM) is used with basic features representations for sentences. The great successes of deep neural networks in fields like natural language processing, computer vision and speech recognition in both supervised and unsupervised learning problems was a motivation to build a neural-based model for such classification problem [9]. Recently, Convolutional Neural Networks (CNNs) showed remarkable results in text modeling for classification and features extraction [10], [11], so that it could be used for this task like in [12]. Other works focus on recurrent based models like in [13] on which Long Short-Term Memory (LSTM) is used to encode the sentences embedded into a Siamese network structure [14]. Shown in Figure 2, this is the general pipeline for the discrimination task.

In conclusion, currently, deep neural networks are heavily used for the identification task, alongside the currently advanced words and sentences representations.

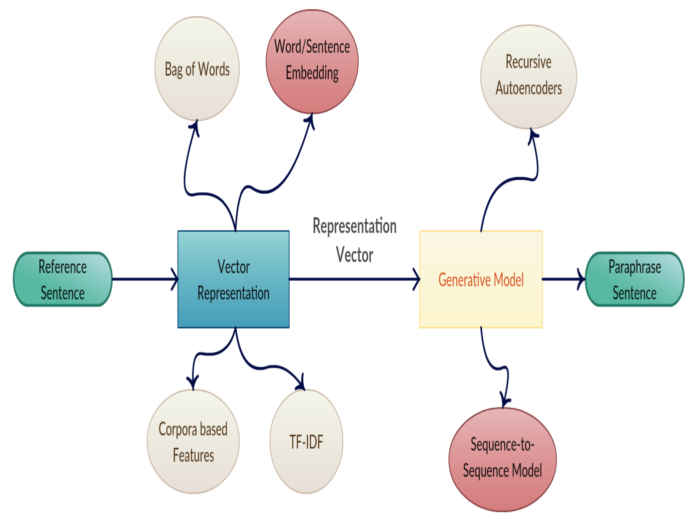


Fig. 3. Generation Model Pipeline

B. Generation Task

In the generation task, given a reference sentence S_1, where $S_1 = \{w_1, w_2, .., w_n\}$, the target is to generate candidate(s) sentence(s) that are semantically equivalent to the reference sentence. It is considered to a text generation problem.

Classically, some lexical-based features and wording replacement are used to generate alternatives to the reference sentence. For instances, paraphrases are generated using templates extracted from WikiAnswers repositories like in [15] and lexical-based rules like in [16]. They make use of WordNet to get words hypernyms and synonyms for replacements, however, these techniques suffer from the generation of poor candidate paraphrases.

Recently, the great successes of the deep generative models [17] such as Variational Autoencoders (VAEs) [18] and Generative Adversarial Networks (GANs) [19] had a great impact in unsupervised learning problems. Several research worked on generation realistic texts either for task specific problems, such as machine translation [20], [21], and question generation [22]. Generic text generation has been investigated using VAEs [23] and GANs [24]. As these models depend on the hidden representation of sentences during the training, the produced texts are randomized and uncontrollable. Serious attempts were made recently to control the generated sentences [25], [26], using some conditional features such as the sentence's polarity and syntax-tree [25], [26]. Typically, this problem is considered to be sequence-to-sequence problem [27] on which the target is to generate sequence(s) of words given other sequence of words [28], [29], [30]. Shown in Figure 3, this is the general pipeline for the generation task with highlights of the most used techniques nowadays.

III. Evaluation Measures

Evaluation metrics are performed on the discrimination and generation tasks. As the discrimination task is a supervised learning problem, metrics such as accuracy, precision, recall and f1-score are used to evaluate the trained models. This is very different from the generation task on which BLEU (Bilingual Evaluation Understudy) [31], ROUGE (Recall Oriented Understudy for Gisting Evaluation) [32], METEOR (Metric for Evaluation for Translation with Explicit Ordering) [33] and Translation Error Rate (TER) [34] are used for approximate all natural language generation tasks.

IV. Datasets

Compared to other tasks, the datasets for the text paraphrasing task aren't large. This supports the importance of having a robust and generalized text paraphrasing models to help on creating datasets with large diversity for the problem itself and other problems in general such as Sentiment Analysis and Named Entity Recognition.

A. MSCOCO

Microsoft has recently released a dataset for images captions [35]. The dataset comes with 120K images that are captioned with short and medium size texts. For each image, five captions are provided that describe the images, and these five captions are written by different five annotators. As the annotators are describing the same things, the captions could represented as paraphrases to each other.

B. PPDB

PPDB [36], [37] is widely known dataset for paraphrase generation. It comes with wide sizes, however the most used size is PPDB 2.0 Large dataset. For some phrases, PPDB has one-to-many paraphrases.

C. Quora Questions Pairs

Quora questions pairs [38] is a dataset produced by Quora. The dataset contains approximately 400K pairs of sentences. Sentences are questions that are labeled by whether the pair is duplicate, has the same semantic meaning, or not. The duplicate questions are considered to be paraphrases if they are duplicates.

D. SNLI

The SNLI dataset [39] consists of approximately 570k sentences. These sentences are generated by human annotators and manually labeled with either they are entailment, contradiction and neutral. For the paraphrasing task, the focus is on the neutral sentences as they describe the paraphrases sentences to each other. SNLI is heavily used in natural language generation tasks and to capture the semantics of languages [40].

E. WikiAnswers

WikiAnswers [41] is a large question paraphrase corpus created by crawling the WikiAnswers website. The paraphrases are different questions, which were tagged by the users as similar questions. The dataset contains approximately 18M question pairs aligned by word.

V. Applications

Paraphrasing has numerous applications. It is either used as a preprocessing module to increase the datasets as a *Data Augmentation* technique or embedded in an end-to-end model. For instances, paraphrasing is used in Question Answering (QA) systems [42], where several paraphrases are generated in an end-to-end neural-based model for the given question to increase the diversity and the coverage of the input.

Paraphrasing is also used to accurately evaluate the Machine Translation models [43]. This is done by generating paraphrases that are closer in wording to the translation output based on some lexico-semantic resources such as WordNet. Some researches use paraphrasing to improve the the generated logical form of sentences [44] that is known as Semantic Parsing problem and as a representation for the input queries [45] in Semantic Search.

Paraphrasing has also an important application which is plagiarism detection. In this task the target is check whether two texts are copied or altered to another or not. It is a typical paraphrasing application. It is mainly could be used for author copyrights ownership. Natural language processing suffers from a lack of resources and datasets that could be used to train the models [26], such problem decreases the model's generalization. Several researches use paraphrasing as a data augmentation technique to enrich the datasets [46], [47], [48]. Recently, the concept of paraphrasing is used to reformulate the questions to lead for better questions generation and to increase the diversity of questions intents [49].

VI. Proposal: Conditional Text Paraphrasing Task

This work proposes a *Conditional Text Paraphrasing* task as an addition to the original task. As paraphrasing only focuses on the semantics of the sentences, there is a need for more paraphrasing specifications to control the recognition and generation processes. On conditional paraphrasing, the target is to either detect or generate according to specific condition or constraint. The conditions are divided to Morphology-based, Syntax-based and Readability-based categories. This work is driven by several attempts to control the text generation [25], [26]. The target of this work is to set and organize the problem as a task that is closely related to the original task. The research is represented on some ways to control the paraphrase generation by creating a taxonomy that defines the task. This may help other researches when they work on it, also, this taxonomy could be expanded for more concrete tasks if needed.

Feasible sentences generation is considered to be a challenging problem in natural language processing because of several obstacles that occur while modeling text. For instances, representing the text in a way of hidden representation that captures the semantics of text and its structure is hard because of the complexity of text and its language. Controlling the text generation means that we need to make the latent representation of sentence capture the semantics, structure and the other disentangled representation of the embedded attribute. For images, some attempts were done to control the generated image with disentangled attribute [50], [51]. In general, as the nature of text is a discrete data; training the models that depend

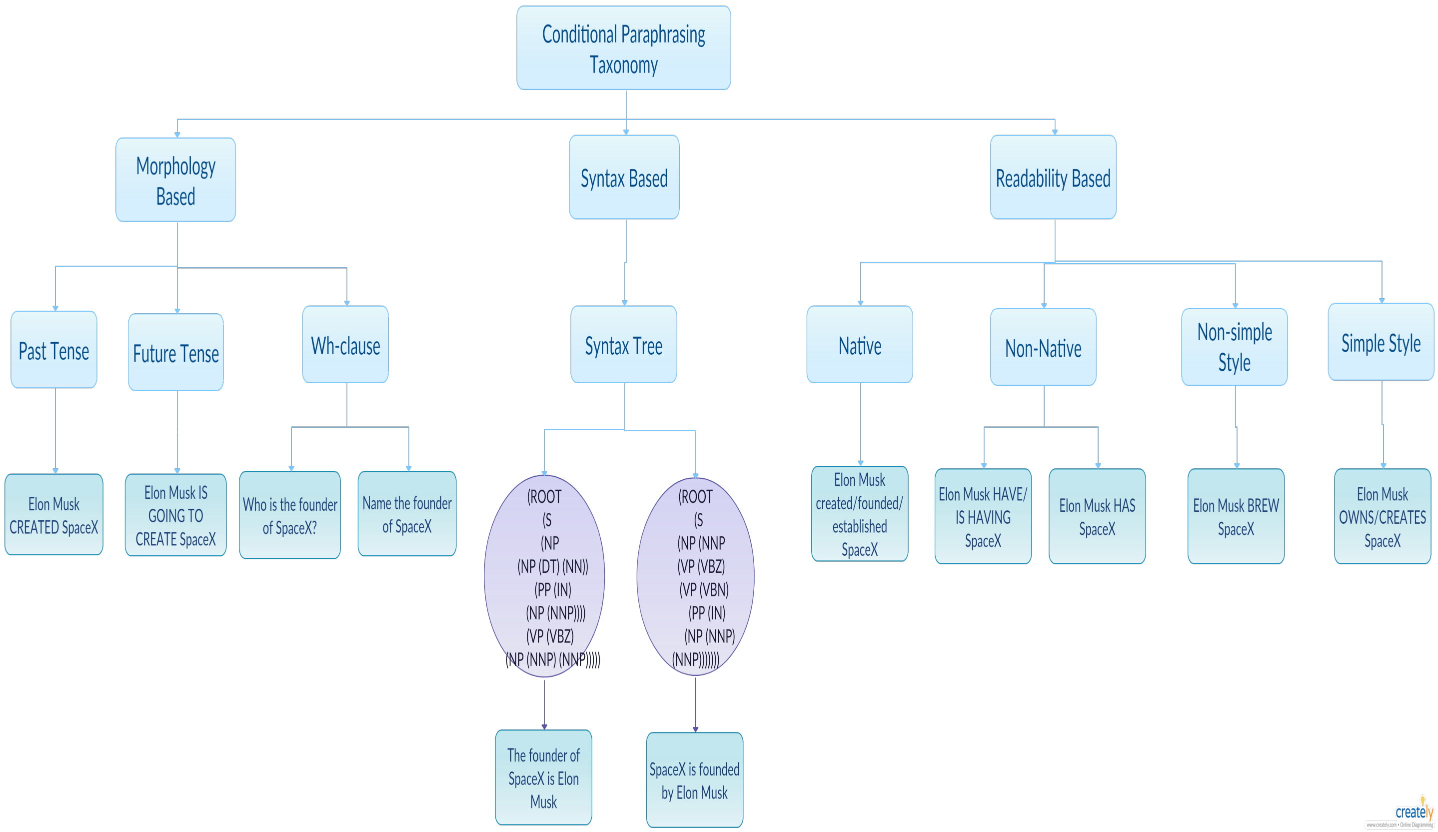


Fig. 4. Conditional Text Paraphrasing Taxonomy. *Reference Text: Elon Musk is the founder of SpaceX*

on optimization and back-propagation becomes hard to reach the global minima. The GAN-based model created by [25] tackled this problem by allocating one dimension of the latent representation to encode the disentangled attributes, such as the polarity in sentiment analysis task, and generates samples with desired sentence semantics. The research focused on disentangled representations of polarity (positive, negative) and Tense (past, present, future) attributes, however, our research believes that this could be generalized further more by expanding the additive attributes that could be conditioned to the model to force it to generate more intensive and concrete samples. These conditions are divided to Morphology-based, Syntax-based and Readability-based categories. These conditioned features control the sentence generation and also the discrimination. In the discrimination task, the conditioned features could be viewed as classes that any paraphrases could be classified to one of them. The taxonomy tree is proposed for the conditional text paraphrasing on Figure 4.

Conditional Text Paraphrasing is divided to three sub-classes. These sub-classes controls the paraphrasing of the output sentences, however, these sub-classes are all considered to be paraphrasing on their semantic meaning. In other words, they are all paraphrases that represent the same meaning, but they differ on their morphology, syntax and readability aspects. For instance, consider the given sentence *Elon Musk is the founder of SpaceX*. The following sentences are candidate paraphrases *Elon Musk created SpaceX*, *Elon Musk is going to create SpaceX*, but these sentences are in the past and future tenses respectively. Also, the sentences *The founder of SpaceX is Elon Musk*, which syntax tree is *(ROOT (S (NP (NP (DT) (NN)) (PP (IN) (NP (NNP)))) (VP (VBZ) (NP (NNP) (NNP)))))*, and *SpaceX is the founded by Elon Musk*, which syntax tree is *(ROOT (S (NP (NNP (VP (VBZ) (VP (VBN) (PP (IN) (NP (NNP) (NNP)))))))*, are paraphrases but each of them have different structure and parse tree. These varieties are divided to three sub-classes, Morphology-based Past Tense, Future Tense, Wh-clause, Syntax-based Syntax-tree and Readability-based Native, Non-native, Simple-style, Non-simple Style. The last sub-class focuses more on the style of writing. As humans, it is easy to us to identify whether the text is complex or written by a non-native person, so, the target is to be able to either generate or detect texts that have different styles.

Formally, for the generation task in conditional paraphrasing, given a reference sentence S_1, where $S_1 = \{w_1, w_2, .., w_n\}$, and in addition, a conditioned class is given as a feature C, the target is to generate candidate(s) sentence(s) that are semantically equivalent to the reference sentence and also applies the conditioning property C.

On the other hand, for the discrimination task, given two sentences (S_1, S_2), where $S_1 = \{w_1, w_2, .., w_n\}$ and $S_2 = \{w_1, w_2, .., w_m\}$, the target is to classify whether they are paraphrases or not, further more, it is possible to check what type of paraphrases that they are conditioned on. This transforms the original classification problem to multi-class classification problem. Shown in Figure 5, which is derived from the above taxonomy, the number of parent classes is two, number of conditioning classes is four and the number of conditioning sub-classes is seven. That makes the total number

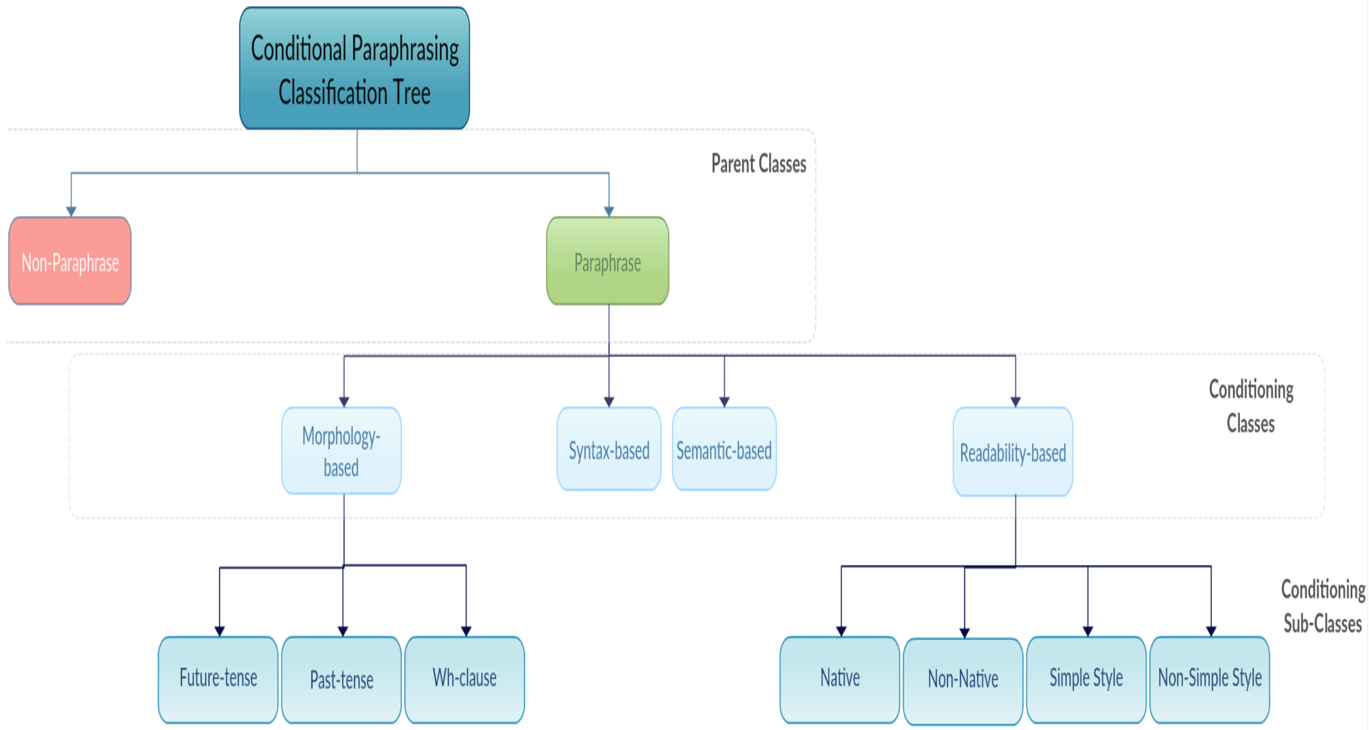


Fig. 5. Conditional Paraphrasing Classification Tree

of classes thirteen.

VII. Conditional Text Paraphrasing Applications

Conditional paraphrasing has several application at which it could be used. Mainly, it could be used in question-answer generation and domain-specific data augmentation. The constraints could be encoded and fed to the generation model as a conditional feature to handle the generated texts like such approaches [25], [26]. Regarding to the discrimination model, the problem turned to be a multi-classification problem. Given that, several transfer learning [52] and multi-task learning [53] could be applied to domain-specific objectives.

Conditional text paraphrasing could be very helpful in learning distributed representations of words and sentence. Currently, supervised learning based models dominates the field of sentences and words representation [40], so, the constraints, specifically syntax-based features, that are provided could help in much better representations for text that would effect on better natural language understanding systems.

VIII. Conclusion

This work showed a survey for text paraphrasing and its recent researches and efforts that work in the directions of paraphrase generation and discrimination. Also, it proposed our definition and taxonomy of conditional text paraphrasing task and how its great impact on existed applications and problems. For the future work, we are looking forward to working on datasets for conditional text paraphrasing, alongside doing several experiments for this task.

References

[1] S. Fernando and M. Stevenson, "A semantic similarity approach to paraphrase detection," in *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, 2008, pp. 45–52.

[2] V. Ngoc Phuoc An, S. Magnolini, and O. Popescu, "Paraphrase identification and semantic similarity in twitter with simple features," 2015.

[3] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, 2003, pp. 133–142.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[5] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[6] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in neural information processing systems*, 2015, pp. 3294–3302.

[7] C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B. Dhoedt, "Learning semantic similarity for very short texts," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1229–1234.

[8] C. Brockett and W. B. Dolan, "Support vector machines for paraphrase identification and corpus construction," in *Proceedings of the third international workshop on paraphrasing (IWP2005)*, 2005.

[9] B. Agarwal, H. Ramampiaro, H. Langseth, and M. Ruocco, "A deep network model for paraphrase detection in short text messages," *Information Processing & Management*, vol. 54, no. 6, pp. 922–937, 2018.

[10] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[11] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2042–2050. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969055

[12] W. Yin and H. Schütze, "Convolutional neural network for paraphrase identification," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 901–911.

[13] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity." in *AAAI*, vol. 16, 2016, pp. 2786–2792.

[14] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 539–546.

[15] A. Fader, L. Zettlemoyer, and O. Etzioni, "Open Question Answering Over Curated and Extracted Knowledge Bases," in *KDD*, 2014.

[16] P. A. Duboue and J. Chu-Carroll, "Answering the question you wish they had asked: The impact of paraphrasing for question answering," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, 2006, pp. 33–36.

[17] Z. Hu, Z. Yang, R. Salakhutdinov, and E. P. Xing, "On unifying deep generative models," *arXiv preprint arXiv:1706.00550*, 2017.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[21] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[22] N. Duan, D. Tang, P. Chen, and M. Zhou, "Question generation for question answering," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 866–874.

[23] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[24] L. Yu, W. Zhang, J. Wang, and Y. Y. Seqgan, "sequence generative adversarial nets with policy gradient. arxiv preprint," *arXiv preprint arXiv:1609.05473*, vol. 2, no. 3, p. 5, 2016.

[25] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," *arXiv preprint arXiv:1703.00955*, 2017.

[26] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," *arXiv preprint arXiv:1804.06059*, 2018.

[27] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[28] A. Prakash, S. A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu, and O. Farri, "Neural paraphrase generation with stacked residual lstm networks," *arXiv preprint arXiv:1610.03098*, 2016.

[29] A. Gupta, A. Agarwal, P. Singh, and P. Rai, "A deep generative framework for paraphrase generation," *arXiv preprint arXiv:1709.05074*, 2017.

[30] Z. Li, X. Jiang, L. Shang, and H. Li, "Paraphrase generation with deep reinforcement learning," *arXiv preprint arXiv:1711.00279*, 2017.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[32] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[33] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[34] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*, vol. 200, no. 6, 2006.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[36] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "Ppdb: The paraphrase database," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 758–764.

[37] E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, 2015, pp. 425–430.

[38] Z. Chen, H. Zhang, X. Zhang, and L. Zhao, "Quora question pairs," 2018.

[39] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[40] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[41] A. Fader, L. Zettlemoyer, and O. Etzioni, "Paraphrase-driven learning for open question answering," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 1608–1618.

[42] L. Dong, J. Mallinson, S. Reddy, and M. Lapata, "Learning to paraphrase for question answering," *arXiv preprint arXiv:1708.06022*, 2017.

[43] D. Kauchak and R. Barzilay, "Paraphrasing for automatic evaluation," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 455–462.

[44] J. Berant and P. Liang, "Semantic parsing via paraphrasing," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2014, pp. 1415–1425.

[45] P. Arthur, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Semantic parsing of ambiguous input through paraphrasing and verification," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 571–584, 2015.

[46] R. Barzilay and K. R. McKeown, "Extracting paraphrases from a parallel corpus," in *Proceedings of the 39th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2001, pp. 50–57.

[47] C. Bannard and C. Callison-Burch, "Paraphrasing with bilingual parallel corpora," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 597–604.

[48] C. Callison-Burch, "Syntactic constraints on paraphrases extracted from parallel corpora," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 196–205.

[49] C. Buck, J. Bulian, M. Ciaramita, W. Gajewski, A. Gesmundo, N. Houlsby, and W. Wang, "Ask the right questions: Active question reformulation with reinforcement learning," *arXiv preprint arXiv:1705.07830*, 2017.

[50] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[51] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, 2016, pp. 2172–2180.

[52] S. J. Pan, Q. Yang *et al.*, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[53] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

TokenVote: Secured Electronic Voting System in the Cloud

Fahad Alsolami
Department of Information Technology
King Abdulaziz University , KSA

***Abstract*—With the spread of democracy around the world, voting is considered a way to collectively make decisions. Recently, many government offices and private organizations use voting to make decisions when the opinions of multiple decision makers must be accounted for. Another advancement: cloud computing attracts many individual and organizations due to low cost, scalability, and the ability to leverage big data. These considerations motivate our proposal of the TokenVote scheme. TokenVote is an electronic voting system in the cloud that uses revocable fingerprint biotokens with a secret sharing scheme to provide privacy, non-repudiation, and authentication. The TokenVote scheme spreads shares of secret (vote), embeds them inside the encoding biometric data (i.e. fingerprint), and distributes them over multiple clouds. During the voting process, each voter must provide his/her fingerprint, causing the TokenVote scheme to collect all voting shares from all voters to compute the final voting result. TokenVote does cloud parallel computing for the voting process in an encoded mode to prevent disclosure of the shares of voting and the fingerprint itself. Our experiments show that TokenSign has a significant performance and comparable accuracy when compared with two baselines.**

Keywords*—*Cloud; Fingerprint; Voting; Security

I. Introduction

With the worldwide spread of democracy, voting is no longer an alternative; it is a necessity. The importance of voting is recognized by many individuals, organizations, and countries, not only for presidential elections, but also for making decisions in organizations. Making decisions within organizations has been made more difficult since many decisions must be made through consensus rather than by a single decision maker [1] [2]. Rather than allowing a debate in a meeting to continue unresolved for a long time, people can make use of an electrotonic system to speed up the process. Many electronic voting systems have been introduced in literature. Gibbard [3] proves the conjecture by proving a game theory where a voting scheme is a game form and individual actions are strategies, while Sako et al [4] proposes a voting scheme that proves a receipt-free system. For secure electronic voting system, Fujioka et al [5] proposes a secure voting scheme where the scheme ensures that no one can disclose the intermediate voting result. In addition, Peralto et al [6] proposes a computerized voting system that can identify participants and prevent duplicated and fraudulent voting.

Despite the great advantages of electronic voting systems, security and privacy are serious problems for voters and organizations/administrations. Thus, many electronic voting schemes are proposed in literature to provide security and privacy. Okamoto [7] proposes a secured voting scheme which solves fairness, anonymity, receipt, and privacy. Lee et al [8] proposes an electronic voting protocol that achieves receipt-freeness. Bannet et al [9] presents a voting system that demonstrates many bugs that can occur in a voting system. Other voting systems using biometrics data for authentication purposes have also been proposed [10] [11] [12] [13] [14]. Some use fingerprint data, while others use both fingerprint and face recognition. Other schemes use a secret sharing scheme [15] to distribute the voting secret among all voters [16] [17] [18]. Even though these schemes have solved many issues in electronic voting system, other issues remain research challenges.

In this paper, we propose the TokenVote scheme, which provides security and privacy for both voters and administrations. TokenVote scheme has multiple purposes, such as presidency elections, organizations elections, and formal meeting decision-making in both government and private organizations. Specifically, our contribution is to design, implement, and evaluate a TokenVote scheme that uses the revocable fingerprint biotokens (Biotope) [19], Bipartite token [20], and the secret-sharing scheme [15]. During the enrollment process, TokenVote encodes the biometric data (i.e. fingerprint). Then, TokenVote embeds a shared secret (i.e. voting) inside the encoded fingerprint data. TokenVote then distributes all shares of a vote over multiple clouds, so no single cloud stores the threshold required to recover the result of a vote. During the voting process, TokenVote matches the fingerprint data of voters in encoded mode, then computes the final voting result. This whole process being conducted in encoding form which provides security and privacy for voters. On the other hand, voters must provide their fingerprint data to vote which provides non-repudiation and authentications for administrations.

The rest of this paper is organized as follows: in section II, we briefly describe previous related work. The objectives of TokenVote are given in section III. Our proposed TokenVote algorithm is presented in section IV. In section V, the description of the experimental design is given. The experimental evaluation and results are provided in section VI. Finally, the conclusion is drawn in section VII.

II. Background

A. Voting system and security concerns

A great deal of literature has been devoted to designing a voting system that satisfy all purposes. Gibbard [3] proves the conjecture by proving a game theory where a voting scheme is set up in a game form and individual actions are strategies. The author discusses manipulation in a voting system and how to prove it. Fujioka et al [5] proposes a secure voting scheme that provides privacy for voters, ensuring voting fairness. The scheme ensures that no one can disclosure the intermediate voting result. Sako et al [4] proposes a voting scheme that proves to be receipt-free, so voters can hide their votes from a powerful adversary. To achieve their goal, they replace the physical voting booth with a hardware assumption. Peralto et al [6] invents a computerized voting system that can identify participants and prevent duplicate and fraudulent voting. Jakobsson et al [21] proposes a new solution that provides proof of correct operation of the voting system. They use randomized partial checking to check the subset of input/output data instead of completely correct operation. However, to provide secure voting system, Okamoto [7] proposes a secured voting scheme that solves fairness, anonymity, receipt, and privacy. To achieve its goal, the scheme has four steps: authorization, voting, claiming, and counting. Lee et al [8] proposes an electronic voting protocol that achieves receipt-freeness. Bannet et al [9] presents a voting system to demonstrate many bugs that might occur in a voting system.

B. Using Biometric Authentication in Voting System

Authentication in a voting system proves that the participant giving his/her vote is the right person. Many researchers discuss the use of biometrics data as authentication tools in voting system. Ahammad et al [14] proposes an electronic voting machine based on fingerprint identification to provide security for voting system. Their system consists of four phases: enrollment, voting, election result demonstration, and database restoration. In enrollment phase, voters enroll their fingerprint data in the voting system to do matching in voting phase when voters come to vote. In the election result phase, the voting system calculates all votes; in the database restoration phase, the system deletes the current voting result for next voting event. Hof [12] proposes electronic voting with biometric authentications. They evaluate all biometrics (fingerprint, iris, face etc.) against another authentication tool, like a password or card. They also discuss the security issues related to biometric data. Kumar et al [10] proposes an electronic voting system based on fingerprint authentication. Their system requires successful matching of fingerprints to vote. Kumar et al [13] proposes a framework for electronic voting that maintains authentication security using fingerprint. Najam et al [11] proposes an electronic voting system based on fingerprint and face recognition.

C. Using Secret Sharing Scheme for Voting System

Secret sharing schemes have been used in electronic voting, as documented in literature. Schoenmakers [17] uses secret sharing schemes with cryptographic tools to secure electronic voting. Their electronic voting has two protocols: distribution and reconstruction protocols. The distribution protocols have two steps. The distribution step allows the dealer to create and distribute the shares of a secret among all voters. The verification-of-the-shares step allows any participant to use the public key of the encryption method to verify the share. Similarly, the reconstruction protocols have two steps: decrypting the share and pooling the share. Nair et al [18] proposes an electronic voting system (EVS) that uses a secret sharing scheme and secures multi party computation to provide security. The electronic voting system (EVS) has four modules: polling station, communication server, chief election commissioner, and collection center. The polling station has the voting machines and voting panel. The chief election is responsible for managing the candidate information in the voting panel. The communication server manages all activities and coordinates all modules. The collection center manages the collection centers. Liu et al [16] proposes an electronic voting scheme that uses a secret sharing scheme and k-anonymity to provide security and coercion-resistance. Their scheme ensures voters can verify the correctness without knowing others information.

III. TokenVote Objectives

The main goal of this paper is to explore a cloud electronic voting scheme which protects not only the voting information but also the biometric data (i.e. fingerprint). TokenVote protects each vote during its journey from the voter to administration who computes and declares the final voting. In this section, we explore the objectives of TokenVote in privacy, non-repudiation, and authentication.

A. Non-Repudiation and Authentication

TokenVote scheme uses biometric data (i.e. fingerprint) and a secret sharing scheme [15] to achieve its goal in non-repudiation and authentication. In the voting process, a user must provide his/her biometric data (i.e. fingerprint). Thus, a voter cannot deny his/her vote, providing non-repudiation. This strengthen the TokenVote scheme as any organizations/administrations can make sure any voter cannot deny his/her voting at a later time. Regarding the authentication objective, biometric data (i.e. fingerprint) is considered a highly regarded authentication tool for organizations. TokenVote scheme requires that any voter must enroll his/her fingerprint to participate in voting. In the voting process, the voters must provide their fingerprint again for matching, authentication and participation during voting. This allows organizations/administrations to verify who has participated in voting.

B. Security and Privacy

TokenVote provides security and privacy for each voters voting information and for the biometric data (i.e., fingerprint). TokenVote scheme uses bipartite tokens [20] to do matching/voting in an encrypted domain. For each voting share, TokenVote scheme uses secret sharing scheme to split the vote into multiple shares where each share is in encoded mode during all of voting process. No one knows the information hidden inside each voting share from the start stage until the last stage of voting process. The TokenVote provides security and privacy, not only for voting information but also for biometric data (i.e., fingerprint).

C. Scalability and Performance

The TokenVote scheme uses the cloud for many objectives including scalability and performance. The TokenVote scheme uses secret sharing scheme to split each vote into multiple shares of a vote where the size of each share is equal to the original size of a vote. The increased data size needs a scalable environment for computing. Thus, cloud computing

Plain Text Data			Encrypted Data			Share of a vote
k	j	θ_{kj}	d_{kj}	β_1	β_2	

Fig. 1. The pair table data layout stores in the cloud. Each row of the pair table data contains plain text data (k, j, θ_{kj}), encoding data (d_{k_j}, β_1, β_2) and share of vote

is a great option for big data. Therefore, the TokenVote scheme distributes all shares of a vote over multiple clouds where no single cloud stores the whole shares of a vote. In matching/voting process, the TokenVote scheme uses threading to do matching/voting in parallel, resulting in improved voting performance.

IV. Design of TokenVote Scheme Algorithm

In our design, we present the architecture of TokenSign scheme in enrollment and matching/signing process. The TokenSign scheme consist of two protocols: single protocol and group protocol. Single protocol is used to perform a signature for one person while group protocol is used to perform a signature for a group of people.

A. Enrollment Process

First, TokenVote creates a vote (secret) that is distributed among all participants. Second, the TokenVote scheme uses secret sharing scheme [15] to split and distribute the vote (secret) between all voters where each voter has one share of the vote (secret). In this distribution, the TokenVote determines the maximum total shares, set equal to the maximum number of voters, and the threshold shares, set equal to the minimum number of voters. Third, the TokenVote scheme collects the fingerprint data from voters and applies the NIST Bozorth Matcher Algroithm [22] to create minutia files and pair-tables. The TokenVote scheme uses the revocable fingerprint biotokens (Biotope) [19] to transform the plaintext of fingerprint data (i.e., pair-table) in the encrypted fingerprint data. The TokenVote applies Bipartite token [20] to store each share of vote inside a voters encrypted fingerprint data. Finally, TokenVote distributes each share of votes (Biotoken share) in multiple clouds where each cloud does not store all shares of votes that needs to recover the final voting result. Algorithm 1 shows the details of the TokenVote enrollment process steps and figure 1 shows the layout share of vote store in the cloud.

B. Voting Process

In the voting process, the TokenVote scheme requires collection of fingerprint data from each voter who is enrolled as his/her fingerprint data is needed for the individual to participate in a vote. Second, the TokenVote scheme follows the same steps in the enrollment process where it creates minutia points, minutia files, pair-tables, and encodes the pair-tables for the probe fingerprint data. Then, the TokenVote scheme matches the encoding probe pair-table against an encoding gallery pair-table for all voters simultaneously. If the authenticating matching is successful between a voters probe and gallery fingerprint data, the share of voting is released

Data: 1. Gallery fingerprint image g_i, Where i=1,2,3,. . . ,n.
2. Voting Secret v_i.
Result: 1. Encrypted gallery fingerprint (pair-table t_i).
2. Share of Vote hidden inside Encrypted Gallery Fingerprint.
for (*each Voting Secret v_i && each gallery fingerprint image g_i*) {
  for (*each Voting Secret v_i*) {
    split the voting secret v_i into multiple shares N_i using SSS [15];
    determine all shares N_i and the threshold K_i using SSS [15] ;
    check (shares N_i == maximum number of voters R_i);
    check (threshold K_i == minimum number of voters R_i);
    check (shares N_i >= threshold K_i);
  for (*each gallery fingerprint impression g_i*) {
    create minutia points m_i using NIST Bozorth [22];
    create minutiae file f_i using NIST Bozorth [22];
    create the gallery pair-table t_i using NIST Bozorth [22];
    encode the gallery pair-table t_i using Biotope [19];
    hide a secret of voting share N_i inside the encoding gallery pair-table t_i using Bipartite [20] ; upload the encoding gallery fingerprint (pair-table t_i) to the cloud;

Algorithm 1: Enrollment process algorithm of TokenVote scheme

Data: Probe fingerprint image p_i where i=1,2,3,. . . ,n.
Result: Vote Result.
for (*each probe fingerprint impression p_i*) {
  create minutia points m_i using NIST Bozorth [22];
  create minutiae file f_i using NIST Bozorth [22];
  create the probe pair-table t_i using NIST Bozorth [22];
  encode the probe pair-table t_i using Biotope [19];
for (*all encoding probe pair-table t_i && encoding gallery pair-table t_i in the cloud*) {
  match each encrypted probe pair-table t_i in parallel against all encrypted gallery pair-table t_i ;
  if *(match == true)* **then**
    release the shared vote for each voter;
    collect all shared votes from all voters;
  if *(shared votes >= minimum number of voters R_i)* **then**
    compute the all shared votes from all voters using SSS [15];
    declare the voting;
Return the vote result.

Algorithm 2: Voting process algorithm of TokenVote

TABLE I. THE SPEED RESULTS WHERE THE VALUE OF P-VALUE FROM T-TEST REJECTS THE NULL HYPOTHESIS HO

	Cloud-ID-Screen	Bipartite	TokenVote
AVE	26.836	26.93	19.93
STD	0.22	0.31	0.24

for this voter. The TokenVote scheme conducts this process in parallel for all voters and releases all shares of voting from all voters. Finally, the TokenVote scheme applies a secret sharing scheme [15] to compute all shares of voting. If the number of shares is greater than or equal to the minimum number of voters (threshold shares of voting), the TokenVote scheme releases the voting secret and declares the voting result. Algorithm 2 shows the details of TokenVote voting process steps.

V. EXPERIMENTAL DESIGN

In our experiment we design a decision-making scenario in an organization where the decision has been taken from three levels of management. Each level of management has ten people as decision makers who propagate their voting decision from a lower level group to the upper level group. At the third level which is the final stage, the voting is computed and released. We compare our scheme against the Bipartite Biotokena algorithm [20] as the first baseline and Cloud-ID-Screen [23] as the second baseline. We conduct our experiment in Amazon Web Service cloud, so we use eight clouds: North Virginia, North California, Ohio, London, Paris, Ireland, Tokyo, and Sydney. During the enrollment process, TokenVote uses the fingerprint dataset (FV C2002Db2 a) [24] and follows the steps as explained in Section 4-A. Then, the TokenVote scheme uses the programming lanaguages C++ and Python to upload the gallery encoded fingerprint data to multiple AWS S3 cloud storages. The uploading process is done in parallel by using threading. During the voting process, we connect Amazon AWS S3 with Amazon AWS EC2 instance by using the Python boto library to do parallel matching. Finally, we did the matching/voting process twenty times in parallel and took the average of all these runs.

VI. EXPERIMENT EVALUATION

In our hypothesis: we aim to prove that if we distribute a vote among multiple people, we can do parallel voting to speed up the voting process while getting comparable accuracy against the baselines. To prove our hypothesis, we run two experiments, accuracy and performance, and evaluate the results.

A. Accuracy Evaluation

We run our accuracy experiment to evaluate the false accept rate (FAR) and the genuine accept rate (GAR) for both the baseline and TokenVote scheme. Then, we compare the result of our TokenVote scheme against the two baselines. Figure 2 shows the ROC curve result where our TokenVote scheme achieves GAR equal to 97 and FAR equal to 0. This promising result proves our hypothesis.

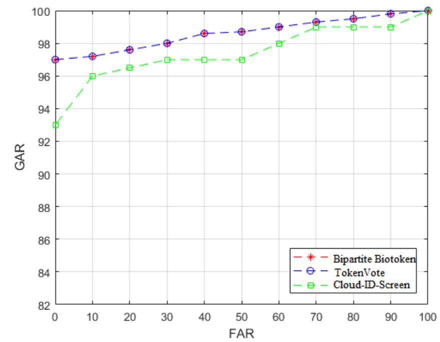


Fig. 2. The ROC curve comparing accuracy of our TokenVote scheme against the two baselines

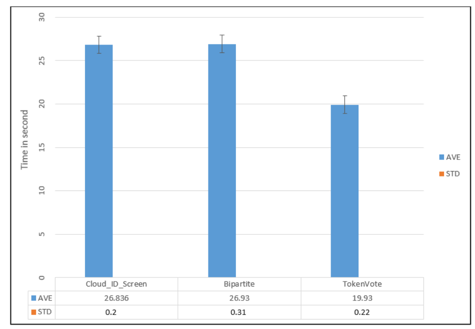


Fig. 3. The speed result comparing the performance of our TokenVote scheme against the two baselines

B. Performance Evaluation

We run an identification (1:N) experiment to evaluate the performance of our TokenVote scheme against the two baselines. We run our experiment in parallel for our TokenVote scheme and the two baselines. Then, we compare the result of our TokenVote scheme against the two baselines. Figure 3 shows the speed result where our TokenVote scheme achieves promising result over the two baselines. For formal testing, the null hypothesis Ho is that the time of matching for the baseline is less than or equal to TokenVote, i.e., the baseline performs better. Table I shows the speed results where the value of P-value from t-test rejects the null hypothesis Ho and proves our claim that the TokenVote achieves better performance over the baseline.

VII. CONCLUSION

In this paper we design, implement, and evaluate the TokenVote scheme. The TokenVote is a cloud electronic voting system uses the revocable fingerprint biotoken and secret sharing scheme. Thus, TokenVote provides privacy and security not only for the voters but also for administration. Moreover, TokenVote uses cloud computing and threading to provide scalability and performance. For future work we will use smart devices for voting since all smart devices use biometric for authentication.

REFERENCES

[1] F. C. Lunenburg, "Decision making in organizations." *International journal of management, business, and administration*, vol. 15, no. 1, pp. 1–9, 2011.

[2] P. Drucker, "The effective executive." 2016.

[3] A. Gibbard, "Manipulation of voting schemes: A general result," *Econometrica*, vol. 41, pp. 587–601, 1973.

[4] K. Sako and J. Kilian, "Receipt-free mix-type voting scheme: A practical solution to the implementation of a voting booth," in *Proceedings of the 14th Annual International Conference on Theory and Application of Cryptographic Techniques*, ser. EUROCRYPT'95. Berlin, Heidelberg: Springer-Verlag, 1995, pp. 393–403.

[5] A. Fujioka, T. Okamoto, and K. Ohta, "A practical secret voting scheme for large scale elections," in *Proceedings of the Workshop on the Theory and Application of Cryptographic Techniques: Advances in Cryptology*, ser. ASIACRYPT '92. London, UK, UK: Springer-Verlag, 1993, pp. 244–251.

[6] P. Ryan G, "Computerized voting system," Mar 1999.

[7] T. Okamoto, *An electronic voting scheme*. Boston, MA: Springer US, 1996, pp. 21–30.

[8] B. Lee and K. Kim, "Receipt-free electronic voting scheme with a tamper-resistant randomizer," in *Proceedings of the 5th International Conference on Information Security and Cryptology*, ser. ICISC'02. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 389–406.

[9] J. Bannet, D. W. Price, A. Rudys, J. Singer, and D. S. Wallach, "Hack-a-vote: Security issues with electronic voting systems," *IEEE Security Privacy*, vol. 2, no. 1, pp. 32–37, Jan 2004.

[10] D. A. Kumar, T. Ummal, and S. Begum, "A novel design of electronic voting system using fingerprint," 2011.

[11] S. N. Syed, A. Z. Shaikh, and S. Naqvi, "A novel hybrid biometric electronic voting system: Integrating finger print and face recognition," *CoRR*, vol. abs/1801.02430, 2017.

[12] S. Hof, "E-voting and biometric systems," 2004.

[13] S. Kumar and M. Singh, "Design a secure electronic voting system using fingerprint technique." *International Journal of Computer Science Issues (IJCSI)*, vol. 10, no. 4, 2013.

[14] e. a. Ahammad, Ifthekhar, "Towards a secure and automated platform for fingerprint-based electronic voting machine." *International Journal of Intelligent Machines and Robotics*, vol. 1, no. 1, pp. 34–44, 2018.

[15] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.

[16] Q. Zhao and Y. Liu, "E-voting scheme using secret sharing and k-anonymity," in *Advances on Broad-Band Wireless Computing, Communication and Applications*, L. Barolli, F. Xhafa, and K. Yim, Eds. Cham: Springer International Publishing, 2017, pp. 893–900.

[17] B. Schoenmakers, "A simple publicly verifiable secret sharing scheme and its application to electronic voting," in *In CRYPTO*. Springer-Verlag, 1999, pp. 148–164.

[18] D. G. Nair, V. P. Binu, and G. S. Kumar, "An improved e-voting scheme using secret sharing based secure multi-party computation," *CoRR*, vol. abs/1502.07469, 2014.

[19] T. E. Boult, W. J. Scheirer, and R. Woodworth, "Revocable fingerprint biotokens: accuracy and security analysis," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[20] W. J. Scheirer and T. E. Boult, "Bipartite biotokens: Definition, implementation, and analysis," in *Advances in Biometrics*, M. Tistarelli and M. S. Nixon, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 775–785.

[21] M. Jakobsson, A. Juels, and R. L. Rivest, "Making mix nets robust for electronic voting by randomized partial checking," in *Proceedings of the 11th USENIX Security Symposium*. Berkeley, CA, USA: USENIX Association, 2002, pp. 339–353.

[22] C. Watson, M. Garris, E. Tabassi, C. Wilson, R. McCabe, S. Janet, and K. Ko, *User's Guide to NIST Biometric Image Software (NBIS)*, 2007.

[23] F. Alsolami, B. Alzahrani, and T. Boult, "Cloud-id-screen: Secure fingerprint data in the cloud," in *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, Jan 2018, pp. 1–8.

[24] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, "Handbook of fingerprint recognition," 2009.

A New Uncertainty Measure in Belief Entropy Framework

Moïse Digrais Mambé[1,4], Tchimou N'Takpé[2,4], Nogbou Georges Anoh[3,4], Souleymane Oumtanaga[1,4]
[1]Institut National Polytechnique Félix Houphouët-Boigny, Yamoussoukro, Côte d'Ivoire
[2]Université Nangui Abrogoua, Abidjan, Côte d'Ivoire
[3]Université Virtuelle de Côte d'Ivoire, Abidjan
[4]Laboratoire de Recherche en Informatique et Télécommunication, Abidjan, Côte d'Ivoire

***Abstract*—Belief entropy, which represents the uncertainty measure between several pieces of evidence in the Dempster-Shafer framework, is attracting increasing interest in research. It has been used in many applications and is mainly based on the theory of evidence. To quantify uncertainty, several measures have been proposed in the literature. These measures, sometimes in extended or hybrid forms, use the Shannon entropy principle to determine uncertainty degree. However, the failure to consider the scale of the frame of discernment framework remains an open issue in quantifying uncertainty. In this paper, we propose a new uncertainty measure that takes into account the power set of the frame of discernment. After analysing the different existing methods, we show the performance and effectiveness of our proposed approach.**

***Keywords*—Dempster Shafer Theory; Belief entropy; Uncertainty; Information management; Deng entropy**

I. Introduction

In recent years, there have been increasing improvements in the management of uncertainty issues in information systems [1], [2], [3], [4]. Several theories including Shannon's entropy [5], probability theory [6], [7], possibility theory [8], [9] and Dempster-Shafer evidence theory [10], [11] have been developed. Dempster-Shafer theory in particular, provides effective tools for modelling and processing uncertain information [12], [13], [14], [15]. It has been widely used in various applications including information fusion [16], [17], decision making [18], [19], diagnosis and fault detection [20], [21], target recognition [22], [23], and so on. In applying this theory, many challenges are increasingly being addressed. We can mention, conflict management through different sources of information [24], [25], Consideration of the relationship between various pieces of evidence before the fusion step [26], [27], the problem in body of evidence (BOE) generation [28], [29] and finally, the consideration of the frame of discernment (FOD) in uncertainty management [24], [30].

In the Dempster-Shafer framework, there are several hybrid or extended uncertainty measures. These measures use the Shannon entropy principle for the uncertainty quantifiying. Also, these measures include Hohle's confusion measure [31], Dubois & Prade's weighted Hartley entropy [32], Yager's dissonance measure [33], Klir & Ramer's discord measure [34], Klir & Parviz's strife measure [35], George & Pal's conflict measure [36], Deng entropy [37] and modified Deng entropy proposed by Zhou et al. [30].

In this work, we are mainly interested in Deng entropy and modified Deng entropy. The Deng entropy, a very effective measure, compared to various measures in some cases, has been used in several fields of application[38], [37]. However, one of the main limitations of this uncertainty measure is related to not considering the scale of FOD. To address this limit, some authors, including Zhou et al., have proposed a modified Deng measure [30]. Nevertheless, although effective in some cases, the problem related to the scale of FOD is still perceptible. In this paper, we propose a new uncertainty measure by extending the modified Deng entropy. This new measure improves the performance of the measures proposed by Deng and Zhou takes into consideration the power set of the FOD. After analysing the different existing methods, we show the performance and effectiveness of our proposed approach.

The rest of this document is organized as follows : section 2 provides a brief overview of Dempster-Shafer theory and Shannon entropy. Section 3 presents some uncertainty measures in the Dempster-Shafer framework and and some limitations. Section 4 describe the new uncertainty measure. Section 5 presents, using numerical examples, the effectiveness of the new measure. The conclusion and some perspectives related to this work are presented in Section 6.

II. Preliminaries

A. Dempster-Shafer Theory

Dempster-Shafer Theory [10], [11], also known as belief theory or evidence theory, has many advantages for processing uncertain information. We present some basic concepts related to this theory.

1) Formalism: Let $\Omega = \{\omega_1, \ldots, \omega_N\}$ a set of N mutually exclusive and exhaustive events. Ω represents the frame of discernement FOD. A mass function is defined on the power set of Ω, noted 2^Ω with :

$$2^\Omega = \{\emptyset, \{\omega_1\}, \ldots, \{\omega_N\}, \ldots, \{\omega_1, \ldots, \omega_i\}, \ldots, \Omega\} \quad (1)$$

In Ω, a mass function assigns to each subset a value between 0 and 1 representing its elementary belief mass defined by :

$$m : 2^\Omega \rightarrow [0, 1] \quad (2)$$

Such as :

$$m(\emptyset) = 0 \quad \text{et} \quad \sum_{A \subseteq \Omega} m^\Omega(A) = 1 \quad (3)$$

When $m(A) > 0$, A is called a focal element. Belief is the amount of trust that supports a hypothesis A of the power set 2^{Ω}. It is most often called the Body Of Evidence BOE or Basic Probability Assignment BPA or Basic Belief Assignment BBA [39] and is characterized by all the focal elements and their associated mass value (Eq.4):

$$(\aleph, m) = \{(A, m(A)) : A \in 2^{\Omega}, m(A) > 0\} \tag{4}$$

where $\aleph$ represents a subset of the power set 2^{Ω} and each proposition $A \in \aleph$ are focal elements.
A BOE can also be represented by its associated belief Bel and plausibility Pl functions defined as follows :

$$Bel(A) = \sum_{0 \neq B \subseteq A} m(B) \quad \text{and} \quad Pl(A) = \sum_{B \cap A = \phi} m(B) \tag{5}$$

2) Combination rules: In Dempster-Shafer theory, two independent mass functions, noted m_1 and m_2, can be combined with the Dempster's combination rule [40], defined as follows:

$$m(A) = \frac{1}{1-k} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{6}$$

where k represents the degree of conflict between m_1 and m_2. k is defined as follows:

$$k = \sum_{B \cap C = \phi} m_1(B) m_2(C) \tag{7}$$

B. Shannon entropy

In information theory, Shannon entropy (E_s) [5] is used to measure the volume of information in a system, process or message. This measure determines the expected value of the information contained in a message. The measure is defined as follows :

$$E_s(m) = -\sum_{i=0}^{n} p_i \log_b p_i \tag{8}$$

where n represents the quantity or number of basic states, p_i represents the probability of the state i with $\sum_{i=0}^{n} p_i = 1$, b is the basis of the logarithm, it most often takes the value 2.

III. Related Works

In this section, we present some uncertainty measures in the Dempster-Shafer framework. In these measures, X represents FOD, A and B are the focal elements. $\mid A \mid$ refers to the cardinality of A.

A. Uncertainty measures

Some uncertainty measures are represented in the table I. These measures include Hohle's confusion measure [31], Dubois & Prade's weighted Hartley entropy [32], Yager's dissonance measure [33], Klir & Ramer's discord measure[34], Klir & Parviz's strife measure [35], George & Pal's conflict measure [36], Deng entropy [37] and modified Deng entropy proposed by Zhou et al. [30]. In this study, we are particularly interested in the entropies proposed by Deng and Zhou et al.

B. Problem formulation

In Demspter-Shafer theory, uncertain information should not only be modelled by mass functions, FOD is also a source of uncertainty [38]. This paper recall this problem by using the Zhou et al.'s [30] example. Some measures such as Deng entropy (E_d) and Modified Deng entropy (E_z) are calculated.

Example 3.2: Consider two $BOEs$ m_1 and m_2, representing respectively results of two reliable sensors in a target identification problem as follows :

$$m_1 : m_1(\{a, b\}) = 0.4,\ m_1(\{c, d\}) = 0.6$$
$$m_2 : m_2(\{a, c\}) = 0.4,\ m_2(\{a, b\}) = 0.6$$

Deng entropy is calculated as follows :

$$\begin{aligned} E_d(m_1) &= -\sum_{A \subseteq X} m_1(A) \log_2\left(\tfrac{m_1(A)}{2^{|A|}-1}\right) \\ &= -0.4 \times \log_2\left(\tfrac{0.4}{2^2-1}\right) - 0.6 \times \log_2\left(\tfrac{0.6}{2^2-1}\right) \\ E_d(m_1) &= 2.5559 \end{aligned} \tag{9}$$

$$\begin{aligned} E_d(m_2) &= -\sum_{A \subseteq X} m_2(A) \log_2\left(\tfrac{m_2(A)}{2^{|A|}-1}\right) \\ &= -0.4 \times \log_2\left(\tfrac{0.4}{2^2-1}\right) - 0.6 \times \log_2\left(\tfrac{0.6}{2^2-1}\right) \\ E_d(m_2) &= 2.5559 \end{aligned} \tag{10}$$

Despite the difference in $FODs$ $(i.e. X_1 = \{a, b, c, d\}$ et $X_2 = \{a, b, c\})$, The results obtained from the Deng measure about the BOE m_1 are similar to the uncertainty measure of BOE m_2. Intuitively, the uncertainty measure of BOE m_1 should be bigger than that of BOE m_2. This, because the FOD related to the BOE m_1 contains more elements than the FOD related to the BOE m_2. At this main limitation, some authors, notably Zhou et al. [30] have proposed the modified Deng entropy.

Modified Deng entropy is calculated as follows :

$$\begin{aligned} E_z(m_1) &= -\sum_{A \subseteq X} m_1(A) \log_2\left(\tfrac{m_1(A)}{2^{|A|}-1} e^{\frac{|A|-1}{|X|}}\right) \\ &= -0.4 \times \log_2\left(\tfrac{0.4}{2^2-1} e^{\frac{2-1}{4}}\right) - 0.6 \times \log_2\left(\tfrac{0.6}{2^2-1} e^{\frac{2-1}{3}}\right) \\ E_z(m_1) &= 2.1952 \end{aligned} \tag{11}$$

$$\begin{aligned} E_z(m_2) &= -\sum_{A \subseteq X} m_2(A) \log_2\left(\tfrac{m_2(A)}{2^{|A|}-1} e^{\frac{|A|-1}{|X|}}\right) \\ &= -0.4 \times \log_2\left(\tfrac{0.4}{2^2-1} e^{\frac{2-1}{3}}\right) - 0.6 \times \log_2\left(\tfrac{0.6}{2^2-1} e^{\frac{2-1}{3}}\right) \\ E_z(m_2) &= 2.0750 \end{aligned} \tag{12}$$

As can be seen, the modified Deng entropy gives almost different results for each BOE. In this measure, we can see that the degree of uncertainty calculated from the different $BOEs$ is reduced compared to the Deng measure. However, although the modified Deng entropy takes into account the

TABLE I. SOME UNCERTAINTY MEASURES

Measures	Expression
Hohle's confusion measure [31]	$C_h(m) = -\sum_{A\subseteq X} m(A)\log_2 Bel(A)$
Yager's dissonance measure [33]	$E_y(m) = -\sum_{A\subseteq X} m(A)\log_2 Pl(A)$
Dubois & Prade's weighted Hartley entropy [32]	$E_{dp}(m) = -\sum_{A\subseteq X} m(A)\log_2 \mid A \mid$
Klir & Ramer's discord measure [34]	$D_{kr}(m) = -\sum_{A\subseteq X} m(A)\log_2 \sum_{B\subseteq X} m(B)\frac{\lvert A\cap B\rvert}{\lvert B\rvert}$
Klir & Parviz's strife measure [35]	$S_{kp}(m) = -\sum_{A\subseteq X} m(A)\log_2 \sum_{B\subseteq X} m(B)\frac{\lvert A\cap B\rvert}{\lvert A\rvert}$
George & Pal's conflict measure [36]	$C_{gp}(m) = -\sum_{A\subseteq X} m(A)\log_2 \sum_{B\subseteq X} m(B)[1-\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}]$
Deng entropy [37]	$E_d(m) = -\sum_{A\subseteq X} m(A)\log_2(\frac{m(A)}{2^{\lvert A\rvert}-1})$
Modified Deng entropy [30]	$E_z(m) = -\sum_{A\subseteq X} m(A)\log_2(\frac{m(A)}{2^{\lvert A\rvert}-1}e^{\frac{\lvert A\rvert-1}{\lvert X\rvert}})$

number of elements in the FOD, this measure often gives counter-intuitive measures. We present another problem in example 3.3.

Example 3.3 : Consider the $BOEs$ m_3 and m_4 in the FOD $X = \{a, b\}$ as follows :

$$m_3 : m_3(\{a, b\}) = 1$$
$$m_4 : m_4(\{a\}) = m_4(\{b\}) = 0.5$$

Modified Deng entropy is calculated as follows :

$$E_z(m_3) = -1 \times \log_2(\frac{1}{2^2-1}e^{\frac{2-1}{2}}) = 0.8636$$
$$E_z(m_4) = -0.5\times\log_2(\frac{0.5}{2^1-1}e^{\frac{1-1}{2}})-0.5\times\log_2(\frac{0.5}{2^1-1}e^{\frac{1-1}{2}}) = 1$$

Deng entropy is calculated as follows :

$$E_d(m_3) = -1 \times \log_2(\frac{1}{2^2-1}) = 1.5850$$
$$E_d(m_4) = -0.5 \times \log_2(\frac{0.5}{2^1-1}) - 0.5 \times \log_2(\frac{0.5}{2^1-1}) = 1$$

As can be seen, example 3.3 defines two $BOEs$ m_3 and m_4 where BOE m_3 represents a case of total uncertainty. Intuitively, the uncertainty level of the BOE m_3 must be bigger than that of the BOE m_4, which is in contradiction with the modified Deng measure (E_z). However, the Deng measure (E_d) better distinguishes total uncertainty with an uncertainty measure of BOE m_3 bigger than that of BOE m_4.

Thus, after analyzing examples 3.2 and 3.3, it can be seen that example 3.2 presents a case of variable FOD with the same number of elements in the focal elements. And example 3.3 presents a case where the FOD does not vary. In contrast, in example 3.3, the measure proposed by Deng has better results compared to the Zhou et al. results. Therefore, how to quantify optimally uncertainty by taking into account the limits observed in the measures proposed by Deng and Zhou et al.?

To solve this problem, we propose a new uncertainty measure by extending the measures proposed by Deng and Zhou et al.

IV. NEW UNCERTAINTY MEASURE

In the Dempster-Shafer framework, the new uncertainty measure (E_{Nm}) we propose is as follows:

$$E_{Nm}(m) = -\sum_{A\subseteq X} m(A)\log_2(\frac{m(A)}{2^{|A|}-1}e^{\frac{|A|-1}{2^{|X|}}}) \tag{13}$$

where m is the mass function defined on X. A is the focal element of X and $\mid A \mid$ represents the cardinality of A.The particularity of this measure is that it takes into account the number of elements of the power set represented by $2^{|X|}$.

A simple transformation of the new entropy is as follows:

$$E_{Nm}(m) = \sum_{A\subseteq X} m(A)\log_2(2^{|A|}-1) - \sum_{A\subseteq X} m(A)\log_2 m(A)$$
$$-\sum_{A\subseteq X} m(A)\log_2 e^{\frac{|A|-1}{2^{|X|}}} \tag{14}$$

As can be seen in this expression, the first two terms refer to Deng entropy [37]. These are respectively the measure of the total non-specificity in the mass function m, and the measure of the discord of the mass function between focal elements. The third term, the exponential factor, $e^{\frac{|A|-1}{2^{|X|}}}$, is the main factor in this contribution. The choice of this factor is based on the exponential factor (i.e. $e^{\frac{|A|-1}{|X|}}$) proposed by Zhou et al. [30], which represents the measure of uncertain information in a BOE. Compared to the Zhou et al. measure, the new measure takes into account the number of elements in the power set represented by $2^{|X|}$. Thus, the new measure is intended to be more generic in resolving the limitations of the measures proposed by Deng and Zhou et al. in examples 3.2 and 3.3 respectively. Let's go back to example 3.2, the new entropy is as follows:

$$\begin{aligned} E_{Nm}(m_1) &= -\sum_{A\subseteq X} m_1(A)\log_2\left(\frac{m_1(A)}{2^{|A|}-1}e^{\frac{|A|-1}{2^{|X|}}}\right) \\ &= -0.4\times\log_2\left(\frac{0.4}{2^2-1}e^{\frac{2-1}{2^4}}\right) - 0.6\times\log_2\left(\frac{0.6}{2^2-1}e^{\frac{2-1}{2^3}}\right) \\ E_{Nm}(m_1) &= 2.4657 \end{aligned} \quad (15)$$

$$\begin{aligned} E_{Nm}(m_2) &= -\sum_{A\subseteq X} m_2(A)\log_2\left(\frac{m_2(A)}{2^{|A|}-1}e^{\frac{|A|-1}{2^{|X|}}}\right) \\ &= -0.4\times\log_2\left(\frac{0.4}{2^2-1}e^{\frac{2-1}{2^3}}\right) - 0.6\times\log_2\left(\frac{0.6}{2^2-1}e^{\frac{2-1}{2^3}}\right) \\ E_{Nm}(m_2) &= 2.3755 \end{aligned} \quad (16)$$

The results of the different uncertainty measures of the two $(BOEs)$ m_1 and m_2 are summarized in the table II.

TABLE II. UNCERTAINTY MEASURES CALCULATED IN EXAMPLE 3.2

BOEs	Deng entropy	Modified Deng entropy	New entropy
m_1	2.5559	2.1952	2.4657
m_2	2.5559	2.0750	2.3755

In this table, like Zhou et al.'s proposed measure, the new measure gives different measures for each of $BOEs$ m_1 and m_2. However, the new measure gives bigger values compared to those of Zhou et al. Table III takes the measures from table II, provides the distance $d(m_1, m_2)$ between the measures of $BOEs$ m_1 and m_2 to determine the observed information loss between these $BOEs$. Finally, the BOE $m1$ measure is added to the calculated distance. In this case, the new measure compared to the Zhou et al. measure, takes into account the loss of perceived information in the Deng measure.

TABLE III. EFFECTIVENESS OF THE NEW MEASURE

BOEs	Deng entropy [37]	Modified Deng entropy [30]	New entropy
m_1	2.5559	2.1952	2.4657
m_2	2.5559	2.0750	2.3755
$d(m_1, m_2)$	0	0.1202	0.0902
$m_1 + d(m_1, m_2)$	2.5559	2.3154	2.5559

Thus, using the example 3.3, the new measure of the different $BOEs$ m_3 and m_4 is as follows :

$$E_{Nm}(m_3) = -1\times\log_2\left(\frac{1}{2^2-1}e^{\frac{2-1}{2^2}}\right) = 1.2243$$
$$E_{Nm}(m_4) = -0.5\times\log_2\left(\frac{0.5}{2^1-1}e^{\frac{1-1}{2^2}}\right) - 0.5\times\log_2\left(\frac{0.5}{2^1-1}e^{\frac{1-1}{2^2}}\right)$$
$$E_{Nm}(m_4) = 1$$

The results of the $(BOEs)$ m_3 and m_4 are summarized in the table IV. In this table, the new entropy (E_{Nm}) is also represented. And as can be seen, only the entropy proposed by Zhou et al. gives counter-intuitive results. In this case of example where the FOD does not vary, the new measure is close to the merits of the Deng measure.

TABLE IV. UNCERTAINTY MEASURES CALCULATED IN EXAMPLE 3.3

BOEs	Deng entropy [37]	Modified Deng Entropy [30]	New Entropy
m_3	1.5850	0.8636	1.2243
m_4	1	1	1

Thus, the proposed new measure responds to the limitations of Deng and Zhou et al. by taking a more generic character and an efficient quantification of uncertainty.

V. PROOF AND DISCUSSIONS

In this section, we first present some fundamental properties of the new uncertainty measure. Then, using numerical examples, we show the concordance of the new entropy with some basic entropy including Shannon's entropy (E_s), Deng entropy (E_d) and the entropy proposed by Zhou et al. (E_z). Finally, we discuss the superiority of the new entropy compared to the above-mentioned entropies.

A. Concordance with Shanon entropy

The proposed entropy measure is identical to the basic entropy, the Shannon entropy (Eq. 17), when we have a Bayesian mass function (i.e. a single element in the BOE or $\mid A \mid\equiv 1$) as follows.

$$\begin{aligned} E_{Nm}(m) &= -\sum_{A\subseteq X} m(A)\log_2\left(\frac{m(A)}{2^{|A|}-1}e^{\frac{|A|-1}{2^{|X|}}}\right) \\ &= -\sum_{A\subseteq X} m(A)\log_2\left(\frac{m(A)}{2^1-1}e^{\frac{1-1}{2^{|X|}}}\right) \\ &= -\sum_{A\subseteq X} m(A)\log_2 m(A) \\ E_{Nm}(m) &= E_s(m) \end{aligned} \quad (17)$$

In addition, another proof of concordance of the new entropy with different uncertainty measures is provided in the case where the (FOD) has only one element (i.e. total uncertainty case) as shown in the following example.

Example 5.1 : Consider an information processing system in which information I reported by a sensor has a belief equal to one hundred percent. In $X = \{I\}$, the mass function can be noted as $m(\{I\}) = 1$. The calculation of the entropies of Shannon (E_s), Deng (E_d), Zhou et al (E_z) and the new entropy measure (E_{Nm}) are defined as follows:

$$E_s(m) = -1\times\log_2 1 = 0$$
$$E_d(m) = -1\times\log_2\frac{1}{2^1-1} = 0$$
$$E_z(m) = -1\times\log_2\left(\frac{1}{2^1-1}e^{\frac{1-1}{1}}\right) = 0$$
$$E_{Nm}(m) = -1\times\log_2\left(\frac{1}{2^1-1}2^{\frac{1-1}{2^1}}\right) = 0$$

B. Superiority of the new uncertainty measure

To show the superiority of the new uncertainty measure, recall the example mentioned in [37].

Exemple 5.3 : Consider a mass function m in a FOD X such that : $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$, $m(\{6\}) = 0.05$, $m(\{3, 4, 5\}) = 0.05$, $m(E) = 0.8$ et $m(X) = 0.1$. E represents a subset of elements varying from 1 to 14 as shown in the table V.

TABLE V. ENTROPY MEASURES WITH A SUBSET E OF VARIABLE ELEMENTS

Elements in E	Deng Entropy [37]	Modified Deng entropy [30]	New Entropy
$E = \{1\}$	2.6623	2.5180	2.6622
$E = \{1, 2\}$	3.9303	3.7090	3.9302
$E = \{1, 2, 3\}$	4.9082	4.6100	4.9080
$E = \{1, ..., 4\}$	5.7878	5.4127	5.7876
$E = \{1, ..., 5\}$	6.6256	6.1736	6.6254
$E = \{1, ..., 6\}$	7.4441	6.9151	7.4439
$E = \{1, ..., 7\}$	8.2532	7.6473	8.2530
$E = \{1, ..., 8\}$	9.0578	8.3749	9.0575
$E = \{1, ..., 9\}$	9.8600	9.1002	9.8597
$E = \{1, ..., 10\}$	10.6612	9.8244	10.6608
$E = \{1, ..., 11\}$	11.4617	10.5480	11.4613
$E = \{1, ..., 12\}$	12.2620	11.2714	12.2616
$E = \{1, ..., 13\}$	13.0622	11.9946	13.0617
$E = \{1, ..., 14\}$	13.8622	12.7177	13.8617

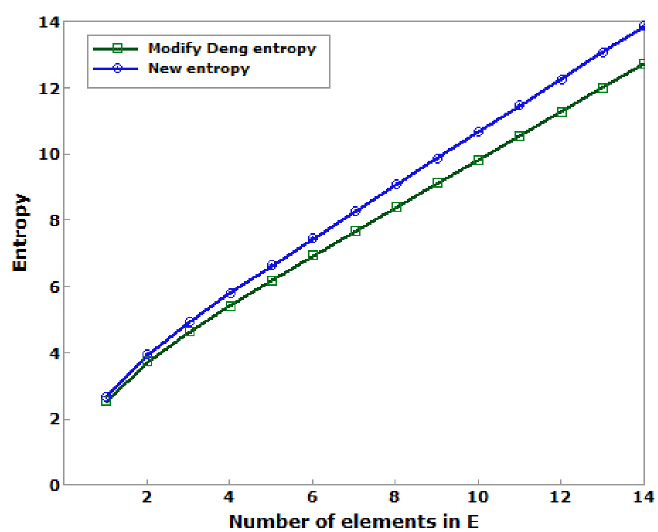


Fig. 1. Comparison between the entropy proposed by Zhou et al. and the new entropy

As can be seen in Figure 1, the new and modified entropy of Deng proposed by Zhou et al. increase monotonously with increasing size of the subset E. However, the values of the new entropy are significantly bigger than that of Zhou et al. measure. As shown in example 3.3. the measure proposed by Zhou et al. records losses of information especially in such a case where FOD that does not vary. Moreover, the new entropy gives results almost identical to the Deng measurement (figure 2), hence the effectiveness of the approach when the FOD does not change.

Moreover, the new entropy gives results almost identical to the Deng measure (figure 2), hence the effectiveness of the new approach when the FOD does not change. The new measure does not differ from the merits of the Deng measure.

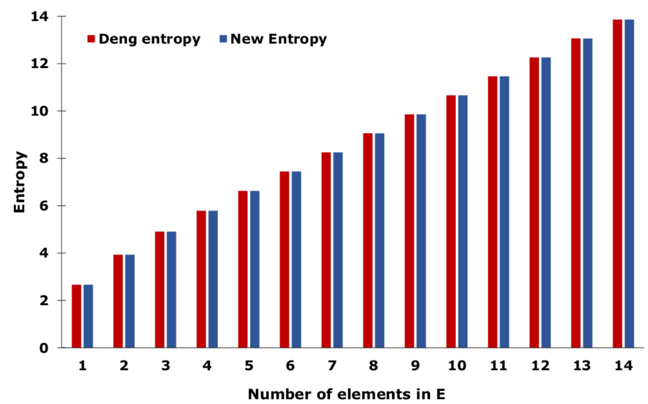


Fig. 2. Comparison between Deng entropy and new entropy

Figure 3 shows the degree of uncertainty of the BOE using other different uncertainty measures including Hohle's confusion measure [31], Dubois & Prade's weighted Hartley entropy [32], Yager's dissonance measure [33], Klir & Ramer's discord measure [34], Klir & Parviz's strife measure [35], George & Pal's conflict measure [36]. In this figure, we can observe that only the entropies of Dubois & Prade [32], Deng [37], Zhou [30], and the new entropy increase monotonously with the increase in the size of subset E. Also with the increase in size of E, there is either a declination or a change in the pace of other uncertainty measures. Hence the effectiveness of the new measure.

VI. CONCLUSION

Quantifying uncertainty in information systems is very important for evaluating the quality of information. Several methods based on entropy of beliefs have been proposed in the literature, but these give counter-intuitive results, particularly in the case of variable FOD with BOE. In this paper, we have proposed a new measure to address these deficiencies. This measure extends the measures proposed by Deng and Zhou et al. From numerical examples and mathematical properties, we have shown the effectiveness of the new measure which gives more information in the power set of the FOD. Our future studies will focus on the actual application of the news in several areas including decision making, fault diagnosis and detection, and so on.

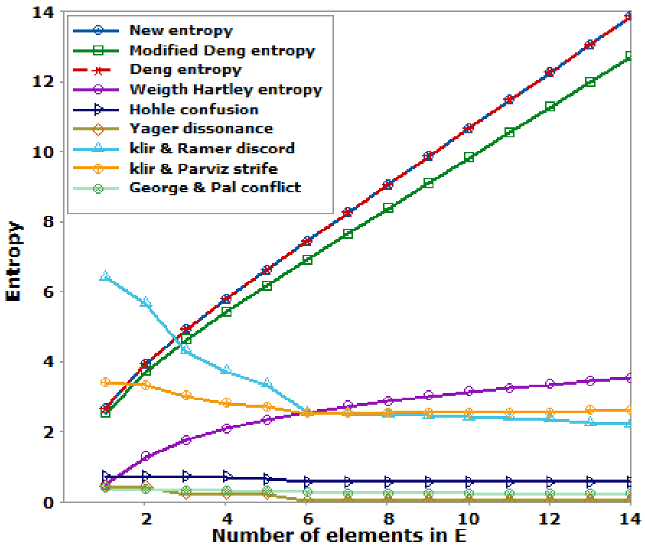


Fig. 3. Comparison with other uncertainty mesures

REFERENCES

[1] Y. Deng, "D numbers: theory and applications," *Journal of Information &Computational Science*, vol. 9, no. 9, pp. 2421–2428, 2012.

[2] G. J. Klir, "Generalized information theory," *Fuzzy Sets and Systems*, vol. 40, no. 1, pp. 127–142, Mar. 1991.

[3] L. A. Zadeh, "Toward a generalized theory of uncertainty (GTU)an outline," *Information Sciences*, vol. 172, no. 1, pp. 1–40, 2005.

[4] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of big data on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98–115, Jan. 2015.

[5] C. E. Shannon, "A note on the concept of entropy," *Bell System Tech. J*, vol. 27, no. 3, pp. 379–423, 1948.

[6] W. Feller, *An introduction to probability theory and its applications*. John Wiley & Sons, 2008, vol. 2.

[7] J. M. Steele, "Probability theory: Formal," 2015.

[8] D. Dubois and H. Prade, "Possibility Theory," in *Computational Complexity: Theory, Techniques, and Applications*, R. A. Meyers, Ed. New York, NY: Springer New York, 2012, pp. 2240–2252.

[9] L. N. Kanal and J. F. Lemmer, *Uncertainty in artificial intelligence*, 2014, vol. 4.

[10] A. P. Dempster, "Upper and Lower Probabilities Induced by a Multivalued Mapping," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967.

[11] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[12] D. Harmanec and G. J. Klir, "Measuring Total Uncertainty in Dempster-Shafer Theory: A Novel Approach," *International Journal of General Systems*, vol. 22, no. 4, pp. 405–419, 1994.

[13] R. D. Shachter, L. Kanal, M. Henrion, and J. F. Lemmer, *Uncertainty in Artificial Intelligence 5*, 2017, vol. 10.

[14] H. B. Sta, "Quality and the efficiency of data in Smart-Cities," *Future Generation Computer Systems*, vol. 74, pp. 409–416, 2017.

[15] A. Gaur, B. Scotney, G. Parr, and S. McClean, "Smart city architecture and its applications based on iot," *Procedia computer science*, vol. 52, pp. 1089–1094, 2015.

[16] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang, "Sensor fusion using Dempster-Shafer theory [for context-aware HCI]," in *Instrumentation and Measurement Technology Conference, 2002. IMTC/2002. Proceedings of the 19th IEEE*, vol. 1. IEEE, 2002, pp. 7–12.

[17] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[18] M. Bauer, "Approximation algorithms and decision making in the Dempster-Shafer theory of evidenceAn empirical study," *International Journal of Approximate Reasoning*, vol. 17, no. 2-3, pp. 217–237, 1997.

[19] D. Yong, S. WenKang, Z. ZhenFu, and L. Qi, "Combining belief functions based on distance of evidence," *Decision support systems*, vol. 38, no. 3, pp. 489–493, 2004.

[20] O. Basir and X. Yuan, "Engine fault diagnosis based on multi-sensor information fusion using DempsterShafer evidence theory," *Information Fusion*, vol. 8, no. 4, pp. 379–386, 2007.

[21] A. Paksoy and M. Göktürk, "Information fusion with dempster-shafer evidence theory for software defect prediction," *Procedia Computer Science*, vol. 3, pp. 600–605, 2011.

[22] D. M. Buede and P. Girardi, "A target identification comparison of Bayesian and Dempster-Shafer multisensor fusion," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 27, no. 5, pp. 569–577, 1997.

[23] M. Beynon, B. Curry, and P. Morgan, "The dempster–shafer theory of evidence: an alternative approach to multicriteria decision modelling," *Omega*, vol. 28, no. 1, pp. 37–50, 2000.

[24] E.-P. Lim, J. Srivastava, and S. Shekhar, "An evidential reasoning approach to attribute value conflict resolution in database integration," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 5, pp. 707–723, 1996.

[25] J. Yang, H.-Z. Huang, Q. Miao, and R. Sun, "A novel information fusion method based on Dempster-Shafer evidence theory for conflict resolution," *Intelligent Data Analysis*, vol. 15, no. 3, pp. 399–411, 2011.

[26] K. Yamada, "A new combination of evidence based on compromise," *Fuzzy sets and Systems*, vol. 159, no. 13, pp. 1689–1708, 2008.

[27] L. A. Zadeh, "A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination," *AI magazine*, vol. 7, no. 2, p. 85, 1986.

[28] P. Xu, X. Su, S. Mahadevan, C. Li, and Y. Deng, "A non-parametric method to determine basic probability assignment for classification problems," *Applied intelligence*, vol. 41, no. 3, pp. 681–693, 2014.

[29] W. Jiang, J. Zhan, D. Zhou, and X. Li, "A method to determine generalized basic probability assignment in the open world," *Mathematical Problems in Engineering*, vol. 2016, 2016.

[30] D. Zhou, Y. Tang, and W. Jiang, "A modified belief entropy in Dempster-Shafer framework," *PLOS ONE*, vol. 12, no. 5, p. e0176832, 2017.

[31] U. HOHLE, "Entropy with respect to Plausibility Measures," *Proc. of 12th IEEE Int. Symp. on Multiple Valued Logic, Paris, 1982*, 1982.

[32] D. Dubois and H. Prade, "Properties of measures of information in evidence and possibility theories," *Fuzzy Sets and Systems*, vol. 24, no. 2, pp. 161–182, 1987.

[33] R. R. Yager, "Entropy and Specificity in a Mathematical Theory of Evidence," *International Journal of General Systems*, vol. 9, no. 4, pp. 249–260, 1983.

[34] G. J. Klir and A. Ramer, "Uncertainty in the Dempster-Shafer Theory: A Critical Re-examination," *International Journal of General Systems*, vol. 18, no. 2, pp. 155–166, 1990.

[35] G. J. Klir and B. Parviz, "A note on the Measure of Discord," in *Uncertainty in Artificial Intelligence*, D. Dubois, M. P. Wellman, B. D'Ambrosio, and P. Smets, Eds. Morgan Kaufmann, Jan. 1992, pp. 138–141.

[36] T. George and N. R. Pal, "Quantification of Conflict in Dempster-Shafer Framework: A new approach," *International Journal of General Systems*, vol. 24, no. 4, pp. 407–423, 1996.

[37] Y. Deng, "Deng entropy," *Chaos, Solitons & Fractals*, vol. 91, pp. 549–553, Oct. 2016.

[38] J. Wang, F. Xiao, X. Deng, L. Fei, and Y. Deng, "Weighted Evidence Combination Based on Distance of Evidence and Entropy Function," *International Journal of Distributed Sensor Networks*, vol. 12, no. 7, p. 3218784, Jul. 2016.

[39] P. P. Shenoy and G. Shafer, "Axioms for Probability and Belief-Function Propagation," in *Classic Works of the Dempster-Shafer Theory of Belief Functions*, ser. Studies in Fuzziness and Soft Computing, R. R. Yager and L. Liu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 499–528.

[40] T. Denoeux, “Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence,” *Artificial Intelligence*, vol. 172, pp. 234–264, 2008.

A Secure User Authentication Scheme with Biometrics for IoT Medical Environments

YoHan Park
Division of IT Convergence
Korea Nazarene University
Korea, Republic

***Abstract*—Internet of Things (IoT) is a ubiquitous network that devices are interconnected and users can access those devices through the Internet. Recently, medical healthcare systems are combined with these IoT networks and provide efficient and effective medical services to medical staff and patients. However, the security threats are increased simultaneously as the requirements of medical services in IoT medical environments are increased. It is essential to provide security of the networks from malicious attacks.**

In 2018, Roy et al. proposed a remote user authentication and key agreement scheme with biometrics in IoT medical environments. Unfortunately, we analyze Roy et al.'s scheme and demonstrate that their scheme does not withstand various attacks, such as replay attacks and password guessing attacks. Then we propose a user authentication scheme to overcome these security drawbacks. The proposed scheme withstands various attacks from adversaries in IoT medical environments and provide better security functionalities of those of Roy et al. We then prove the authentication and session key of the proposed scheme using *BAN* logic and analyze that our proposed scheme is secure against various attacks.

Keywords*—*IoT medical environments; Cryptanalysis; User authentication; BAN logic

I. Introduction

With the rapid development of mobile devices and wireless networks, users can access various services conveniently at any time and anywhere [1], [2]. These changes affect the healthcare environment, enabling medical devices to communicate with each other and communicate that information to the users. Those devices are also interconnected with medical servers and medical staff [3]. The changes that those developments have brought on to the daily lives of human beings are enormous. The spread of IoT medical technology enables people to utilize advanced medical services such as e-healthcare [4], [5]. The telecare medical information system (TMIS) is one of the advanced information medical system [6], [7]. Medical staff can treat patients and diagnose a case of them in the distance with the aid of medical devices and store the information of patients to a medical server. Remote monitoring can be possible efficiently with IoT connected medical devices. Sensors attached to patients can capture health data and share it through wireless connection with medical staff. The IoT technology in medical environments makes the healthcare system easy to be managed and gives a lot of possibilities of medical services.

However, the IoT environment has an enormous threat to security and privacy due to its heterogeneous and dynamic nature [8]. To make the IoT-based medical system widely accepted, security problems should be resolved in advance. Especially, user authentication is an essential prerequisite among all the security concerns to provide integrity, access control, and availability for IoT environments [9]–[11]. Without secure authentication methods, the external party can directly access user's information which are more valuable and even critical than general information. Hence, it is necessary to provide an authentication process between a user and service providers before permitting a user to access the services.

There are many authentication schemes to provide security of users medical information. To provide user security against inside attackers, Chen *et al*. proposed a dynamic ID-based authentication scheme for TMIS [12]. However, [12] was vulnerable to guessing attacks and tracking attacks. Jiang *et al*. [3] demonstrated that and [12]' scheme leaked out personal information. Then Jiang *et al*. proposed an authentication scheme which can withstand anonymity and untraceability of users. But There scheme was attacked by Kumari *et al*. [6]. They said [3] was vulnerable to password guessing attack, user impersonation attack, and so on. [7] also showed the security flawless of [3]. Many authentication schemes try to provide patients to utilize medical services securely.

Roy et al. [13] proposed a three factor remote user authentication scheme in IoT medical environments. They insisted that their scheme is resist to various attacks. Unfortunately, this paper demonstrates Roy et al.'s scheme fails to provide security against a number of attacks, such as replay attacks and offline password guessing attacks. And we show that their scheme does not provide perfect forward secrecy. Subsequently, we propose a secure three factor remote user authentication scheme to solve these security vulnerabilities.

A. Threat model

The Dolev-Yao threat (DY) model [14] is widely used in evaluating the security of a protocol [15]. Under the DY model, we assume that the capabilities of adversaries $\mathcal{A}$ are as follows.

- $\mathcal{A}$ has total controlled over the communication channel connecting the users and the remote server in login/authentication phase. Thus the adversary can intercept, insert, delete, or modify any message transmitted via a public channel.
- $\mathcal{A}$ can have a lost or stolen smart card, and extract the information stored in a smart card by means of analyzing the power consumption of the smart card [16], [17].
- $\mathcal{A}$ can perform various attacks including offline password guessing attack, replay attack, and man-in-the-middle

attack. Especially, $\mathcal{A}$ can guess identity and password simultaneously [18].

B. Contributions

The contributions made in the paper are listed below:

1) We analyze security weaknesses of Roy et al.'s scheme [13] and demonstrate that it is vulnerable to replay attack, offline password guessing attack. In addition, we show that their scheme does not provide perfect forward secrecy.
2) To overcome these security weaknesses, we propose an enhanced secure authentication scheme in IoT medical network. The proposed scheme prevents various attacks such as password guessing attack, user impersonation attack and replay attack from malicious adversaries.
3) Our scheme provides secure mutual authentication and perfect forward secrecy, and we prove the secure mutual authentication of our scheme using the BAN logic.

C. Paper Structure

The rest of the paper is organized as follows. In Section 2, we review Roy et al.'s scheme followed by the cryptanalysis of Roy et al.'s scheme in Secion 3. In Section 4, we propose a secure remote user authentication scheme in IoT medical networks to withstand the security pitfalls found in the authentication scheme of Roy et al.'s scheme, and then security and efficiency of the proposed scheme are analyzed with related existing schemes in Section 5. Finally, Section 6 concludes the paper.

II. Review of Roy et al.'s scheme

In this section, we review Roy et al.'s remote user authentication scheme. It is composed of four phases: registration, login, authentication and key establishment, and password change. Table I describe the notations used throughout the paper.

TABLE I. Notations

Notation	Meaning
U_i	user i
ID_i	identity of U_i
PW_i	password of U_i
B_i	biometric template of U_i
SC_i	smart card of user U_i
S_j	medical server j
$Gen(\cdot)$	Generation function
$Rep(\cdot)$	Reproduction function
$E_k(\cdot)/D_k(\cdot)$	encryption/decryption using key k
X_S	master secret key of S_j
TS_i	timestamp
$\|\|$	concatenate operation
$\bigoplus$	XOR operation
$h(\cdot)$	hash function

A. Registration phase

If a new user U_i wants to access the medical service, U_i must register with the remote server S_j first. The Roy et al.'s user registration phase is illustrated in Figure 1, and the detailed steps of this registration phase are as follows:

1) U_i chooses ID_i and PW_i, and imprints a biometric template B_i. U_i generates parameters $< \alpha_i, \beta_i > \leftarrow Gen(B_i)$.
2) U_i generates a random number θ_i and compute $TID_i = h(h(ID_i) \oplus h(\theta_i))$ and $RPB_i = h(ID_i||\alpha_i||h(PW_i))$. Then U_i sends TID_i to S_j via a secure channel.
3) S_j computes $\delta_i = h(h(TID_i)||h(X_S))$. S_j sends a smart card SC_i and δ_i to U_i via a secure channel.
4) U_i computes the parameters Y_1, Y_2 and Y_3 as follows:

$$\begin{aligned} Y_1 &= h(\delta_i) \oplus h(RPB_i||\theta_i) \\ Y_2 &= h(h(ID_i)||PW_i) \oplus \theta_i \\ Y_3 &= h(\alpha_i||PW_i||\theta_i) \end{aligned}$$

5) Finally, U_i store the parameters $< Y_1, Y_2, Y_3, h(\cdot), \beta_i >$ in a smart card SC_i.

B. Login phase

When the authenticated user U_i wants to use a medical service, U_i sends request messages of accessing the medical service to the remote server S_j. Roy et al.'s scheme also supposed that U_i and S_j must authenticate each other before sending the request message. The Roy et al.'s login phase is illustrated as follows:

1) U_i chooses ID_i and PW_i, and imprints biometrics B_i'. Then SC_i generates α_i, and computes θ_i' and Y_3', and then compares Y_3' with Y_3 to check a user credential as follow:

$$\begin{aligned} \alpha_i &\leftarrow Rep(B_i', \beta_i) \\ \theta_i' &= Y_2 \oplus h(h(ID_i)||PW_i) \\ Y_3' &= h(\alpha_i||PW_i||\theta_i') \\ \text{verifies } Y_3' &\stackrel{?}{=} Y_3 \end{aligned}$$

2) SC_i generates two random number RN_i and θ_i^* and computes $RPB_i, \mu_i(= h(\delta_i)), TID_i, D_1$ and H_1 as follows:

$$\begin{aligned} RPB_i &= h(ID_i||\alpha_i||h(PW_i)) \\ \mu_i &= Y_1 \oplus h(RPB_i||\theta_i) \\ TID_i &= h(h(ID_i) \oplus h(\theta_i)) \\ D_1 &= E_{\mu_i}(ID_i||\theta_i||\theta_i^*||TS_i||RN_i) \\ H_1 &= h(h(ID_i \oplus \theta_i)||\theta_i^*||TID_i||TS_i||RN_i) \end{aligned}$$

Then, U_i sends a request message $Msg_1 =< TID_i, D_1, H_1 >$ to S_j via a public channel.

C. Authentication and key establishment phase

U_i and S_j authenticate and generate a session key. Figure 2 illustrates the authentication and key establishment phase, which performs as follows:

1) S_j computes δ_i, μ_i', and decrypts D_1 and obtain $(ID_i, \theta_i, \theta_i^*, TS_i, RN_i)$ as follow:

$$\begin{aligned} \delta_i &= h(h(TID_i)||h(X_S)) \\ \mu_i &= h(\delta_i) \\ D_{\mu_i}(D_1) &= \{ID_i||\theta_i||\theta_i^*||TS_i||RN_i\} \end{aligned}$$

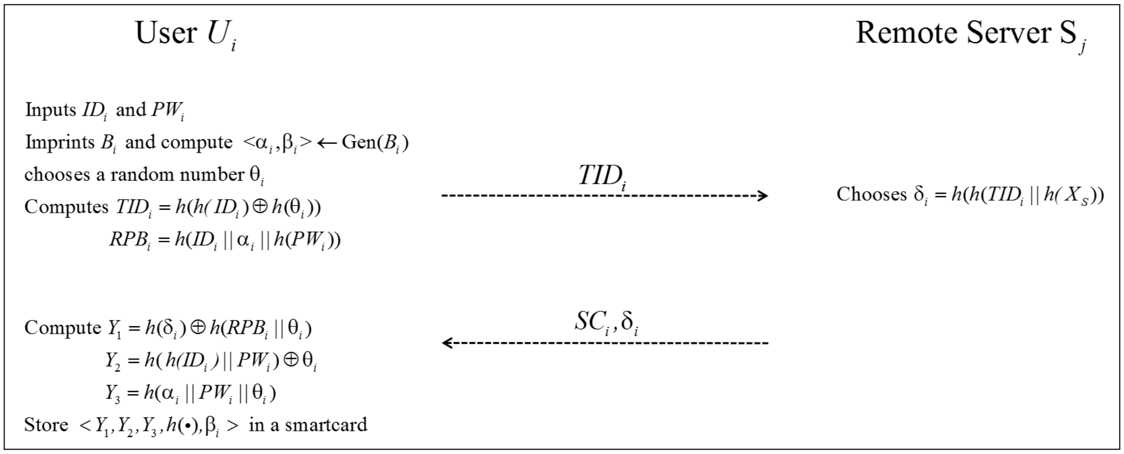


Fig. 1. User registration phase of Roy et al.'s scheme

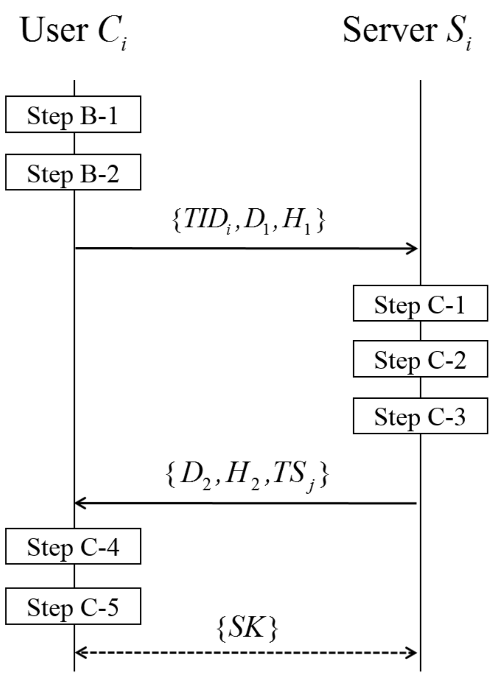


Fig. 2. Authentication and key establishment of Roy et al.'s scheme

2) S_j retrieves TS_i^* and checks $TS_i^* - TS_1 \leq \Delta T$. If it is true, S_j checks the validity of H_1 and TID_i using the decrypted parameters of D_1 as follows:

$$
\begin{aligned}
H_1 &\stackrel{?}{=} h(h(ID_i) \oplus \theta_i)||\theta_i^*||TID_i||TS_i||RN_i) \\
TID_i &\stackrel{?}{=} h(h(ID_i) \oplus h(\theta_i))
\end{aligned}
$$

If both verifications are successful, proceed to the next step.

3) S_j computes $TID_i, \delta_i^1, \lambda_i, D_2, SK_{S_j,U_i}$ and H_2 as follows:

$$
\begin{aligned}
TID_i^1 &= h(h(ID_i) \oplus h(\theta_i) \oplus h(\theta_i^*)) \\
\delta_i^1 &= h(h(TID_i^1)||h(X_S)) \\
\lambda_i &= h(TID_i^1) \\
D_2 &= E_{\lambda_i}(\delta_i^1||\theta||RN_j) \\
SK_{S_j,U_i} &= h(\delta_i^1||\theta_i^*||RN_i||RN_j||TS_i||TS_j) \\
H_2 &= h(TID_i^1||\delta_i^1||SK_{S_j,U_i}||TS_j||RN_j) \\
&\quad \text{(where, } TS_j \text{ is the time stamp.)}
\end{aligned}
$$

Then, S_j sends the replay message $Msg_2 =< D_2, H_2, TS_j >$ to U_i via a public channel.

4) U_i retrieves TS_j^* and checks $TS_j^* - TS_j \leq \Delta T$. If it is true, U_i computes TID_i^1, δ_i and decrypts D_2 as follows:

$$
\begin{aligned}
TID_i^1 &= h(h(ID_i) \oplus h(\theta_i) \oplus h(\theta_i^*)) \\
\lambda_i &= h(TID_i^1) \\
D_{\lambda_i}(D_2) &= \{\delta_i^1||\theta_i||RN_j\}
\end{aligned}
$$

5) Finally, U_i generates a session key $SK_{U_i,S_j} = h(\delta_i^1||\theta_i^*||RN_i||RN_j||TS_i||TS_j)$. Then U_i checks the validity of H_2 by comparing it with the computed value $h(TID_i^1||\delta_i^1||SK_{U_i,S_j}||TS_j||RN_j)$. If it is true, U_i accepts SK_{U_i,S_j} as the current session key, then updates new parameters Y_1^*, Y_2^* and Y_3^* as follows:

$$
\begin{aligned}
Y_1^* &= h(\delta_i) \oplus h(RPB_i||\theta_i^*) \\
Y_2^* &= h(h(ID_i)||PW_i) \oplus \theta_i^* \\
Y_3^* &= h(\alpha_i||PW_i||\theta_i^*)
\end{aligned}
$$

D. Password change phase

To provide a password change requirement, U_i performs following steps.

1) U_i inserts SC_i and inputs ID_i, PW_i and B_i.
2) SC_i computes θ_i from Y_2 and Y_3 using θ_i as given in step 2 of login phase.
3) If it is correct, U_i input a new password PW_i^{new} and compute new parameters Y_1^{new}, Y_2^{new} and Y_3^{new} as follows:

$$\begin{aligned} RPB_i^{new} &= h(ID_i||\alpha_i||h(PW_i^{new})) \\ Y_1^* &= Y_1 \oplus h(RPB_i||\theta_i) \oplus h(RPB_i^{new}||\theta_i) \\ Y_2^* &= h(h(ID_i)||PW_i^{new}) \oplus \theta_i \\ Y_3^* &= h(\alpha_i||PW_i^{new}||\theta_i) \end{aligned}$$

III. Cryptanalysis of Roy et al.'s Scheme

In this section, we demonstrate that Roy et al.'s scheme cannot prevent replay attacks and offline password guessing attacks. We also show that their scheme cannot provide perfect forward secrecy, and an adversary can trace users freely. We assumed that an adversary $\mathcal{A}$ could steal or obtain the user's smart card SC_i. In addition, an adversary $\mathcal{A}$ could extract information $\{Y_1, Y_2, Y_3\}$ from the smart card and could get previous session messages transmitted through public network. The description of the security weaknesses of Roy et al.'s scheme is as follows.

A. Reply attack

If the adversary $\mathcal{A}$ obtains the transmitted parameter TID_i, $\mathcal{A}$ can attempt to reuse it as it's registration message, and then $\mathcal{A}$ can get δ_i used as a secret key between a user and a server. The procedure of replay attack is as follow.

1) $\mathcal{A}$ captures the transmitted parameter TID_i and sends it to S_j.
2) S_j which received TID_i from $\mathcal{A}$ computes $\delta_i = h(H(TID_i)||X_S)$ which is exactly same as that of U_i.
3) S_j sends δ_i to $\mathcal{A}$ via a secure channel.
4) $\mathcal{A}$ computes a session key $\mu_i = h(\delta_i)$ and may use it to decrypt D_1.

The result of this attack indicates that Roy et al.'s scheme is vulnerable to replay attack.

B. Offline password guessing attack

If the adversary $\mathcal{A}$ somehow steals SC_i of U_i, $\mathcal{A}$ can attempt to guess the password of U_i, and then $\mathcal{A}$ can guess identity and password of U_i successfully. The procedure of offline password guessing attack is as follows:

1) From the password dictionary space D_{PW}, the adversary $\mathcal{A}$ randomly chooses the password PW_i^*, and picks up the identity ID_i^* from the identity dictionary space D_{ID}.
2) $\mathcal{A}$ calculates $\theta_i^* = Y_2 \oplus h(h(ID_i^*)||PW_i^*)$
3) $\mathcal{A}$ calculates $TID_i^* = h(h(ID_i^*) \oplus h(\theta_i^*)$
4) To check the correctness of PW_i^*, $\mathcal{A}$ examines whether $TID_i^* = TID_i$, where TID_i is previous transmitted parameter. If it is correct, $\mathcal{A}$ guesses identity and password of U_i correctly.

Therefore, Roy et al.'s scheme is vulnerable to offline password guessing attack.

C. Lack of perfect forward secrecy

We assume that $\mathcal{A}$ intercepts and store messages transmitted in the previous session, and a session key μ_i is compromised by $\mathcal{A}$. In Roy et al.'s scheme, D_1 is computed as $D_1 = E_{\mu_i}(ID_i||\theta_i||\theta_i^*||TS_i||RN_i)$. Once μ_i is revealed to $\mathcal{A}$, $\mathcal{A}$ can decrypt the previous encrypted messages using μ_i. Therefore, Roy et al.'s scheme does not support perfect forward secrecy.

IV. Proposed Scheme

In this section, we present the secure biometric based remote user authentication scheme for IoT medical networks that overcomes the security weaknesses of Roy et al's scheme. To provide perfect forward secrecy, we refer Reddy et al.'s technique [19]. The proposed scheme consists of four phases as in the Roy's scheme, namely 1) Registration phase, 2) Login phase, 3) Authentication and key establishment phase, and 4) password change. It is worth noticing that the password change phase of the proposed scheme remains same as that of Roy et al.'s scheme.

A. Registration phase

If a new user U_i wants to access the medical service, U_i must register with the remote server S_j first. User registration phase in the proposed scheme is illustrated in Figure 3, and the detailed steps of this registration phase are as follows:

1) U_i chooses ID_i and PW_i, and imprints a biometric template B_i. U_i generates parameters $< \alpha_i, \beta_i > \leftarrow Gen(B_i)$.
2) U_i generates a random number θ_i and compute $TID_i = h(h(ID_i) \oplus h(\theta_i))$ and $RPW_i = h(ID_i||\alpha_i||h(PW_i))$. Then U_i sends TID_i, RPW_i to S_j via a secure channel.
3) S_j chooses two master key X_{S_1} and X_{S_2}. Then S_j computes $\delta_i = h(h(TID_i)||RPW_i||h(X_{S_1})$, $B_i = h(\delta_i), C_i = h(X_{S_2} \oplus \delta_i)$. Then S_j sends a smart card SC_i and δ_i, B_i, C_i to U_i via a secure channel.
4) U_i computes the parameters Y_1, Y_2 and Y_3 as follows:

$$\begin{aligned} Y_1 &= B_i \oplus h(RPW_i||\theta_i) \\ Y_2 &= h(h(ID_i)||PW_i||\alpha_i) \oplus \theta_i \\ Y_3 &= h(\alpha_i||PW_i||\theta_i) \end{aligned}$$

5) Finally, U_i store the parameters $< Y_1, Y_2, Y_3, B_i, C_i, h(\cdot), \beta_i >$ in a smart card SC_i.

B. Login phase

When the authenticated user U_i wants to use a medical service, U_i sends request messages of accessing the medical service to the remote server S_j. The proposed scheme also supposed that U_i and S_j must authenticate each other before sending the request message. Login phase in the proposed scheme is illustrated as follows:

1) U_i chooses ID_i and PW_i, and imprints biometrics B_i'. Then SC_i generates α_i, and computes θ_i' and Y_3', and then compares Y_3' with Y_3 to check a user credential as follow:

$$\begin{aligned} \alpha_i &\leftarrow Rep(B_i', \beta_i) \\ \theta_i' &= Y_2 \oplus h(h(ID_i)||PW_i||\alpha_i) \\ Y_3' &= h(\alpha_i||PW_i||\theta_i') \\ \text{verifies } Y_3' &\stackrel{?}{=} Y_3 \end{aligned}$$

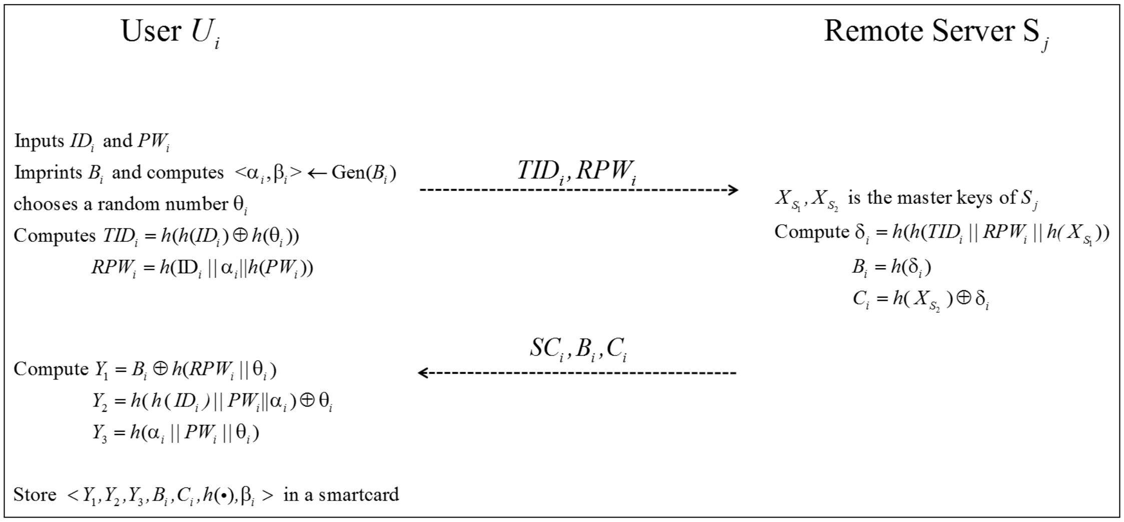


Fig. 3. User registration phase of the proposed scheme

2) SC_i generates two random number r_i and RN_i and computes RPW_i, K_i, PID_i, D_1 and H_1 as follows:

$$\begin{aligned}
TID_i &= h(h(ID_i)\oplus h(\theta_i)) \\
RPW_i &= h(ID_i||\alpha_i||h(PW_i)) \\
B_i &= Y_1 \oplus h(RPW_i||\theta_i) \\
K_i &= h(C_i)\oplus r_i \\
PID_i &= TID_i \oplus r_i \\
M_i &= h(B_i)\oplus r_i \\
D_1 &= E_{K_i}(ID_i||\theta_i||r_i||TS_i||RN_i) \\
H_1 &= h(h(ID_i\oplus\theta_i)||r_i||TID_i||TS_i||RN_i)
\end{aligned}$$

Then, U_i sends a request message $Msg_1 =< PID_i, C_i, M_i, D_1, H_1 >$ to S_j via a public channel.

C. Authentication and key establishment phase

U_i and S_j authenticate and generate a session key. Figure 4 illustrates the authentication and key establishment phase, which performs as follows:

1) S_j computes δ_i, r_i, K_i, and decrypts D_1 using K_i and obtain $(ID_i, \theta_i, r_i, TS_i, RN_i)$ as follow:

$$\begin{aligned}
\delta_i &= C_i \oplus h(X_{S_2}) \\
B_i &= h(X_{S_1})\oplus \delta_i \\
r_i &= M_i \oplus h(B_i) \\
K_i &= r_i \oplus h(C_i) \\
TID_i &= PID_i \oplus r_i \\
D_{K_i}(D_1) &= \{ID_i||\theta_i||r_i||TS_i||RN_i\}
\end{aligned}$$

2) S_j retrieves TS_i^* and checks $TS_i^* - TS_1 \le \Delta T$. If it is true, S_j checks the validity of H_1 and TID_i using the decrypted parameters of D_1 as follows:

$$\begin{aligned}
H_1 &\stackrel{?}{=} h(h(ID_i)\oplus\theta_i)||r_i||TID_i||TS_i||RN_i) \\
TID_i &\stackrel{?}{=} h(h(ID_i)\oplus h(\theta_i))
\end{aligned}$$

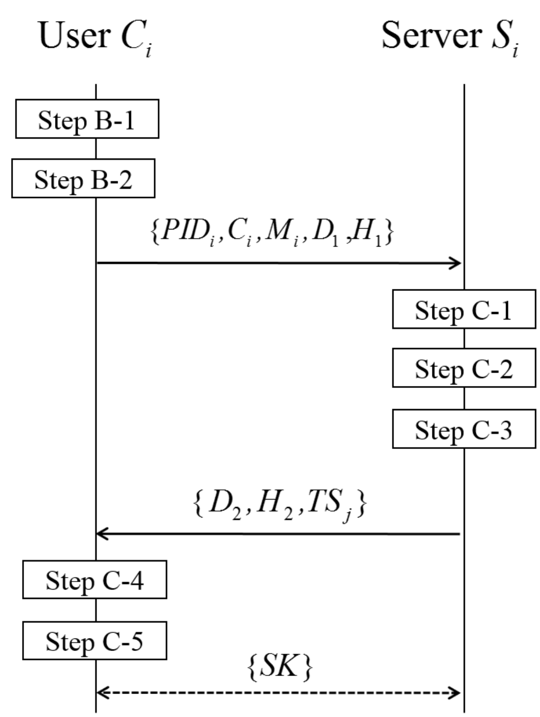


Fig. 4. Authentication and key establishment of the proposed scheme

If both verifications are successful, proceed to the next step.

3) S_j computes $TID_i, \delta_i^1, \lambda_i, D_2, SK_{S_j,U_i}$ and H_2 as follows:

$$\begin{aligned}
TID_i^1 &= h(h(ID_i) \oplus h(\theta_i) \oplus h(r_i)) \\
\delta_i^1 &= h(h(TID_i^1)||h(X_{S_1})) \\
\lambda_i &= h(TID_i^1) \\
D_2 &= E_{\lambda_i}(\delta_i^1||\theta_i||RN_j) \\
SK_{S_j,U_i} &= h(\delta_i^1||r_i||RN_i||RN_j||TS_i||TS_j) \\
H_2 &= h(TID_i^1||\delta_i^1||SK_{S_j,U_i}||TS_j||RN_j) \\
& \quad \text{(where, } TS_j \text{ is the time stamp.)}
\end{aligned}$$

Then, S_j sends the replay message $Msg_2 =< D_2, H_2, TS_j >$ to U_i via a public channel.

4) U_i retrieves TS_j^* and checks $TS_j^* - TS_j \leq \Delta T$. If it is true, U_i computes TID_i^1, δ_i and decrypts D_2 as follows:

$$\begin{aligned}
TID_i^1 &= h(h(ID_i) \oplus h(\theta_i) \oplus h(r_i)) \\
\lambda_i &= h(TID_i^1) \\
D_{\lambda_i}(D_2) &= \{\delta_i^1||\theta_i||RN_j\}
\end{aligned}$$

5) Finally, U_i generates a session key $SK_{U_i,S_j} = h(\delta_i^1||r_i||RN_i||RN_j||TS_i||TS_j)$. Then U_i checks the validity of H_2 by comparing it with the computed value $h(TID_i^1||\delta_i^1||SK_{U_i,S_j}||TS_j||RN_j)$. If it is true, U_i accepts SK_{U_i,S_j} as the current session key, then updates new parameters Y_1^*, Y_2^* and Y_3^* as follows:

$$\begin{aligned}
Y_1^* &= B_i \oplus h(RPB_i||r_i) \\
Y_2^* &= h(h(ID_i)||PW_i||\alpha_i) \oplus r_i \\
Y_3^* &= h(\alpha_i||PW_i||r_i)
\end{aligned}$$

V. Analysis

We analyse security and efficiency of the proposed authentication scheme. To prove the security of our proposed scheme, we perform the formal security analysis using the the BAN logic [20]. Furthermore, We perform the informal security analysis in order to verify the security of the proposed scheme is secure with high probability.

A. BAN logic security analysis

The notations of the BAN logic are given in Table II:

TABLE II. Notations of the BAN Logic

Notation	Description
$P\| \equiv X$	P **believes** the statement X
#X	The statement X is **fresh**
$P \lhd X$	P **sees** the statement X
$P\| \sim X$	P once **said** X
$P \Rightarrow X$	P **controls** the statement X
$< X >_Y$	Formula X is **combined** with the formula Y
$\{X\}_K$	Formula X is **encrypted** by the key K
$P \overset{K}{\leftrightarrow} Q$	P and Q communicate using K as the **shared key**
SK	Session key used in the current authentication session

1) Postulates of BAN logic: The postulates of the BAN logic are given below:

1. Message meaning rule :

$$\frac{P \Big| \equiv P \overset{K}{\leftrightarrow} Q, \quad P \lhd \{X\}_K}{P \ |\equiv Q \ | \sim X}$$

2. Nonce verification rule :

$$\frac{P \ |\equiv \#(X), \quad P \ | \equiv Q \Big| \sim X}{P \ |\equiv Q \ | \equiv X}$$

3. Jurisdiction rule :

$$\frac{P \ |\equiv Q \ | \Longrightarrow X, \quad P \ |\equiv Q \ | \equiv X}{P \Big| \equiv X}$$

4. Freshness rule :

$$\frac{P \Big| \equiv \#(X)}{P \Big| \equiv \#\,(X,Y)}$$

5. Belief rule :

$$\frac{P \Big| \equiv (X,Y)}{P \Big| \equiv X.}$$

2) Goals: We have the following goals to demonstrate the secure mutual authentication of proposed protocol:

Goal 1: $S \ |\equiv (S \overset{SK_{U_i,S_j}}{\longleftrightarrow} U)$

Goal 2: $S \ |\equiv U \ |\equiv (S \overset{SK_{U_i,S_j}}{\longleftrightarrow} U)$

Goal 3: $U \ |\equiv (S \overset{SK_{U_i,S_j}}{\longleftrightarrow} U)$

Goal 4: $U \ |\equiv S \ |\equiv (S \overset{SK_{U_i,S_j}}{\longleftrightarrow} U)$

3) Idealized Forms: The idealized forms of the transmitted messages are given below:

Msg_1: $U \rightarrow S$: $(ID_i, TID_i, RN_i, r_i, TS_i, \theta_i)_{K_i}$

Msg_2: $S \rightarrow U$: $(\delta_i^1, RN_j, TS_j, SK_{U_i,S_j})_{\lambda_i}$

4) Assumptions: We make the following initial assumptions to perform the BAN logic proof:

A_1: $S \ |\equiv \#(RN_i)$

A_2: $U \ |\equiv \#(RN_j)$

A_3: $S \ |\equiv (S \overset{K_i}{\longleftrightarrow} U)$

A_4: $U \ |\equiv (S \overset{\lambda_i}{\longleftrightarrow} U)$

A_5: $S \ |\equiv U \Rightarrow (SK_{U_i,S_j})$

A_6: $U \ |\equiv S \Rightarrow (SK_{U_i,S_j})$

5) Proof Using BAN Logic: The detailed steps of the main proof are as follows:

Step 1: According to Msg_1, we can obtain

$$S_1 : S \lhd (ID_i, TID_i, RN_i, r_i, TS_i, \theta_i)_{K_i}$$

Step 2: In conformity with the message meaning rule with S_1 and A_3, we can get

$$S_2 : S \mid\equiv U \mid\sim (ID_i, TID_i, RN_i, r_i, TS_i, \theta_i)_{K_i}$$

Step 3: According to the freshness rule with A_1, we can get

$$S_3 : S \mid\equiv U \mid\equiv \#(ID_i, TID_i, RN_i, r_i, TS_i, \theta_i)_{K_i}$$

Step 4: According to the nonce verification rule with S_2 and S_3, we can obtain

$$S_4 : S \mid\equiv U \mid\equiv (ID_i, TID_i, RN_i, r_i, TS_i, \theta_i)_{K_i}$$

Step 5: According to the belief rule with S_3 and S_4, we can get

$$S_5 : S \mid\equiv U \mid\equiv (RN_i, r_i, TS_i)$$

Step 6: Because of $SK_{U_i,S_j} = h(\delta_i^1 || r_i || RN_i || RN_j || TS_i || TS_j)$ from the S_5 and A_2, we can get. where δ_i^1, RN_j are random number selected by S_j and TS_j is current timestamp.

$$S \mid\equiv U \mid\equiv (S \overset{SK_{U_i,S_j}}{\longleftrightarrow} U) \qquad \textbf{(Goal 2)}$$

Step 7: According the jurisdiction rule with S_6 and A_5, we can obtain

$$S \mid\equiv (S \overset{SK_{U_i,S_j}}{\longleftrightarrow} U) \qquad \textbf{(Goal 1)}$$

Step 8: According to Msg_2, we can obtain

$$S_8 : U \lhd (\delta_i^1, RN_j, TS_j, SK_{U_i,S_j})_{\lambda_i}$$

Step 9: In conformity with the message meaning rule with S_8 and A_4, we can get

$$S_9 : U \mid\equiv S \mid\sim (\delta_i^1, RN_j, TS_j, SK_{U_i,S_j})_{\lambda_i}$$

Step 10: According to the freshness rule with A_2, we can get

$$S_{10} : U \mid\equiv S \mid\equiv \#(\delta_i^1, RN_j, TS_j, SK_{U_i,S_j})_{\lambda_i}$$

Step 11: According to the nonce verification rule with S_9 and S_{10}, we can obtain

$$S_{11} : U \mid\equiv S \mid\equiv (\delta_i^1, RN_j, TS_j, SK_{U_i,S_j})_{\lambda_i}$$

Step 12: According to the belief rule with S_{10} and S_{11}, we can get

$$S_{12} : U \mid\equiv S \mid\equiv (S \overset{SK_{U_i,S_j}}{\longleftrightarrow} U) \qquad \textbf{(Goal 4)}$$

Step 13: According the jurisdiction rule with S_{12} and A_6, we can obtain

$$U \mid\equiv (S \overset{SK_{U_i,S_j}}{\longleftrightarrow} U) \qquad \textbf{(Goal 3)}$$

B. Security analysis against various attacks

Replay attack. Our scheme does not send a real identity ID_i in public channels. $\mathcal{A}$ is required to know TID_i and θ_i to derive ID_i from PID_i, however, $\mathcal{A}$ cannot obtain both TID_i and θ_i. Because r_i is hidden to $\mathcal{A}$ and ID_i, θ_i is only known to an authentic user U_i. Furthermore, PID_i changes in every session, therefore, $\mathcal{A}$ cannot reuse TID_i or PID_i to get any information from S_j in registration phase as we have shown in chapter 3-A.

Resisting off-line Identity and password guessing attack. $\mathcal{A}$ may attempt to guess ID_i from PID_i and Y_2. Suppose $\mathcal{A}$ obtains these values and a smart card SC_i. To find ID_i from PID_i, $\mathcal{A}$ have to know r_i and compute TID_i first, and then guess ID_i and θ_i concurrently, The guessing probability, when ID_i consist of n characters and the hash value is 160 bits, is roughly $1/2^{6n+160}$ and it is a computationally infeasible problem [21]. Therefore, it is infeasible to guess an identity correctly in our scheme.

Resisting off-line password guessing attack. $\mathcal{A}$ may attempt to guess PW_i from Y_2 and Y_3. The probability of guessing PW_i from Y_2 and Y_3 is same as above. $\mathcal{A}$ who somehow gets RPW_i is also required to guess ID_i, PW_i and α_i concurrently, and the probability is more complicated. Therefore, it is infeasible to guess a password correctly.

Forward secrecy and session key exposure. Three keys K_i, λ_i, and SK exist in the proposed scheme. Ephemeral key K_i is computed as $r_i \oplus h(C_i)$. Though $\mathcal{A}$ somehow knows K_i, he/she cannot compute previous ephemeral keys K_i, because r_i changes in every session and is hidden to $\mathcal{A}$. Likewise $\mathcal{A}$ somehow knows λ_i, he/she cannot compute previous ephemeral keys, because r_i changes in every session and is hidden to $\mathcal{A}$. Session key contains random parameters $\{r_i, RN_i, RN_j, TS_i TS_j\}$. Therefore, our scheme provides forward secrecy and withstands the session key exposure.

User anonymity. Our scheme does not send a real identity ID_i in public channels. $\mathcal{A}$ is required to compute TID_i to derive ID_i, however, $\mathcal{A}$ cannot even obtain TID_i because of r_i is hidden. And TID_i changes dynamically into PID_i, thus $\mathcal{A}$ cannot trace U_i using identity information. Therefore, our scheme provides user anonymity.

Resisting user impersonation attack. $\mathcal{A}$ who obtains a smart card SC_i of U_i and tries to access S_j is needed to generate and send a valid login request message $\{PID_i, C_i, M_i, D_1, H_1\}$ to S_j. To compute those values, $\mathcal{A}$ needs to know TID_i, B_i and compute PID_i, M_i, however, $\mathcal{A}$ does not know these parameters. Thus, $\mathcal{A}$ cannot compute valid login messages and finally H_1. Therefore, our scheme withstands the user impersonation attack.

Resisting server impersonation attack. $\mathcal{A}$ needs to compute valid reply messages D_2 and H_2 to masquerade as a server, however, he/she cannot compute valid reply messages because $\mathcal{A}$ cannot get δ_i^1. Therefore, our scheme withstands the server impersonation attack.

Resisting man-in-the-middle attack. $\mathcal{A}$ who knows public channel information between U_i and S_j and has a smart card SC_i can establish a secure channel when $\mathcal{A}$ knows unique information of U_i, such as PID_i, C_i, M_i....... However, as we mentioned above, $\mathcal{A}$ cannot compute those values because r_i is hidden to $\mathcal{A}$ and guess ID_i, PW_i, and α_i. Therefore, our proposal withstands the man-in-the-middle attack.

Resisting stolen smart card attack. $\mathcal{A}$ who somehow possesses a valid smart card SC_i of U_i may attempt to get authentication credentials. But, $\mathcal{A}$ cannot have any advantage because all the parameters are protected with a one-way has function. $\mathcal{A}$ also cannot obtain or compute any login information using SC_i without ID_i, PW_i and α_i. Guessing ID_i and PW_i concurrently is impractical as mentioned above. Therefore, our scheme withstands the stolen smart card attack.

We compare the functionality features of the proposed scheme with Roy et al.'s scheme in Table III. $\circ$ indicates the scheme provides the property or is secure against the attack; $\times$ indicates the scheme does not provide the property or is vulnerable to the attack.

TABLE III. COMPARISONS OF THE FUNCTIONALITY FEATURES

	Roy et al.' scheme [13]	Proposed scheme
replay attack	$\times$	$\circ$
ID guessing attack	$\times$	$\circ$
password guessing attack	$\times$	$\circ$
forward secrecy	$\times$	$\circ$
user anonymity	$\circ$	$\circ$
efficient password change	$\circ$	$\circ$
user impersonation attack	$\circ$	$\circ$
server impersonation attack	$\circ$	$\circ$
man-in-the-middle attack	$\circ$	$\circ$
stolen smart card attack	$\circ$	$\circ$

C. Performance

We compare the cost of computation with Roy et al's scheme in Table IV. T_h indicates the computation time for hash function; T_F indicates fuzzy extraction; XOR are not considered because it can be ignored comparing with T_h. The computation cost of ours is almost similar to [13], and the proposed scheme enhances the security.

TABLE IV. COMPARISONS OF THE COMPUTATION COSTS

	Roy et al.'s scheme [13]		Proposed scheme	
	User	Server	User	Server
Registration	$10T_h + T_F$	$3T_h$	$9T_h + T_F$	$5T_h$
Login	$11T_h + T_F$	0	$13T_h + T_F$	0
Authentication	$10T_h$	$19T_h$	$9T_h$	$19T_h$
Total	$31T_h + 2T_F$	$22T_h$	$31T_h + 2T_F$	$24T_h$

VI. CONCLUSIONS

Several biometric-based remote user authentication schemes using smart card have been proposed in the last few years. Unfortunately, most of them could not provide secure authentication and suffer from various attacks. This paper showed the security flaws of Roy et al.'s scheme. Roy et al's scheme is prone to replay attacks and offline guessing attacks. Furthermore, their scheme does not support perfect forward secrecy. We proposed a secure user authentication scheme in IoT medical environments for better security functionality than that of Roy et al. Our scheme withstands various attacks, such as replay and guessing attacks. In addition, our scheme provide perfect forward secrecy to provide secure authentication. In addition, the proposed scheme provides a dynamic identity mechanism and withstands various attacks by the malicious server.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1D1A3B07050409) and was supported by the Korea Nazarene University Research Grants, 2018.

REFERENCES

[1] J. L. Tsai, N. W. Lo, and T. C. Wu, *"Novel anonymous authentication scheme using smart cards"*, IEEE Transactions on Industrial Informatics, IEEE, Computer Networks, 101(4), 192-202, 2016.

[2] M. S. Hossain and G. Muhammad, *"Cloud-assisted industrial internet of things (IIoT) âĂŞ enabled framework for health monitoring"*, 9(4), 2004-2013, 2013.

[3] Q. Jiang, J. Ma, Z. Ma, and G.Li, *"A privacy enhanced authentication scheme for telecare medical information systems,"* J. Med. Syst., 37, 2013.

[4] D. Mishra, S. Kumari, M. K. Khan, and S. Mukhopadhyay, *"An anonymous biometric-based remote-user authentication key agreement scheme for multimedia systems"*, International Journal of Communication Systems, 2015.

[5] Y. H. Park and Y. H. Park, Y. *"A Selective Group Authentication Scheme for IoT-Based Medical Information System."* Journal of medical systems, 41(4), 48, 2017.

[6] S. Kumari, M. K. Khan, and R. Kumar, *"Cryptanalysis and improvement of 'a privacy enhanced scheme for telecare medical information systems'"*, Journal of medical systems, 37(4), 2013.

[7] T. Cao and J. Zhai, *"Improved dynamic id-based authentication scheme for telecare medical information systems"*, Journal of medical systems, 37(2), 2013.

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

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

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

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

[12] H. M. Chen, J. W. Lo, and C. K. Yeh, An efficient and secure dynamic ID-based authentication scheme for telecare medical information systems, J. Med. Syst., 36(6), pp. 3907-3915, Dec., 2012.

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

[14] D. Dolev and A. Yao, *"On the security of public key protocols"*, IEEE Trans. Inf. Theory, 29(2), 198-208, 1983.

[15] K. S. Park, Y. H. Park, Y. H. Park, and A. K. Das, *"2PAKEP: Provably Secure and Efficient Two-Party Authenticated Key Exchange Protocol for Mobile Environment,"* IEEE Access, 6, 2018.

[16] T. S. Messerges, E. A. Dabbish, and R. H. Sloan, *"Examining smart-card security under the threat of power analysis attacks",* IEEE Trans. Comput., 51(5), 541-552, 2002.

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

[18] G. Xu, S. Qiu, H. Ahmad, G. Xu, Y. Guo, M. Zhang, and H. Xu, *"A Multi-Server Two-Factor Authentication Scheme with Un-Traceability Using Elliptic Curve Cryptography",* Sensors, 18(7), 2018.

[19] A. G. Reddy, E. J. Yoon, A. K. Das, and K. Y. Yoo, *"Lightweight authentication with key-agreement protocol for mobile network environment using smart cards,"* IET Information Security, 10(5), 2016.

[20] M. Burrows, M. Abadi, and R. Needham, *"A logic of authentication,"*, ACM Trans. Comput. Syst., 8(1), 1990.

[21] A. K. Das and A. Goswami, *"A secure and efficient uniqueness-and-anonymity-preserving remote user authentication scheme for connected health care,"* Journal of medical systems, 37(3), 2013.

Round the Clock Vehicle Emission Monitoring using IoT for Smart Cities

Jagadish Nayak
Electrical and Electronics Engineering Department
BITS Pilani, Dubai Campus
Dubai, United Arab Emirates

***Abstract*—Emissions from the vehicles contribute the major part of pollution in this world. Most of the countries have stringent rules to check the emission level through their transport authorities. So as to have zero emission, continuous monitoring of emission level is required. Smart cities need to maintain zero pollution throughout the year. In this paper, an IoT (Internet of Things) based system is proposed for continuous tracking and warning system. The prototype developed is connected to the exhaust of the vehicle and data is collected in the cloud, which can be further processed for a warning system. The device is tested with some vehicles and the results are comparable with the existing emission testing systems used in the market. This device can be used by all vehicle manufacturing companies by embedding it in their products.**

***Keywords*—IoT; Emission; Sensor; Carbon; Tracking; Smartcity**

I. Introduction

The proposed work in this paper will emphasize on vehicle emissions monitoring using Internet of Things (IoT). The motive behind the project, which led towards designing an emission level tracking system in motor vehicles, for mainly due to the upcoming Smart City initiative in most of the countries, and the increased concerns in carbon emissions all over the world. Aiming towards a city that provides mobility, health, safety and productivity is important, but alongside sustainability must be taken into consideration. The main objectives of proposed work is to understand Vehicle emissions and other growing concerns of smart city due to the Carbon Monoxide emission and to develop a round the clock Vehicle Emissions Sensing System (EMSENS), which will regularly track emissions throughout its operation. The concerned authority will warn the vehicle owner regarding CO emission of the vehicle.

As the years go by, people have begun to occupy every possible habitable space. This rise in population had led to a development of distinct types of cities. Every city or town now defined characteristics and are currently evolving socially and economically. Environment challenges and increased greenhouse-gas emissions which are forcing cities to develop sustainability strategies for energy generation and distribution, transportation, water management, urban planning, and eco-friendly (green) buildings. Figure 1 shows the components of a smart city. The smart cities must have least pollution. One of the sources of pollution is CO emissions from the vehicles, which need to be almost zero. But it is practically possible, when there is proper system to track the emission from vehicle all the time.

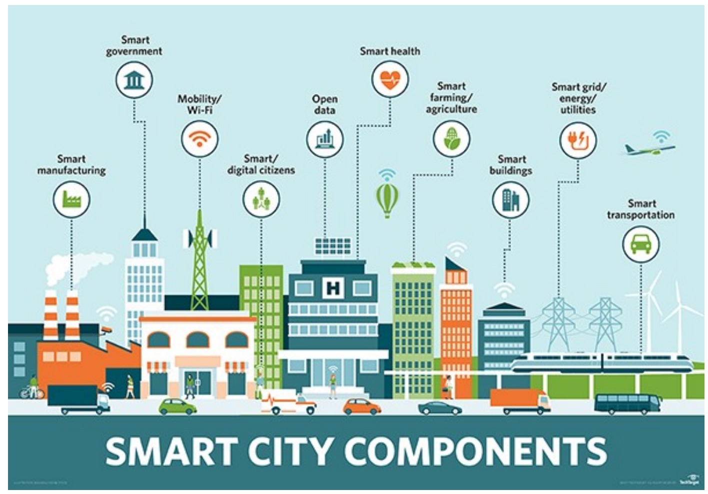


Fig. 1. Components of a Smart City (Source - Learn Internet Governance)

Environmental issues in the world have been on the rise over the years. With the development and bloom of the industry, the demands have also been increasing thus allowing them to be complying to the pressure and thereby giving rise to the already existing environmental conditions [1]. A few other factors which have contributed towards this are the constant depletion and exploitation of the natural resources which they are not really abundant of, coupled with increase in population. Global energy-related carbon dioxide (CO2) emissions are expected to increase by one-third between 2012 and 2040, which is shown in Figure 2, in EIA's International Energy Outlook 2016 (IEO2016) Reference case, largely driven by increased energy use in countries outside of the Organization for Economic Cooperation and Development (OECD). The HC-CO (Hydro Carbon-Carbon Monoxide) tester is designed for checking exhaust emissions of automobiles in service. There is a probe which is placed in the exhaust pipe to determine the emissions. Obviously, the air/fuel mixture delivered to the engine gives the amount of carbon monoxide present. Whereas for HC, they depend on the fuel, and which is either not burned or just partly burned. An emission-controlled automobile running normally has HC emissions below 100 rpm (parts per million) and CO emissions between 0.5 and 1.0%. Non-emission-controlled engines have higher normal emission levels of about 400 ppm HC and 3.0% CO. Readings above the given values is an indication that the vehicle needs to be serviced. It would either be due to some fault in the engine or engine emissions. Tests for HC and CO are run at idle and higher speeds. At higher engine speeds, they must decrease. If the exhaust emissions are normal, no further checking of the emission controls is necessary.

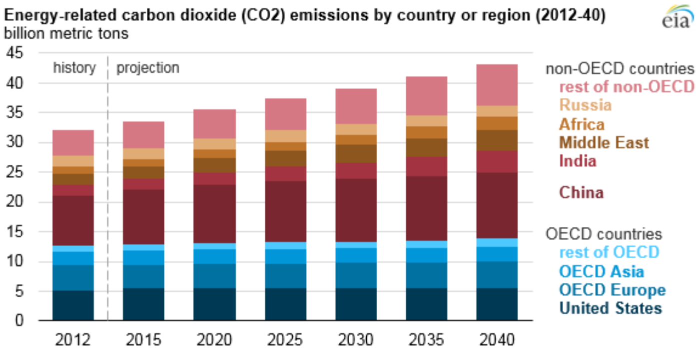


Fig. 2. Energy related Co2 emission by region (Source - U.S. Energy Information Administration, International Energy Outlook 2016)

II. Emission tracking Using IoT

The Internet of things (IoT) is the network of physical devices, vehicles, home appliances and other items embedded with electronics, software, sensors, actuators, and connectivity which enables these objects to connect and exchange data. The "things" in IoT can mean a lot, from a person with implants that automate his heart beat or animals with chips embedded in them which inform owners of their health, sensors in cars alerting drivers about obstacles or anything which can transfer data with minimizing or almost eliminating human to human or human to computer interaction[2]. Figure 3 shows an IoT platform. Some of the work mentioned in the literature[3],[4],[5],[6],[7] proposes methods based on the RFIDs and other methodologies. These systems do not provide any sort of continuous monitoring and warning systems, which is proposed in this paper.

Fig. 3. Internet of Things

The traditional air quality checking framework is very costly and monotonous[8]. Measuring of these otherwise is very time consuming and takes up a lot of power. The prototype that has been discussed gives continuous observation of pollutants which are discharged from the vehicles. In this paper we proposed an IoT based system which will sense the CO emission of an vehicle continuously and updates the data base located in the Transport Authorities. The emission level can also be displayed on the display provided in the car. Warning message can also be sent to the owner of the vehicle regarding the emission level. Proposed system is shown in Figure 4.

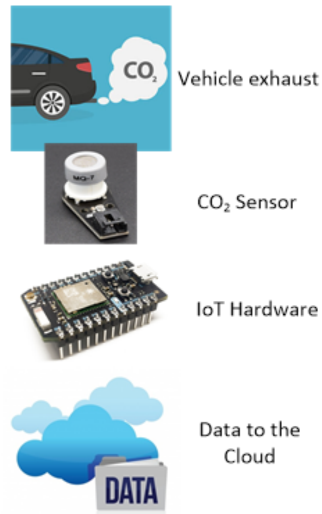


Fig. 4. Proposed prototype model - Design

III. Materials and Method

A. Sensors

The MQ series of Gas Sensors[9] used in this work are highly sensitive, simple and cost-effective sensors useful for sensing gases in the air. Figure 5 shows the typical MQ series of Gas Sensor. There is a wide range of sensors available each of which are made to detect a specific gas like Methane, NOx, SOx, LPG, CNG, Carbon Monoxide and Alcohol. This is a simple-to-use Carbon Monoxide (CO) sensor, suitable for sensing CO concentrations in the air. This sensor works in the temperature ranging from -20^{o}C to 50^{o}C

Fig. 5. Sensor

B. Photon Particle Board

Small and powerful Wi-Fi connected microcontroller Based on Cypress's WICED architecture, the Particle Photon Series combines a powerful STM32 ARM Cortex M3 microcontroller and a Cypress Wi-Fi chip[10]. This is tiny device with many functionalities. For the server, the particle photon is equipped with its own cloud. But there are external cloud services offered as well. We would be able to write the firmware on the web or local IDE, deploy it over the air, and build mobile apps. Figure 6 shows a Photon Particle Board.

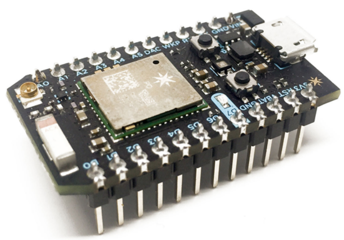


Fig. 6. Photon Particle Board

C. IFTTT(If This Then That)

This is a freeware[11] used to communicate various apps and devices. A trigger (like change of a value, or a new value under the given variable etc.) can be used to bring about the notification. Once the IFTTT detects the trigger, you can perform a variety of function with it. In this project, IFTTT has been incorporated to either send an email to the vehicle owner for high emissions rates or for storing the received values in the server (Google Sheet in this case).

D. Implementation

An enclosed structure consisting of the photon along with the required circuits and resistances (along with 3.3V battery to power up the photon) A PVC pipe with smaller radius than that of the exhaust enclosed within it. Since the sensors have a limitation to temperature, these pipes could be used to reduce the temperature as they travel towards the sensor. Small slits can be made within the box to bring out the sensors which could be hung or placed in front of the exhaust. Another alternative to the PVC would be a water cooler system placed beneath the exhaust, which would help cool down the emissions from the exhaust, but that would be a tedious process considering this would be applicable to all vehicles on the road. Following are the implementation steps.

1) : Connection of sensors to the particle photon: The sensors would be integrated and connected to the IoT device.

2) : Create Particle App. The benefit provided by the photon is the cloud it is equipped with a cloud service. A code can be formulated from common sensors and publish a Particle event.

3) : IFTTT: The event will later act as an IFTTT trigger. Under this the conditions for the sensor emission values can be set. The above setup would be placed near the exhaust of the vehicle.

4) : The trigger in this case would be the notification. If the sensor values exceed the set standards, then an email would be sent to the respective ID (vehicle owner)

IV. Results and Discussions

On testing the system on the vehicle, results were obtained which would be discussed as follows. EMSENS is placed right in front of the vehicle exhaust as shown in Figure 7. The initial setup is done and once the sensors have pre heated the process begins. The particle IDE begins to update with new values of temperature and CO readings.

Fig. 7. EMSENS on the vehicle

The test was done when the car is in neutral phase. It is placed for a period of time for the values to be analyzed. The IFTTT trigger causes the new values to be recorded on new google sheets with respect to the type of reading, which is shown in Figure 8 and 9

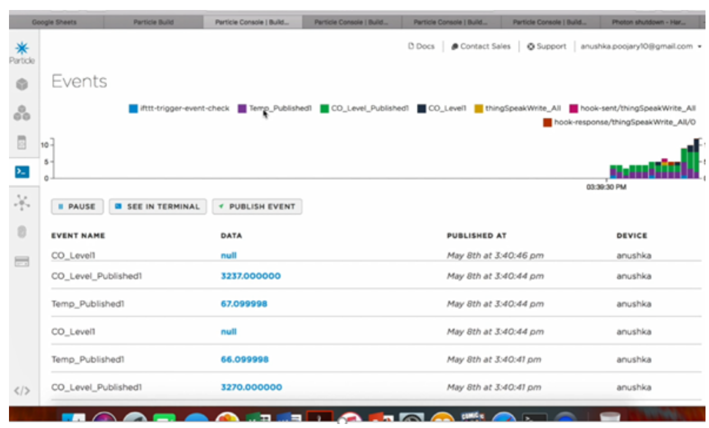


Fig. 8. Particle IDE display

On finding the rough ppm value with respect to the values we have obtained a ppm of 0.0083% was obtained. This when compared to a vehicle passing certificate , the CO ppm percentage was 0.01%. That is almost a 17% deviation from the original. The testing was done on three other vehicles as well giving ppm percentage.

- Mitsubishi Pajero: 0.053%
- Toyota Rav4: 0.01%
- Hyundai i10:0.086%

Gas sensor technologies are still developing and have yet to reach their full potential in capabilities and usage. There are

	A	B	C	D
30	Temp_Published	84.5	anushka	May 8, 2018 at 03:42PM
31	Temp_Published	84.5	anushka	May 8, 2018 at 03:42PM
32	Temp_Published	84.800003	anushka	May 8, 2018 at 03:42PM
33	Temp_Published	85	anushka	May 8, 2018 at 03:42PM
34	Temp_Published	85	anushka	May 8, 2018 at 03:42PM
35	Temp_Published	85	anushka	May 8, 2018 at 03:42PM
36	Temp_Published	85.5	anushka	May 8, 2018 at 03:42PM
37	Temp_Published	85.5	anushka	May 8, 2018 at 03:43PM
38	Temp_Published	85.800003	anushka	May 8, 2018 at 03:43PM
39	Temp_Published	86.099998	anushka	May 8, 2018 at 03:43PM
40	Temp_Published	86.400002	anushka	May 8, 2018 at 03:43PM
41	Temp_Published	86.599998	anushka	May 8, 2018 at 03:43PM
42	Temp_Published	86.599998	anushka	May 8, 2018 at 03:43PM
43	Temp_Published	86.900002	anushka	May 8, 2018 at 03:43PM
44	Temp_Published	87.199997	anushka	May 8, 2018 at 03:43PM
45	Temp_Published	86.5	anushka	May 8, 2018 at 03:43PM
46	Temp_Published	86.199997	anushka	May 8, 2018 at 03:43PM
47	Temp_Published	85.699997	anushka	May 8, 2018 at 03:43PM
48	Temp_Published	85.400002	anushka	May 8, 2018 at 03:43PM
49	Temp_Published	84.900002	anushka	May 8, 2018 at 03:43PM
50	Temp_Published	84.300003	anushka	May 8, 2018 at 03:43PM
51	Temp_Published	83.800003	anushka	May 8, 2018 at 03:43PM
52	Temp_Published	83.300003	anushka	May 8, 2018 at 03:43PM
53	Temp_Published	82.900002	anushka	May 8, 2018 at 03:43PM

Fig. 9. Google sheet records via IFTTT

technologies which give accurate readings without the hassle, but they are quite expensive. For the low-cost method to be effective, various readings will have to be taken and compared against a known value. This way the erroneous readings that could hinder the process can be significantly reduced.

The particle photon is available for 25US$, while the low-cost sensors cost from 10-20US$. The entire system is setup for less than 78US$. Also, by using vehicles more efficiently one can save money. This would either be in terms of fuel expenditure or preventing the wear and tear. Due to the nature of the prototype, the process is automatic with almost no inspection required except for the one present at the servers to observe the values. The main aim of this project is to constantly monitor carbon emissions in the vehicle and notify the user about the levels to notify them when their vehicle is due for service. Since they would be more aware of the emission level of their vehicle, not only are they more aware but also conscious towards the environment. There are concerns with the quality of the data that such sensors can produce. Studies say that when cost is compromised, so is the quality. Thus, important factors like the accuracy and sensitivity are both affected.

Costly sensors usually go through quite a lot of review and evaluation thus making it reliable While the same does not apply for the latter. Another alternative could be to incorporate the sensors within the simple system so give accurate readings while taking in IoT into consideration as well. While we are diverging into an environmentally friendly, some academics are also concerned about the e-waster that is produced post life of these sensors. Another concern to be noted is of IFTTT, is that there tend to bug within the set hence not allowing the applets to run. For Google Sheets it has a limit of only 100 posts a day.

V. Conclusion

The environmental pollution due to the emissions from the vehicles need to be tracked continuously to control the carbon foot print on this earth. Especially in smart cities continuous monitoring is a challenging task. The proposed prototype can track the emission from vehicle and warns the vehicle user to make proper regular maintenance of the vehicle. Transport authority can also take benefit of this device to keep track of vehicle emissions in respective cities.

References

[1] FastFacts U.S. Transportation Sector Greenhouse Gas Emissions 1990 2016 Offce of Transportation and Air Quality EPA-420-F-18-013 July 2018

[2] Peter Waher *Learning Internet of Things*, Packt Publishing, January 2015

[3] S.P. Bangal, Gite Pravin E, Ambhure Shankar G, Gaikwad Vaibhav M, *IoT Based Vehicle Emissions Monitoring and Inspection System*, International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 5, Issue 4, April 2017.

[4] Souvik Manna , Suman Sankar Bhunia , Nandini Mukherjee , *Vehicular pollution monitoring using IoT*, International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014) 9-11 May 2014.

[5] Rose Sweetlin .T, Priyadharshini .D ,Preethi .S, Sikkandar Sulaiman .A , *Control of vehicle pollution through Internet of things (IOT)*, International Journal Of Advance Research, Ideas And Innovations In Technology, Volume-3, Issue-2, 2017.

[6] R.Priyanka, S.K. Thai Bhuvana, Archanaa Raveendran, Dr. R. Kavitha , *Vehicle Pollutants Control Using Sensors And Arduino*, Third International Conference on Sensing, Signal Processing and Security (ICSSS) 4-5 May 2017

[7] Sneha S. Kamalakar and M. S. Vanjale , *Inspection and Notification of Vehicle Emissionin Smart Cities*, Indian Journal of Science and Technology, Vol 9(S1) December 2016

[8] ELAWEJ, Khalifa A. K. , *A framework for the evaluation of air pollution caused by motor vehicles.*, Doctoral, Sheffield Hallam University (United Kingdom),2014

[9] Technical data manual MQ-7 Gas Sensor, HANWEI Electronics Co. Ltd http://www.hwsensor.com

[10] Technical data Sheet Particle Photon Board :https://docs.particle.io/datasheets/

[11] A world that works for you , IFTTT website https://ifttt.com/

The Implementation of an IoT-Based Flood Alert System

Wahidah Md. Shah*, F. Arif†, A.A. Shahrin‡ and Aslinda Hassan*
*Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100 Durian Tunggal
Melaka, Malaysia
†Department of Industrial Engineering
Institut Teknologi Nasional Bandung, Indonesia
‡Iverson Associates Sdn Bhd
Petaling Jaya, Selangor, Malaysia

***Abstract*—Floods are the most damaging natural disaster in this world. On the occasion of heavy flood, it can destroy the community and killed many lives. The government would spend billions of dollars to recover the affected area. It is crucial to develop a flood control system as a mechanism to reduce the flood risk. Providing a quick feedback on the occurrence of the flood is necessary for alerting resident to take early action such as evacuate quickly to a safer and higher place. As a solution, this paper propose a system that is not only able to detect the water level but also able to measure the rise speed of water level and alerted the resident. Waterfall model is adopted as the methodology in this project. Raspberry Pi is used to collect data from the water sensor and transmit the data to GSM module for sending an alert via SMS. The analysis will be done to show how the Raspberry Pi will be integrated with the smartphone to give an alert. The system is tested in an experiments consist of two different environment in order to ensure that the system is able to provide accurate and reliable data. The project is an IoT-based which significantly in line with the Industrial Revolution 4.0, supporting the infrastructure of Cyber-Physical System.**

Keywords*—Flood Alert System; Internet of Things; Cyber-Physical System; IR4.0

I. INTRODUCTION

The worlds climate is changing drastically due to effect from human activities such as pollutions, cut countless trees, excessive gas emission etc. Floods are among the most common damaging natural disasters, that cause significant harm to life, property, and economy. Scientists estimate by 2030, if 4-inch sea level rise, it could potentially caused the severe flooding in many parts of the world [1].

Since Malaysia is located near the equator, the most severe climatic related natural disasters are monsoonal flood. It happens almost every year and causes a lot of damages, property loss, and not to mention the loss of life during the disaster. Recently in Jan 2018, two died and nearly 12,000 evacuated when flood strike in Pahang as reported in [2]. The worst flood in Malaysian history happened in 2014. More than 200,000 people were affected while 21 were killed. The major disasters happened in the several states on the east coast side of Peninsular Malaysia. The estimated cost of damages was over RM1 billion as reported in [3]. The impact of the flood is huge and it is not happening in Malaysia but all over the world.

This paper proposes a flood warning system that can detect the water level and measure the speed of the rise of water level. To give the society an earlier notification to evacuate before the water rises to the dangerous level, the measurement result is sent as the alert to a mobile phone through Short Message Service (SMS). This project is designed on the IoT-based platform, where data from the sensor is collected at the mini-processor and alert is generated and transmitted as SMS to a smartphone. The proposed system is implemented in an experimental setting in two different environment to test its effectiveness.

Section 2 explains some works that have been conducted in the related field. Section 3 presents the methodology used in developing the system that the Waterfall model. Section 4 analyzes the result of the experiment and discusses them accordingly. Section 5 concludes the paper with an outline of contribution and the future work.

II. RELATED WORKS

Some article such as [4] and [5] has reviewed on flood disaster and disaster management in Malaysia which highlight the importance of finding best remedies to prepare if disaster strikes. Futhermore, Leman et al. [5] also reviewed the current disaster management in Malaysia for proposing enhancement the effectiveness. They suggest four phases of action; preparedness, response, recovery and mitigation. Due to this phenomenon, several experts have studied over the years on ways to improve flood control, thereby reducing the aforementioned risks. Better understanding of the flood hazard phenomenon and its potential consequences in our society is crucial for the development of flood risk reduction projects, control policies and other types of flood management strategies. The adoption of IoT based-system has attract attention among researchers since data is collectively sensed by sensors to provide various services without human intervention [6]. Early work by Lo et al. [7] propose to automatically monitor the flood object based on the remote cyber surveillance systems. Image processing methods are being utilized to obtain instant flooding and waterlogging event feedback.

On the other hand, [8] used pressure sensor to read the water level at every second, for detecting the level of water. In the occasion where the water level surpasses a user-defined threshold, an SMS is texted to the residence for warning and quick action. Recent from Jana Priya et al. [9] and Satria et al. [10] demonstrate the idea and implementation of a flood monitoring and alert system using measuring sensor. They proposed a system that measures the height of the water using ultrasonic sensor. While, [11] combine ultrasonic and water level sensor to monitor flood level conditions however it missing alert system.

Inspired by the literature, this work aimed to detect flood with water sensor with different level of measurement with SMS alert. SMS is considered as the most effective way to the scenario Malaysia since smartphone has rapidly become the preferred device for most Malaysian to remain connected. According to the survey [12], the percentage of smartphone users continue to rise from 68.7% in 2016 to 75.9% in 2017. The increment trend shows that the majority of people in Malaysia use mobile phones and it is a good strategy to give an early warning to the society using an SMS via mobile phone.

III. Methodology

Waterfall Model is flows steadily downwards (like a waterfall) through the seven phases; comprises of Planning, Requirement, Design, Implementation, Testing, Maintenance and Documentation. The following are the details of each phase for starting with the backgroup study until the completion of project.

A. Requirement

The first phase is to study the requirement to explore the research area including background study and also related works. From there, the research gap is identified and proceed to identify the hardware and software requirement. Hardwares that involved are water sensor, SEN113104 model and GSM module (model is a USB 3G modem Huawei mobile broadband E173). A breadboard is used to place and set up all the equipment at one proper place, Resistor 10K to reduce the current that flow of the system and also many types of jumper cables to connect the water sensor, raspberry pi and GSM module. The software used is a python language to code the entire program.

B. Design

The design phase presented the system design as presented in Fig. 1. There are two sensors used which placed at two different height. The first one will be placed at a lower level. This is the level of the potential flood that will be anticipated. When the water reaches this level, it triggers the water sensor and data is transmitted to Raspberry Pi and pass to GSM module for generating SMS alert to the residents, as a warning to be cautious and prepared. If the water continues to rise and reaches the second water sensor, it is considered now as dangerous, an alert SMS once again sent to the resident and authorities.

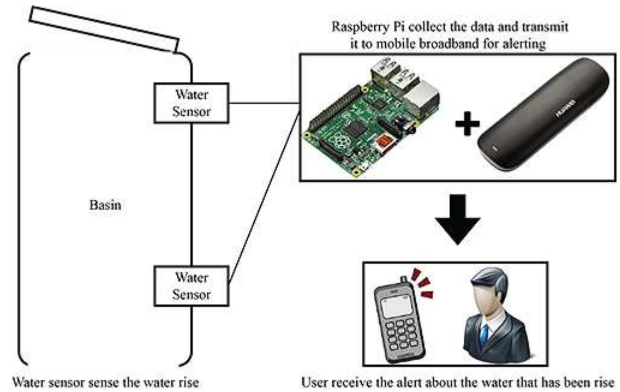


Fig. 1. Flood Alert System Design

C. Implementation

In this phase, Raspberry Pi, water sensors and GSM Module are already being installed and developed based on the design. All the wiring that needs to be installed on the breadboard will be installed. Also, the coding for the Raspberry Pi will be set up in this phase for the water sensor and GSM module. This phase will take a lot of time because there will be many errors and unexpected problems that may occur during the configuration and development of the system. Verification and testing of the system also need to be done in this phase. The experiment has been set up as such a basin is placed with the two water sensors where the sensor will be placed on a certain height. Then, the sensor will be connected to the Raspberry Pi and the Raspberry Pi is connected to a Huawei mobile broadband so that it can send an alert message to the user.

This project is tested actual water where the water sensor will be placed in a basin (see Figs. 2). Then, when the water rises and reaches to the first sensor, automatically it will send an alert message to the authorized user or the head of society to inform about the water rise. Later, when the water reaches the second sensor which is much more higher than the first one, an alert message will be sent to inform that the water is now at a dangerous level, and also it will calculate the time for the water to reach the second sensor with the speed of water rising too.

D. Verification

In this phase, the SMS based flood early warning system using Raspberry Pi is tested to ensure if the system could send the early warning. The sensor will also be tested to see whether the detection of the liquid is okay or not. GSM Module should be checked to see whether it can give an SMS as an alert to the mobile phone. This phase is an important phase because it really needs to verify and test the system before it can be implemented. The performance of the system is tested as discussed in Section 3

E. Maintenance

The maintenance phase is to ensure that the sensor continue working to capture data anytime when water is rising. It is also important to ensure the GSM Module is ready to send SMS alert when needed. Maintenance is needed to be performed

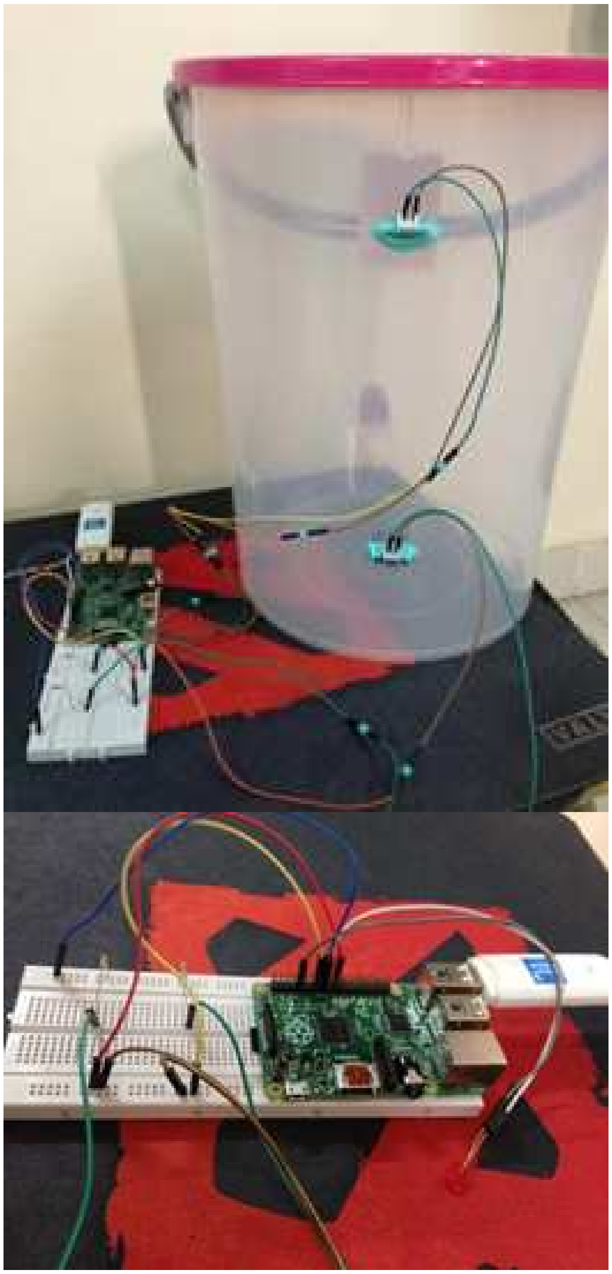


Fig. 2. The prototype of system and experiment set up

from time to time so that, if there any error or malfunction to this system, an appropriate action will be taken. The system needs to always update any new software or any efficient code so that the system is at the top of the performance. This phase is the important part because it will determine the performance and how long will the system withstand.

IV. Result and Discussion

Overall, the system able to send an alert via SMS when the water rises to the pre-determined level. A quick notification through SMS is vital since the system is aimed at alerting people and authorities. Therefore, time delay is used for performance testing and evaluation. Three testings are as shown in Table 1. Entire testing is carried out in a controlled environment where two assumptions made; in the occasion of the water level rise fast or slow; emulating the condition of flash flood and monsoon flood respectively.

TABLE I. Performance Testing Scheduled

Test Environment	Performance Testing		
	Test 1	Test 2	Test 3
A - Full Volume Flow Rate	delay time of water detection by the two sensors	delay time between sent and received messages	delay time of sent water speed
B - Half Volume Flow Rate			

For fast water risen, Full Volume Flow Rate environment is used where the water tab is open at its full speed (Environment A). Meanwhile, Half Volume Flow Rate (Environment B) is set with the water tab is open at half of it so that the water rises slowly. Table 1 summarized the test environment and the performance testing.

A. Test 1 Water Detection

Test 1 is designed to measure the accuracy of time measured by the system which compared with the manual time reading. For each environment, 30 series of reading are captured manually by digital clock and automatically from system time (produced by the system). Since 30 lines of time are quite long, Table II and III only shown the partial result obtained for Environment A and B respectively. The time of water detection at sensor 1 and sensor 2 by the system is compared with the manual reading.

It is not much difference between the manual and system time. For both environment, although some readings are different, on average the difference is not more than 10 seconds. This is expected since the program need some time to acquire data from sensor and pass it to the system. The result appears to verify that the system time is accurate.

TABLE II. Result of Test 1 for Environment A

Sensor 1			Sensor 2		
Manual	System	Time Difference (s)	Manual	System	Time Difference (s)
11:45:53	11:45:53	0:00:00	11:16:03	11:16:06	0:00:03
11:23:43	11:23:43	0:00:00	11:24:59	11:24:59	0:00:00
11:30:51	11:30:52	0:00:01	11:32:03	11:32:03	0:00:00
11:36:17	11:36:18	0:00:01	11:37:27	11:37:28	0:00:01
11:41:51	11:41:51	0:00:00	11:43:02	11:43:02	0:00:00
⋮	⋮	⋮	⋮	⋮	⋮
2:07:33	2:07:35	0:00:02	2:08:43	2:08:47	0:00:04
Minimum		0:00:00	Minimum		0:00:00
Maximum		0:00:02	Maximum		0:00:08
Average		0:00:09	Average		0:00:02

B. Test 2 Delay Time of Received SMS

This test aimed to measure the delay of received SMS relative to the time it is sent. The partial result is tabulated in Table IV and Table V. In both environments, the delay between message sent and received at the sensors are on the average of two seconds. While it is depends on the performance of telecommunication provider, the maximum 7 seconds of delay is acceptable. Resident and authorities can still be alerted in a timely manner for appropriate action.

TABLE III. RESULT OF TEST 1 FOR ENVIRONMENT B

Sensor 1			Sensor 2		
Manual	System	Time Difference (s)	Manual	System	Time Difference (s)
9:31:52	9:31:54	0:00:02	9:35:04	9:35:07	0:00:03
9:39:53	9:39:55	0:00:02	9:42:56	9:42:59	0:00:03
9:48:17	9:48:17	0:00:00	9:51:23	9:51:26	0:00:03
9:55:52	9:55:53	0:00:01	9:59:12	9:59:14	0:00:02
10:03:13	10:03:16	0:00:03	10:06:14	10:06:14	0:00:00
:	:	:	:	:	:
1:23:29	1:23:30	0:00:01	1:26:40	1:26:42	0:00:02
Minimum		0:00:00	Minimum		0:00:00
Maximum		0:00:02	Maximum		0:00:06
Average		0:00:08	Average		0:00:02

TABLE IV. RESULT OF TEST 1 FOR ENVIRONMENT A

Sensor 1			Sensor 2		
Message Sent	Message Receive	Delay (s)	Message Sent	Message Received	Delay (s)
11:14:533	11:14:57	0:00:04	11:16:06	11:16:07	0:00:01
11:23:43	11:23:45	0:00:02	11:24:59	11:25:01	0:00:02
11:30:52	11:30:54	0:00:02	11:32:03	11:32:05	0:00:02
11:36:18	11:36:19	0:00:01	11:37:28	11:37:29	0:00:01
11:41:51	11:41:53	0:00:02	11:43:02	11:43:05	0:00:03
:	:	:	:	:	:
2:07:35	2:07:37	0:00:02	2:08:47	2:08:48	0:00:01
Minimum		0:00:0	Minimum		0:00:01
Maximum		0:00:06	Maximum		0:00:07
Average		0:00:02	Average		0:00:02

TABLE V. RESULT OF TEST 2 FOR ENVIRONMENT B

Sensor 1			Sensor 2		
Message Sent	Message Receive	Delay (s)	Message Sent	Message Received	Delay (s)
9:31:54	9:31:58	0:00:04	9:35:07	9:35:09	0:00:0
9:39:55	9:39:56	0:00:01	9:42:59	9:43:00	0:00:01
9:48:17	9:48:20	0:00:03	9:51:26	9:51:28	0:00:02
9:55:53	9:55:55	0:00:02	9:59:14	9:59:17	0:00:03
10:03:16	10:03:19	0:00:03	10:06:14	10:06:16	0:00:02
:	:	:	:	:	:
1:23:30	1:23:32	0:00:02	1:26:42	1:26:45	0:00:03
Minimum		0:00:011	Minimum		0:00:01
Maximum		0:00:05	Maximum		0:00:04
Average		0:00:02	Average		0:00:02

C. Test 3 Water Speed Calculation

Test 3 is for summarizing the finding on the rate of water increment speed. The value can be used as a guideline for estimating the how quick the flood rise. Table 6 shows that as expected, in environment A the rate of water increment is quicker than environment B. The time needed for the height of water to increased 1 meter is less in environment A compared to B. The information about the rate of water increased is sent via SMS as well once water reaches the second sensor.

TABLE VI. RATE OF WATER INCREMENT SPEED IN ENVIRONMENT A AND B

Water Speed (meter/seconds)		Increased height of water in 1 meter (seconds)	
Environment A	Environment B	Environment A	Environment B
0.000273973	0.000103627	365	965
0.000263158	0.000108696	380	920
0.00028169	0.00010582	355	945
0.000285714	0.0000995025	350	1005
0.00028169	0.00011236	355	890
0.000277778	0.000104167	360	960

V. CONCLUSION

This work enlighting the possibility to provide an alert system to overcome the flood risk. It can also contribute to the authority or company such as firefighter or any government state agency that can help the society about the flooding or any natural disaster. The proposed prototype system has been tested and it works as proposed. It is able to send an alert message to the user with the time of the water rise also with the speed of the water rises for prediction how quick is the flood is happening. It also has been tested in a controlled environment to evaluate the performance. For real implementation, several sensors might be integrated for accurate detection such as pressure sensor [8] and also camera [6].

ACKNOWLEDGMENT

Special thanks of gratitude to the Universiti Teknikal Malaysia Melaka for sponsoring the fund under the UTeM Short Term Grant with the reference number: PJP/2018/FTMK(1B)/S01628. A high appreciation to Digital Forensics and Computer Networking (INSFORNET) research group under Center for Advanced Computing Technology (C-ACT); and Faculty of Information and Communication Technology (FTMK) the use of the existing facilities to complete this research.

REFERENCES

[1] R. Becker, "A future of more extreme floods, brought to you by climate change," May 2017. [Online]. Available: https://www.theverge.com/2017/5/18/15658342/flooding-sea-level-rice-melting-ice-climate-change-extreme

[2] Hermesauto, "Two dead, nearly 12,000 evacuated in malaysia floods," Jan 2018. [Online]. Available: https://www.straitstimes.com/asia/se-asia/two-dead-nearly-12000-evacuated-in-malaysia-floods

[3] "Amaran hujan lebat peringkat jingga di kelantan, terengganu," 2018. [Online]. Available: https://www.bharian.com.my/node/23430

[4] Z. M. Taib, N. S. Jaharuddin, and Z. D. Mansor, "A review of flood disaster and disaster management in malaysia," *International Journal of Accounting & Business Management*, vol. 4, no. 3, 2016.

[5] A. M. Leman, K. A. Rahman, M. N. M. Salleh, I. Baba, D. Feriyanto, L. S. C. Johnson, and S. N. Hidayah M., "A review of flood catastrophic management in malaysia," *ARPN Journal of Engineering and Applied Sciences*, vol. 11, no. 14, Jul 2016.

[6] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2347–2376, Fourthquarter 2015.

[7] S. W. Lo, J. H. Wu, F. P. Lin, and C. H. Hsu, "Cyber surveillance for flood disasters," *Sensors (Switzerland)*, 2015.

[8] S. Azid, B. Sharma, K. Raghuwaiya, A. Chand, S. Prasad, and A. Jacquier, "SMS based flood monitoring and early warning system," *ARPN Journal of Engineering and Applied Sciences*, 2015.

[9] S. J. Priya, S. Akshaya, E. Aruna, J. A. M. Julie, and V. Ranjani, "Flood monitoring and alerting system," *International Journal of Computer Engineering & Technology (IJCET)*, vol. 8, no. 2, p. 15, Mar 2017.

[10] D. Satria, S. Yana, R. Munadi, and S. Syahreza, "Prototype of google maps-based flood monitoring system using arduino and gsm module," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 10, Oct 2017.

[11] M. Madhumathi and R. Grace, "Flood alert management system using iot and microcontroller," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 4, April 2017.

[12] Malaysian Communications And Multimedia Commission, "Hand phone users survey 2017," Tech. Rep., 2017.

Evaluating the Quality of UCP-Based Framework using CK Metrics

Zhamri Che Ani, Nor Laily Hashim, Hazaruddin Harun
School of Computing
Universiti Utara Malaysia
06010 UUM Sintok, Kedah, Malaysia

Shuib Basri, Aliza Sarlan
Department of Computer and Information Sciences
Universiti Teknologi PETRONAS
31750 Tronoh, Perak, Malaysia

***Abstract*—Software effort estimation is one of the most important concerns in the software industry. It has received much attention since the last 40 years to improve the accuracy of effort estimate at early stages of software development. Due to this reason, many software estimation models have been proposed such as COCOMO, ObjectMetrix, Use Case Points (UCP) and many more. However, some of the estimation methods were not designed for object-oriented technology that actively encourages reuse strategies. Therefore, due to the popularity of UCP model and the evolution of the object-oriented paradigm, a UCP-based framework and supporting program were developed to assist software developers in building good qualities of software effort estimation programs. This paper evaluates the quality of the UCP-based framework using CK Metrics. The results showed that by implementing the UCP-based framework, the quality of the UCP-based program has improved regarding the understandability, testability, maintainability, and reusability.**

***Keywords*—ucp-based framework; use case points; ck metrics**

I. Introduction

Software effort is defined as the person months required to make a software application [1]. This definition is close to [2] who define software effort as the number of staff days/weeks/months or even years associated with a project. Software effort estimation (SEE) can broadly be defined as the process of estimating the effort required to develop a software project [3]. It has been focused by many researchers over the past 40 years [4] and nowadays, it has become one of the most important concerns of the software industry [5]–[9].

There are many software estimation models have been proposed to improve the accuracy of effort estimate at early stages of software development [10]–[12]. However, some estimation methods were not designed to work well with object-oriented technology that introduces inheritance and actively encourages reuse strategies. SLIM [13], Checkpoint [14], PRICE-S [15], SEER [16], COCOMO II [10], ObjectMetrix [17], [18] and Use Case Points [19] are among the popular object-oriented estimation models that have been widely used in previous studies.

Use Case Points (UCP) is a software sizing and estimation method adopted from the standard Function Point (FP) method in solving the specific needs of object-oriented systems based on use cases [20], [21]. It was developed by Gustav Karner at Objectory Systems [19]. Previous studies have demonstrated that the accuracy of UCP estimations was quite close to the actual estimates [22]–[25]. Due to the popularity of UCP, in the last two decades, many UCP-based effort estimation techniques have been proposed [24], [26]–[35], either to give more options or to enhance the capability of UCP. Studies also showed that some parts of UCP-based models have similarity in estimating software effort [19], [24], [30], [32]. However, just a few of them are equipped with proper estimation tools. Most of them used MS Excel to calculate the estimates [36], [37].

To date, software developers have very little guidance to develop quality UCP-based software effort estimation using the object-oriented approach. Even though a tool known as U-EST was developed using Java programming language, the design framework was not accessible to the public [38]. The authors also did not claim that the tool was well-designed and able to be reused or extended by other developers. Therefore, a new framework for UCP-based software effort estimation was developed to promote the reusability of UCP. Without reusability, software applications are very hard to maintain or extend [39]–[42].

Therefore, this study aims to evaluate the quality of UCP-based framework using CK Metrics. The framework was designed using UML notations after identifying the class dependencies using Java programming language. The remainder of this paper is structured as follows. Section II and III provide some basic concept of the UCP-based framework and CK Metrics respectively. Section IV describes the experimental research design. Experimentation results and discussion are discussed in section V. Section VII includes conclusion and suggestion for future work.

II. The Proposed UCP-Based Framework

The UCP-based framework is defined as *a general reusable solution for UCP-based software effort estimation design*. This framework is not a finished design that can be transformed directly into source codes. It is a description or template for solving a UCP-based problem that can be used in many different design approaches [43]. This framework was formed based on four UCP-based models namely Use Case Points (UCP) [19], Adapted Use Case Points (AUCP) [30], Industrial use of Use Case Points (IUCP) [24] and Simplified Use Case Points (SUCP) [32]. Fig. 1 illustrates the UCP-based framework.

The UCP-based framework shows how the elements are structured, and how they work together. In other words, the UCP-based framework is more to the abstract level of

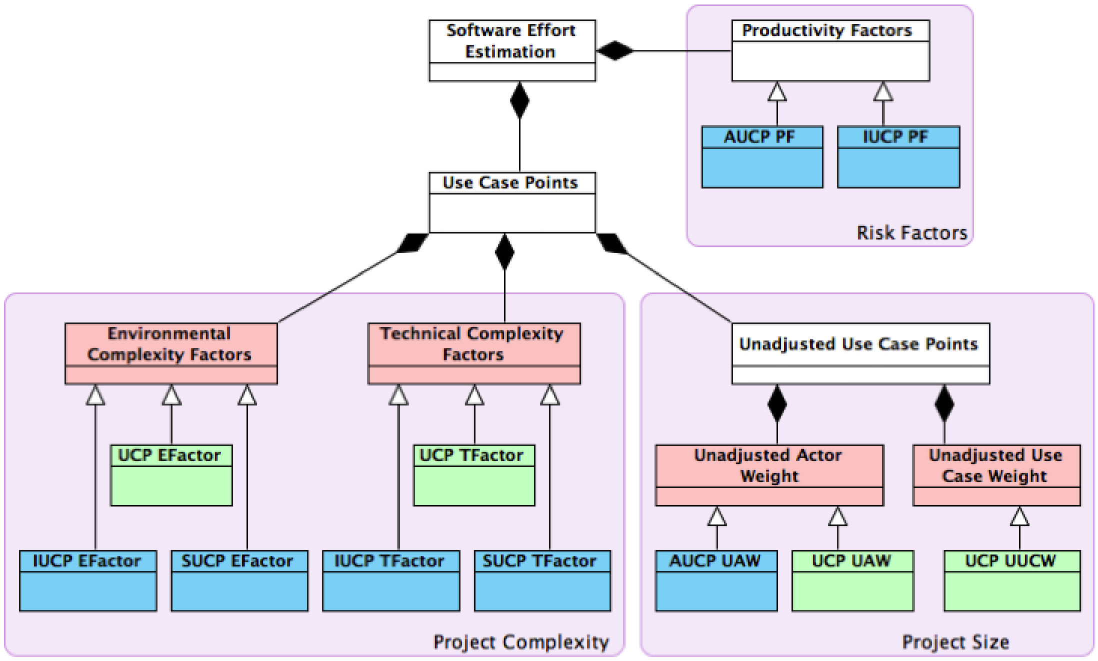


Fig. 1. A UCP-Based Framework

software design, where the abstraction and implementation are independent. The implementations may vary dynamically [44]. This framework captured the important aspects of the UCP-based models to visualize the main packages, classes, and the relationship among them. By using this UCP-based framework, software designers can easily conduct experiments and propose the possible well-designs which can contribute to high-quality software. This framework is also useful especially in software maintenance because it suggests the high-level aspects of UCP-based requirements. By using this framework, the existing UCP-based programs can be changed or extended systematically.

The main elements of the UCP-based framework are classes and use two types of relationships namely association and generalization. The classes determine the general concept of UCP domain knowledge where every software designer familiar and understandable. Classes can be interpreted at various levels in software design. In the early stages of software design, the UCP-based framework captures more logical aspects of the problem. In the later stages, the framework can be extended to any object-oriented design decisions based on the software designer's experience and creativity. In this study, a class is drawn as a rectangle.

Overall, there are 19 classes, and 12 of them are the principal classes where the principal classes are derived from the UCP model. In general, the UCP-based framework is divided into three main components: project size, project complexity, and risk factors. These three main components are based on the estimating principle defined by Garmus and Herron [45]. Project size is composed of six classes. Five of the classes are the principal classes while `AUCP_UAW` class is the extended class. Project complexity includes four principal classes as well as four extended classes namely `IUCP_EFactor`, `SUCP_EFactor`, `IUCP_TFactor`, and `SUCP_TFactor`. Meanwhile, risk factor has only one principal class namely `Productivity_Factors` and two extended classes.

Relationships among classes are drawn as paths connecting class rectangles. Generalization shows the relationship between a more general class and a more specific class which is used for inheritance. In this framework, 11 classes are inherited from their parent classes. For instance, `AUCP_UAW` class is extended from `Unadjusted_Actor_Weight` class. Associations carry information about the relationship among objects in UCP-based domain knowledge. For instance, the `Use_Case_Points` class is associated with `Unadjusted_UseCasePoints` class. All principal classes which are captured from nine steps of UCP [36] are associated with the association relationship. This means that without these key classes the effort estimation cannot be done completely.

III. Chidamber and Kemerer (CK) Metrics

Software metrics play a major role in comparing different versions of object-oriented programs [46]. One of the most popular object-oriented metrics and the most thoroughly investigated is CK Metrics [47]. It was introduced by Chidamber and Kemerer [48] and has been a subject of discussion since the last two decades. The authors themselves and other researchers have carried out a series of experiments to improve the accuracy of the metrics. Even though it was proposed quite a long time ago, the usefulness in analyzing open-source software is still significant [49]–[52], including examining the reusability of software projects [53].

Chidamber and Kemerer proposed six metrics for evaluating the quality of software design namely Weighted Method per Class (WMC), Response For a Class (RFC), Lack of Cohesion in Methods (LCOM), Coupling Between

Objects (CBO), Depth of Inheritance Tree (DIT) and Number of Children (NOC). The definition of each metric can be found in [48]. The metrics have been theoretically validated and widely accepted standard to be used as early quality indicators [52], [54]–[58]. By using these metrics, it will help software designers to make a better decision in designing any object-oriented software applications.

IV. Experimental Setting

The object-oriented approach is very important among all software practitioners. Generally, encapsulation, inheritance, and polymorphism are three main object-oriented principles that are applied throughout the whole object-oriented software engineering process [59]. Instead of these three principles, abstraction is also considered as a key element in designing software applications. By using abstraction, the complexity of a large problem can be minimized. The idea of abstraction is to identify common features in two or more classes and abstract those features out into a higher-level class [60]. A good design, each method in a well-designed class should support abstraction and encapsulation [61]. In this case, all common functionalities of UCP-based models were grouped to form reusable abstract classes.

In this experiment, a new program known as Like-UCP was developed to simulate the UCP-based framework. In order to achieve the consistency between the UCP-based framework and Like-UCP program, the Like-UCP was designed based on four object-oriented programming principles; abstraction, inheritance, encapsulation, and polymorphism. The design must ensure that, for every abstract class of Like-UCP, the subclass must implement the method from the abstract class unless the subclass is also an abstract class.

After completing the development, CK Metrics was used to measure the quality of Like-UCP program. It is impossible to obtain the Like-UCP metrics without using supporting tools. Thus, in this study, CK Java Metrics (CKJM) extended version 2.2 [62] was used to obtain the metrics. CKJM is an open-source program which was also written in Java programming language used to obtain six CK Metrics of the object-oriented programs. In many cases, no single design style can meet all quality attributes simultaneously. Software architects or designers often need to balance among quality attributes. Most of the time, a new program needs refinement, extension, generalization, or improvement [63]. Therefore, if the quality of Like-UCP program did not achieve the desired quality attributes, repetition process from designing the UCP-based framework must be done.

Then, each of the classes was analyzed, and each metric was calculated to obtain the mean value. The mean value of each metric was used as an indicator of quality measurements. Finally, the obtained results were concluded based on the relationship between CK Metrics, Object-oriented Design (OOD) measures, and quality factors. Table I and Table II show the relationship between CK Metrics and OOD measures, and the relationship between CK Metrics and quality factors respectively [64].

V. Experimentation Results and Discussion

The Like-UCP program was developed using Java programming language to evaluate the quality of the UCP-based framework. To visualize the dependencies between classes, class diagram of Like-UCP was generated in Eclipse environment. Fig. 2 shows the class dependencies of Like-UCP program. Basically, each of the classes is a replication of UCP-based framework using the object-oriented approach. As can be seen in Fig. 2, some of the classes such as `UCP_UUCW`, `Unadjusted_UseCasePoints`, and `UseCasePoints` have high dependencies compared to other classes. In other words, these three classes are required by all the identified UCP-based models to form the UCP-based framework.

TABLE I. Relationship between CK Metrics and OOD Measures

OOD Measures	CK Metrics
Class	WMC, RFC, LCOM
Attribute	LCOM
Method	WMC, RFC, LCOM
Inheritance	DIT, NOC
Cohesion	LCOM
Coupling	CBO, RFC

TABLE II. Relationship between CK Metrics and Quality Factors

Quality Factors	CK Metrics
Understandability	RFC, CBO, DIT
Reusability	WMC, CBO, DIT, NOC
Testability	RFC, CBO, NOC
Maintainability	WMC, CBO
Development Effort	WMC, LCOM

To obtain the empirical evidence of quality attributes achieved by Like-UCP program, 22 classes (19 classes derived from the UCP-based framework with three additional driver classes) were analyzed using CKJM-extended-2.2. Table III and Table IV present the descriptive statistics of Like-UCP program and the summary of all the metrics respectively. In this experiment, the total number line of codes (LOC) of Like-UCP is 2556 with the mean value 116.18. The maximum LOC is 380 used by `IUCP` class, and the minimum is 11 used by `UseCasePoints` class. The next sub-sections will discuss the obtained results and compare with the suggestions in [48].

A. *Weighted Method per Class (WMC)*

The number of methods and the complexity of the methods involved indicate how much time and effort is required to develop and maintain the class. The larger the number of methods in a class, the higher the potential impact on subclasses, since subclasses will inherit all the methods defined in the superclass. It was recommended that most classes should have a small number of methods, maximum up to 10 methods in a class [48]. If WMC value is one, it was recommended that to merge the class in some other classes within the same package without influencing the LCOM value, that is without affecting the abstraction and encapsulation of the classes. If WMC value is zero, the possibility of a redesign is high. If WMC over than 20, the class should be refactored to reduce the complexity of the software project. 20 WMC is consistent with the threshold value suggested by Shatnawi [65].

Based on the WMC statistics shown in Table IV, it can be seen that the mean value of the WMC metric is 6.50 which is less than 10. Only four classes, `TechnicalComplexity`, `EnvironmentalComplexity`, `IUCP_PF`, and `IUCP` have more than 10 but still below 20 WMC. None of the classes has

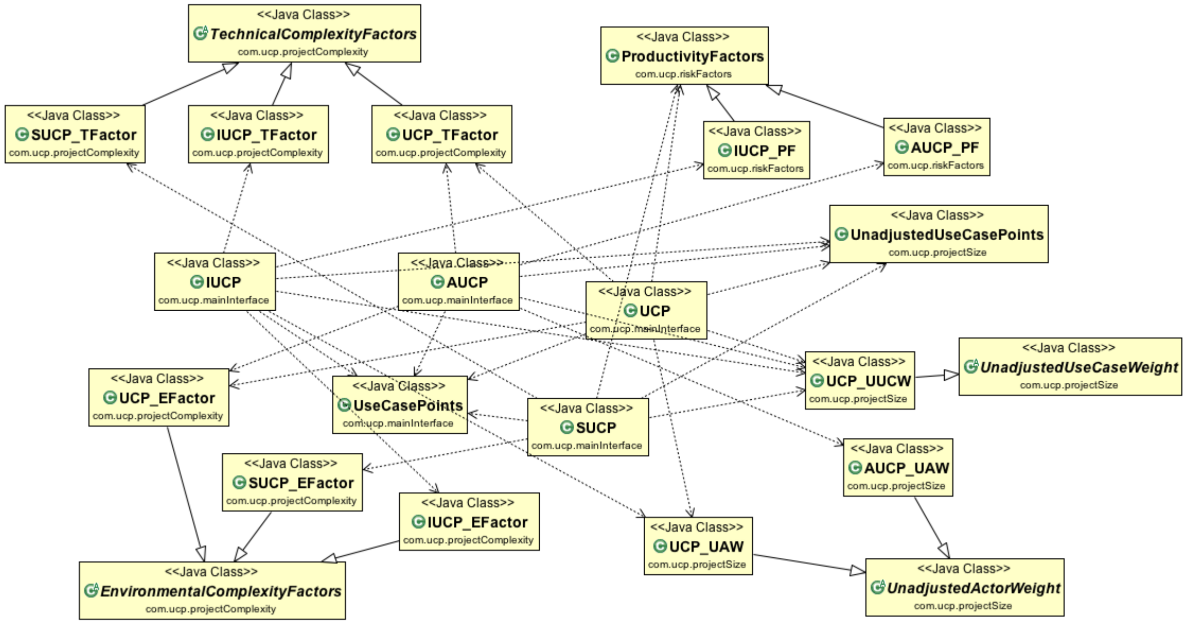


Fig. 2. Class Dependencies of Like-UCP Program

TABLE III. CK METRICS FOR LIKE-UCP PROGRAM

CLASS	LOC	WMC	RFC	LCOM	CBO	DIT	NOC
UCP	281	9	34	24	8	1	0
UseCasePoints	11	2	3	1	4	1	0
UnadjustedUseCasePoints	12	3	4	3	4	1	0
UnadjustedActorWeight	70	9	10	8	2	1	2
AUCP_UAW	24	2	5	1	2	2	0
UCP_UAW	40	2	7	1	3	2	0
UnadjustedUseCaseWeight	91	10	11	3	1	1	1
UCP_UUCW	65	3	9	1	5	2	0
EnvironmentalComplexity	171	15	16	45	3	1	3
IUCP_EFactor	86	2	13	1	2	2	0
UCP_EFactor	102	4	14	6	3	2	0
SUCP_EFactor	38	2	7	1	2	2	0
TechnicalComplexity	265	20	21	96	3	1	3
IUCP_TFactor	126	2	19	1	3	2	0
UCP_TFactor	142	4	19	6	4	2	0
SUCP_TFactor	54	2	9	1	2	2	0
ProductivityFactors	30	7	8	15	5	1	2
AUCP_PF	23	3	4	3	2	2	0
IUCP_PF	127	12	13	28	2	2	0
IUCP	380	12	46	54	9	1	0
AUCP	218	10	33	33	8	1	0
SUCP	200	8	30	18	7	1	0

TABLE IV. SUMMARY OF CK METRICS FOR LIKE-UCP PROGRAM

Metrics	Total	Mean	Maximum	Minimum
LOC	2556	116.18	380	11
WMC	143	6.50	20	2
RFC	335	15.23	46	3
LCOM	350	15.91	96	1
CBO	84	3.82	9	1
DIT	33	1.5	2	1
NOC	11	0.50	3	0

zero or one WMC. These results indicate that all the classes were properly designed.

B. Response For a Class (RFC)

RFC is the number of methods that can be invoked in response to a message in a class. If RFC for a class is large, it means that there is high complexity [48]. If RFC increases, the effort required for testing will increase because the test sequence grows. The overall design complexity of the class also increases and it is difficult to maintain the classes later on. The RFC for a class should usually not exceed 50 although it is acceptable to have RFC up to 100 [66]. However, Shatnawi [65] suggested that the threshold value for RFC is 40. Based on the RFC statistics shown in Table IV, it can be seen that the mean value of the RFC metric is 15.23. None of the classes exceed 50. Fig. 3 shows that only one class which is `IUCP` has more than 40 RFC. Overall, the low values of RFC indicate that the Like-UCP design is less complex.

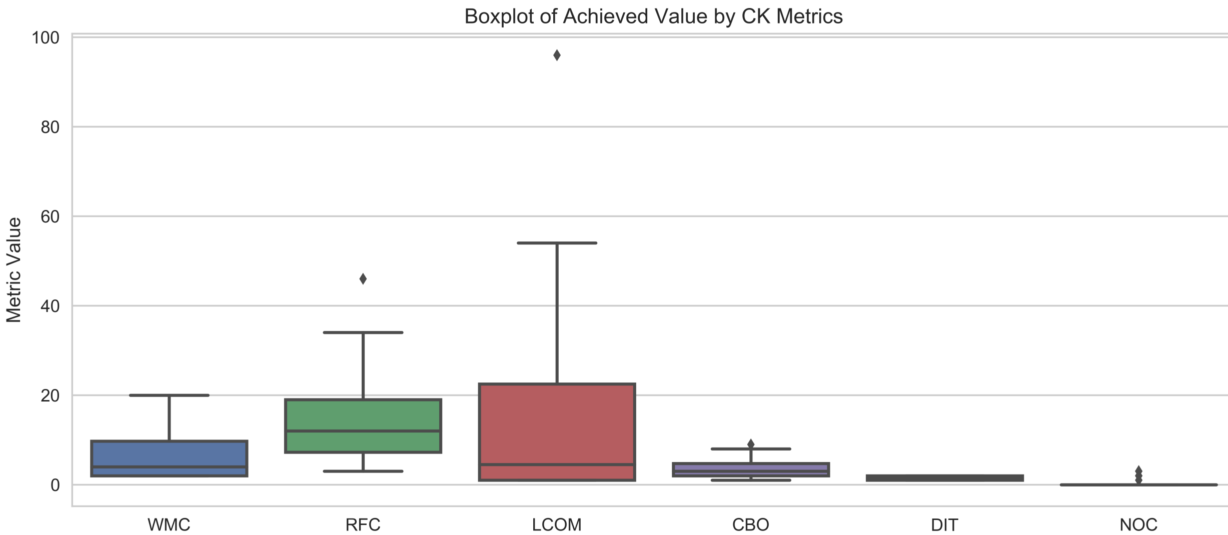


Fig. 3. Boxplot of Achieved Value by CK Metrics

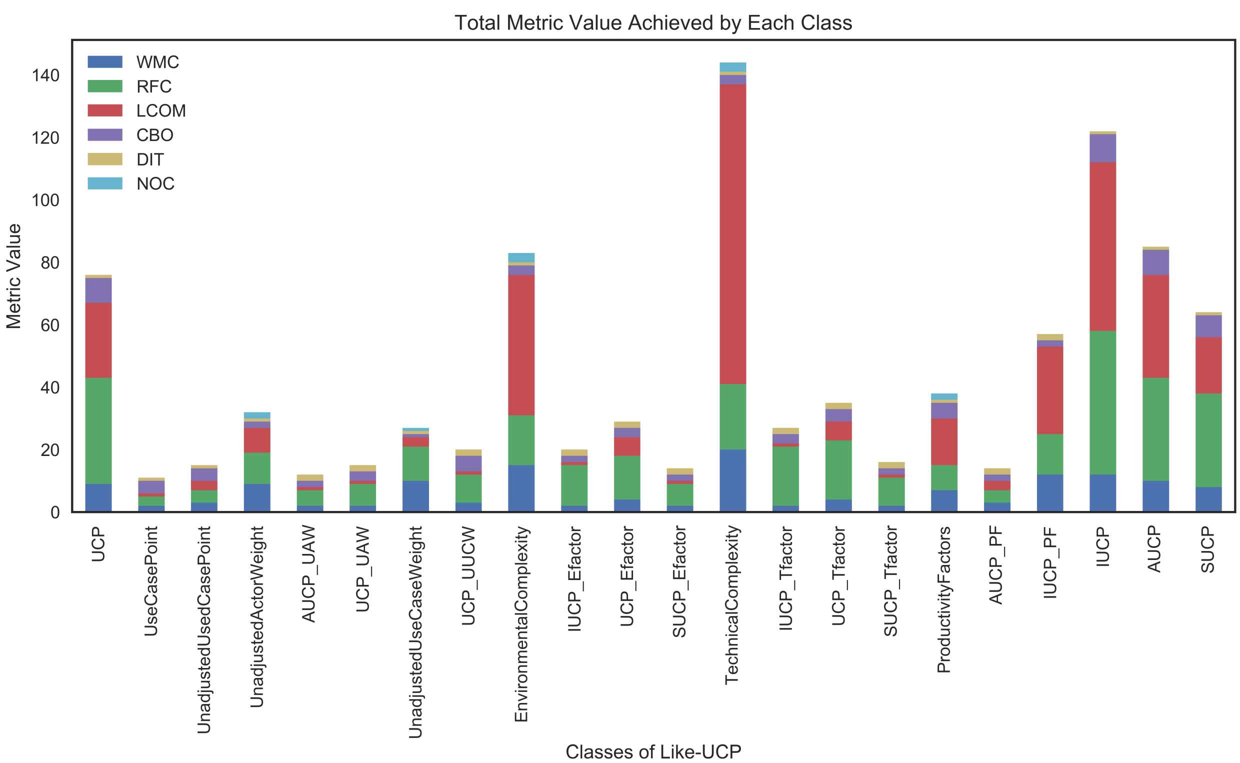


Fig. 4. Total Metric Value Achieved by Each Class

C. Lack of Cohesion in Methods (LCOM)

The cohesiveness of methods inside a class is desirable since it promotes encapsulation and decreases the complexity of the objects. High cohesion decreases complexity, thereby decreasing the probability of errors during the development process [48]. Accordingly, a high LCOM value is a major issue in all the software projects. High LCOM value shows poor encapsulation and abstraction at the class level. The general threshold for LCOM is classified as Good: 0, Regular: range between $1 - 20$ and Bad: > 20 [67]. Based on the LCOM statistics shown in Table IV, the mean value of LCOM is 15.91. This value indicates that most of the classes are normal and less complex. Only six classes, `UCP`, `AUCP`, `IUCP`, `IUCP_PF`, `TechnicalComplexity`, and `EnvironmentalComplexity` have greater than 20. However, as can be seen in Fig. 3, only one class which is `TechnicalComplexity` has a very high value of LCOM. Fig. 4 also shows that the total value of `TechnicalComplexity` class is also the highest among all classes of Like-UCP. The results indicate that the `TechnicalComplexity` class should be re-designed to improve the quality the overall quality of Like-UCP program.

D. Coupling Between Objects (CBO)

In general, objects should be loosely coupled, which implies there ought to be a little dependency between objects. High coupling implies that the sensitivity to changes in other classes is also high which may indicate poor class design [48]. A class which is tightly coupled will cause a large ripple effect when an object is changed and increasing the risk of regressions. Furthermore, testing all the changes will be tedious and hard to check. In contrast, the low CBO values demonstrate that most of the classes refer to a few other classes. In other words, the more independent a class is, the easier to reuse it in other applications. Therefore, it increases the understandability, efficiency, and reusability of the class design [48]. A measure of coupling is useful to determine how complex the testing of various parts of a class design are likely to be. It was suggested that classes with a CBO more than 19 should be examined to reduce coupling [65]. Based on the CBO statistics shown in Table IV, the mean value of CBO is 3.82. None of the CBO values greater than 19. The maximum value of Like-UCP is nine which is equal to the threshold value suggested by Shatnawi [65]. Therefore, it can be concluded that the possibilities of understandability, efficiency, and reusability of the class design are very high.

E. Depth of Inheritance Tree (DIT)

DIT is the length of the longest path from a subclass to the superclass in the inheritance hierarchy. A high value of DIT implies more reusability, but the complexity of the class design may increase due to more methods and classes involved [48]. Classes with bunches of subclasses need to be very carefully modified to avoid regressions in those subclasses. One conceivable solution is that the classes ought to be more abstract with a reasonable number. It was proposed that the maximum value of DIT should be less than 10 [48]. If DIT value is zero or one, it demonstrates poor reusability of the class. Based on the DIT statistics shown in Table IV, the mean value of DIT is 1.5. This value indicates that inheritance was rarely used in Like-UCP program, mainly because of the domain model. Consequently, the program did not have a deep inheritance tree. However, the maximum value of DIT is two which is evaluated as *Good/Common* by DIT threshold [68]. Therefore, with 1.5 DIT, Like-UCP program can be concluded as potentially to be reused and easy to be understood by software developers.

F. Number of Children (NOC)

Inheritance is a form of reuse. Thus, the greater the number of subclasses for a class, the greater reuse. However, a large number of subclasses may introduce inappropriate abstraction of the superclass. It is better to have depth than breadth in the inheritance hierarchy [48]. In other words, high DIT and low NOC is the perfect combination for software design. Low values of DIT and NOC are firmly indicated that reuse through inheritance may not be completely adopted [48]. Based on the NOC statistics shown in Table IV, the majority of the classes (77%) have no subclasses, and the mean value of NOC is 0.50 which is less than the mean value of DIT. Fig. 3 shows that Like-UCP program has a minimal number of outliers. The results indicate that the Like-UCP program may not be using inheritance of methods as a basis for designing classes. However, these results are consistent with the previous case studies [48]. Due to the mean value of NOC smaller than the mean value of DIT, most probably the Like-UCP classes are potential to be reused by other UCP-based programs.

VI. Threats to Validity

Each of metric tools might interpret and produce a different set of results. To avoid this problem, the replication of this study must follow the interpretation that we describe in this paper. However, we encourage readers to use different metric tools with different programming languages.

VII. Conclusion and Future Work

This study presents an evaluation of UCP-based framework using CK Metrics. In order to achieve this objective, Like-UCP program was developed using Java programming language. The program was a replication of the UCP-based framework using the object-oriented approach. Then the developed program was examined using CK Metrics to evaluate whether the Like-UCP has accomplished the desired quality standard. As can be seen in Fig. 4, `TechnicalComplexity` and `EnvironmentalComplexity` have higher metric values. Therefore, these two classes need more treatments in order to improve the quality of Like-UCP program. As mentioned earlier, `IUCP`, `AUCP`, and `SUCP` are the class drivers of Like-UCP. These classes were merely used for program simulation. Thus, the high metric values of these three classes will not affect the overall quality of UCP-based framework.

Overall, based on WMC, RFC, CBO, DIT and NOC obtained by CKJM-extended-2.2, it can be concluded that Like-UCP program was well designed. Low value of LCOM indicates that most of these classes were also less complex. Although this experiment was based on a small sample of programs, the results have provided additional evidence that by implementing the object-oriented approach, the quality of Like-UCP has improved regarding the understandability, testability, maintainability, and reusability of the program. The results also indicate that less effort was required to develop the Like-UCP program. As mentioned earlier, Like-UCP is a replication of UCP-based framework. Therefore, the achievement of these quality attributes also reflects the UCP-based framework. In other words, the UCP-based framework can also be said good qualities regarding the understandability, testability, maintainability, and reusability.

The results also show that the program did not have a deep inheritance tree due to the fewer functionalities in the domain model. However, this can be improved by implementing more UCP-based models in the program. Hence, we intend to additionally investigate this topic by utilizing other object-oriented design methodologies such as implementing SOLID Design Principles and make a comprehensive comparison against these outcomes.

As has been mentioned earlier, U-EST was also developed using Java programming language. However, there was no evidence regarding the quality of the program. Therefore, to date, we are unable to compare the Like-UCP program with other ucp-based programs. To ensure this ucp-based framework has a significant contribution to the software practitioners, we encourage other researchers to simulate this framework using new design approaches and compare the outcome with our results.

Besides CK metrics, there are several object-oriented techniques for evaluating the quality design of object-oriented programs. One of the most commonly used is QMOOD (Quality Model for Object-oriented Design). Therefore, we also plan to investigate this topic further using QMOOD and make a comparison between the two approaches.

Acknowledgment

The researchers acknowledge the financial support (Research Generation University Grant) received from University Utara Malaysia (S/O Code: 13847).

References

[1] S. Tariq, M. Usman, R. Wong, Y. Zhuang, and S. Fong, "On learning software effort estimation," in *3rd International Symposium on Computational and Business Intelligence (ISCBI)*. IEEE, 2015, pp. 79–84.

[2] L. Laird and M. Brennan, *Software measurement and estimation: a practical approach*. New Jersey: John Wiley and Sons, 2006.

[3] L. L. Minku and X. Yao, "Which models of the past are relevant to the present? a software effort estimation approach to exploiting useful past models," *Automated Software Engineering*, vol. 24, no. 3, pp. 499–542, 2017.

[4] C. Jones, "Software cost estimation in 2002," *The Journal of Defense Software Engineering*, vol. 15, no. 6, pp. 4–8, 2002.

[5] S. Basha and D. Ponnurangam, "Analysis of empirical software effort estimation models," *International Journal of Computer Science And Information Security*, vol. 7, no. 3, 2010.

[6] F. Brooks Jr, "Three great challenges for half-century-old computer science," *Journal of the ACM*, vol. 50, no. 1, pp. 25–26, 2003.

[7] M. Jørgensen and D. Sjøberg, "The impact of customer expectation on software development effort estimates," *International Journal of Project Management*, vol. 22, no. 4, pp. 317–325, 2004.

[8] S. W. I. Kuan, "Factors on software effort estimation," *International Journal of Software Engineering & Applications*, vol. 8, no. 1, pp. 23–32, 2017.

[9] A. Trendowicz, J. Münch, and R. Jeffery, "State of the practice in software effort estimation: a survey and literature review," *Software Engineering Techniques*, vol. 4980, pp. 232–245, 2011.

[10] B. Boehm, "Software engineering economics," *IEEE Transactions on Software Engineering*, vol. 10, pp. 4–21, 1984.

[11] B. Boehm, C. Abts, and S. Chulani, "Software development cost estimation approaches-a survey," *Annals of Software Engineering*, vol. 10, no. 1-4, pp. 177–205, 2000.

[12] A. Trendowicz and R. Jeffery, *Software Project Effort Estimation: Foundations and Best Practice Guidelines for Success*. Switzerland: Springer International Publishing, 2014.

[13] L. H. Putnam and W. Myers, *Measures for excellence: reliable software on time, within budget*. Prentice Hall Professional Technical Reference, 1991.

[14] J. Capers, *Applied software measurement*. McGraw-Hill, 1996.

[15] R. Park, "The central equations of the price software cost model," in *4th COCOMO Users Group Meeting*, 1988.

[16] R. Jensen, "An improved macrolevel software development resource estimation model," in *5th ISPA Conference*, 1983, pp. 88–92.

[17] G. Adens and R. Armstrong, "Objectmetrix process," UK:TASSC, Tech. Rep., 2008.

[18] G. Adens, "Interpreting and calibrating metrics," UK:TASSC, Tech. Rep., 2009.

[19] G. Karner, "Resource estimation for objectory projects," *Objective Systems SF AB*, vol. 17, 1993.

[20] R. Clemmons, "Project estimation with use case points," *The Journal of Defense Software Engineering*, pp. 18–22, 2006.

[21] G. Banerjee, "Use case estimation framework," in *Annual IPML Conference*. Citeseer, 2004, pp. 1–12.

[22] B. Anda, H. Dreiem, D. Sjøberg, and M. Jørgensen, "Estimating software development effort based on use cases - experiences from industry," *The Unified Modeling Language. Modeling Languages, Concepts, and Tools*, vol. 2185, pp. 487–502, 2001.

[23] B. Anda, "Comparing effort estimates based on use case points with expert estimates," *Empirical Assessment in Software Engineering (EASE 2002), Keele, UK*, 2002.

[24] E. Carroll, "Estimating software based on use case points," in *20th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*, San Diego, CA, USA, 2005, pp. 257–265.

[25] S. Nageswaran, "Test effort estimation using use case points," in *Quality Week 2001*, San Francisco, California, USA, 2001.

[26] M. R. Braz and S. R. Vergilio, "Software effort estimation based on use cases," in *30th Annual International Computer Software and Applications Conference, COMPSAC'06.*, vol. 1. IEEE, 2006, pp. 221–228.

[27] S. Diev, "Use cases modeling and software estimation: applying use case points," *ACM SIGSOFT Software Engineering Notes*, vol. 31, no. 6, pp. 1–4, 2006.

[28] W. Fan, Y. Xiaohu, Z. Xiaochun, and C. Lu, "Extended use case points method for software cost estimation," in *International Conference on Computational Intelligence and Software Engineering.* IEEE, 2009, pp. 1–5.

[29] M. M. Kirmani and A. Wahid, "Revised use case point (re-ucp) model for software effort estimation," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 3, pp. 65–71, 2015.

[30] P. Mohagheghi, B. Anda, and R. Conradi, "Effort estimation of use cases for incremental large-scale software development," in *27th International Conference on Software Engineering*. IEEE, 2005, pp. 303–311.

[31] N. Nunes, L. Constantine, and R. Kazman, "iucp: Estimating interactive-software project size with enhanced use-case points," *IEEE software*, vol. 28, no. 4, pp. 64–73, 2011.

[32] M. Ochodek, J. Nawrocki, and K. Kwarciak, "Simplifying effort estimation based on use case points," *Information and Software Technology*, vol. 53, no. 3, pp. 200–213, 2011.

[33] K. Periyasamy and A. Ghode, "Cost estimation using extended use case point (e-ucp) model," in *International Conference on Computational Intelligence and Software Engineering.* IEEE, 2009, pp. 1–5.

[34] G. Robiolo, C. Badano, and R. Orosco, "Transactions and paths: two use case based metrics which improve the early effort estimation," in *3rd International Symposium on Empirical Software Engineering and Measurement, ESEM.* IEEE, 2009, pp. 422–425.

[35] A. Srivastava, S. Singh, and S. Q. Abbas, "Advancement of ucp with end user development factor: Aucp," *International Journal of Software Engineering & Applications (IJSEA)*, vol. 6, no. 2, pp. 1–10, 2015.

[36] Ç. Gencel, L. Buglione, O. Demirors, and P. Efe, "A case study on the evaluation of cosmic-ffp and use case points," in *3rd Software Measurement European Forum*, Rome, Italy, 2006.

[37] Z. Mansor, S. Yahya, and N. H. H. Arshad, "Review on traditional and agile cost estimation success factor in software development project," *International Journal of New Computer Architectures and their Applications (IJNCAA)*, vol. 1, no. 4, pp. 942–952, 2011.

[38] S. Kusumoto, F. Matukawa, K. Inoue, S. Hanabusa, and Y. Maegawa, "Estimating effort by use case points: Method, tool and case study," in *10th International Symposium on Software Metrics*, Washington, USA, 2004, pp. 292–299.

[39] C. Alexander, S. Ishikawa, and M. Silverstein, *A pattern language:*

towns, buildings, construction. Oxford University Press, USA, 1977, vol. 2.

[40] M. Grand, *Java enterprise design patterns*. Wiley, 2002.

[41] P. Kuchana, *Software architecture design patterns in Java*. CRC Press, 2004.

[42] C. Lasater, *Design patterns*. Jones & Bartlett Learning, 2006.

[43] Z. C. Ani, S. Basri, and A. Sarlan, "A framework for designing ucp-based effort estimation," *Advanced Science Letters*, vol. 24, no. 2, pp. 995–998, 2018.

[44] K. Lano, J. L. Fiadeiro, and L. Andrade, *Software design using Java 2*. New York: Springer, 2002.

[45] D. Garmus and D. Herron, "Estimating software earlier and more accurately," *Journal of Defense Software Engineering*, vol. 15, no. 6, pp. 18–21, 2002.

[46] A. Kaur, S. Singh, D. K. Kahlon, and P. S. Sandhu, "Empirical analysis of ck & mood metric suit," *Int. Journal of Innovation, Management and Technology*, vol. 1, no. 5, pp. 447–452, 2010.

[47] M. Genero, M. Piattini, and C. Calero, "A survey of metrics for uml class diagrams," *Journal of Object Technology*, vol. 4, no. 9, pp. 59–92, 2005.

[48] S. Chidamber and C. Kemerer, "A metrics suite for object oriented design," *IEEE Transactions on Software Engineering*, vol. 20, pp. 476–493, 1994.

[49] K. Johari and A. Kaur, "Validation of object oriented metrics using open source software system: an empirical study," *ACM SIGSOFT Software Engineering Notes*, vol. 37, no. 1, pp. 1–4, 2012.

[50] T. Honglei, S. Wei, and Z. Yanan, "The research on software metrics and software complexity metrics," in *International Forum on Computer Science-Technology and Applications, IFCSTA'09.*, vol. 1. IEEE, 2009, pp. 131–136.

[51] S. Srivastava and R. Kumar, "Indirect method to measure software quality using ck-oo suite," in *International Conference on Intelligent Systems and Signal Processing (ISSP)*. IEEE, 2013, pp. 47–51.

[52] R. Subramanyam and M. Krishnan, "Empirical analysis of ck metrics for object-oriented design complexity: Implications for software defects," *IEEE Transactions on Software Engineering*, vol. 29, no. 4, pp. 297–310, 2003.

[53] D. P. Darcy and C. F. Kemerer, "Oo metrics in practice," *Software, IEEE*, vol. 22, no. 6, pp. 17–19, 2005.

[54] V. R. Basili, L. C. Briand, and W. L. Melo, "A validation of object-oriented design metrics as quality indicators," *IEEE Transactions on Software Engineering*, vol. 22, no. 10, pp. 751–761, 1996.

[55] L. Cheikhi, R. E. Al-Qutaish, A. Idri, and A. Sellami, "Chidamber and kemerer object-oriented measures: Analysis of their design from the metrology perspective," *International Journal of Software Engineering & Its Applications*, vol. 8, no. 2, 2014.

[56] N. I. Churcher, M. J. Shepperd, S. Chidamber, and C. Kemerer, "Comments on" a metrics suite for object oriented design," *IEEE Transactions on Software Engineering*, vol. 21, no. 3, pp. 263–265, 1995.

[57] M. Hitz and B. Montazeri, "Chidamber and kemerer's metrics suite: a measurement theory perspective," *Transactions on Software Engineering*, vol. 22, no. 4, pp. 267–271, 1996.

[58] G. Succi, W. Pedrycz, S. Djokic, P. Zuliani, and B. Russo, "An empirical exploration of the distributions of the chidamber and kemerer object-oriented metrics suite," *Empirical Software Engineering*, vol. 10, no. 1, pp. 81–104, 2005.

[59] K. Qian, X. Fu, L. Tao, and C.-w. Xu, *Software architecture and design illuminated*. Boston: Jones & Bartlett Learning, 2010.

[60] J. F. Dooley, *Software Development, Design and Coding: With Patterns, Debugging, Unit Testing, and Refactoring (2nd ed.)*. [e-book]. Retrieved from https://www.apress.com.: Apress, 2017.

[61] A. Dingle, *Software Essentials: Design and Construction*. Boca Raton: Chapman and Hall/CRC, 2014.

[62] M. Jureczko and D. Spinellis, *Using Object-Oriented Design Metrics to Predict Software Defects*, ser. Monographs of System Dependability. Wroclaw, Poland: Oficyna Wydawnicza Politechniki Wroclawskiej, 2010, vol. Models and Methodology of System Dependability, pp. 69–81.

[63] S. Lammers, *Programmers at work*. Redmond, WA: Harper & Row Publishers, Inc., 1986.

[64] U. Kulkarni, Y. Kalshetty, and V. Arde, "Validation of ck metrics for object oriented design measurement," in *3rd International Conference on Emerging Trends in Engineering and Technology (ICETET)*. IEEE, 2010, pp. 646–651.

[65] R. Shatnawi, "A quantitative investigation of the acceptable risk levels of object-oriented metrics in open-source systems," *IEEE Transactions on software engineering*, vol. 36, no. 2, pp. 216–225, 2010.

[66] M. Sarker, "An overview of object oriented design metrics," Master's thesis, Department of Computer Science, Umeå University, Sweden, 2005.

[67] K. A. Ferreira, M. A. Bigonha, R. S. Bigonha, L. F. Mendes, and H. C. Almeida, "Identifying thresholds for object-oriented software metrics," *Journal of Systems and Software*, vol. 85, no. 2, pp. 244–257, 2012.

[68] T. G. Filó, M. Bigonha, and K. Ferreira, "A catalogue of thresholds for object-oriented software metrics," in *The First International Conferences on Advanced and Thrends in Software Engineering*, 2015, pp. 48–55.

Tele-Ophthalmology Android Application: Design and Implementation

Rachid Merzougui, Mourad Hadjila
STIC Laboratory
Telecommunications Department
Faculty of Technology
University of Tlemcen, Algeria

Nadia Benmessaoud and Mokhtaria Benaouali
Faculty of Technology
University of Tlemcen,
Algeria

***Abstract*—Diabetic retinopathy is the leading cause of blindness in the world population. Early detection and appropriate treatment can significantly reduce the risk of loss of sight. Medical authorities recommend an annual review of the fundus for diabetic patients. Several screening programs for diabetic retinopathy in the world have been made to implement this recommendation. The purpose of this paper is to implement the idea and principle to facilitate the detection of this disease using tele-ophthalmology and the latest telecommunications technologies. The present paper aims to develop an Android app named "RETINA" which captures retinal photographs using a microscopic lens, process them via several operators based on mathematical morphology after a filtering step by exploiting the library "OpenCV", and stores them in a local or remote MySQL database. The application also offers a small service and facilitates the task for the doctor and ophthalmologist by allowing the inspection of files of remote patients.**

Keywords*—Diabetic retinopathy; Screening; Fundus; Tele-ophthalmology; Android; Mathematical morphology; OpenCV; Database

I. Introduction

The advent of information technology and communications (ICT) has revolutionized the field of health to support the transformation of care delivery approaches [1], [2], [3]. In this sense, they represent an important asset allowing to provide answers to the new challenges health systems face [4], [5]. Indeed, the literature refers to the fact that the sharing and exchange of information via the tools of ICT can contribute to improved safety of the care and quality of care ; that to the extent that these tools enable healthcare professionals to have access to relevant information at all times and where they practice [6], [7]. For organizations, ICTs promise to improve their performance and efficiency, allowing them to manage, share and store large amounts of data [8], [9]. The industrialized countries have invested in ICT as a strategic choice to try to remedy the problem of access to care and the lack and / or the geographical maldistribution of health professionals on their vast territories. Indeed, with the transformations of their health systems, these countries have adopted several socio-health strategies and politico-economic converging towards the establishment of information highways and telemedicine. In fact, most no reform or health policy can be designed without taking into account the role that ICT can play, including telehealth, they appear as one of the essential conditions for the success of the reorganization [10]. Telemedicine seems like a lever for transformation of the health system.

Indeed, it is seen as a way to address the issue of territorial and economic disparities in access to care, because it eliminates geographical barriers while contributing to the emergence of another way of thinking about medical practice and offer of care [2], [11], [12].

Telemedicine is a new tool that gives very attractive to people in remote areas access to medical care and advanced technology from the rest of the population. If he no longer has to travel far for treatment, the patient is more likely to first consult the doctor [13], because, as we know early detection is crucial in many cases.

Simply, telemedicine is remote exchange of information and medical care. It has multiple uses. Essentially, it can be used for all aspects of medicine: prevention, diagnosis, treatment, education and research especially in sub specialties such as ophthalmology, dermatology, pathology and radiology where image is key to differential diagnosis.

In ophthalmology, we use the medicine in recent years in various ways. There was, in Europe, several research projects including the real-time video conferencing, visual electrodiagnostic, transmission of color photographs of the fundus and transmission of videotapes for evaluating the position and mobility of the eye. This technology allows specialists around the world to assess a case and discuss the treatment options field. This is particularly worthwhile when you have doubts about the diagnosis: several specialists in the world can then work together to examine the ambiguous images ophthalmology.

In 2001, Niewenhof et al. [14] concluded data they published on the use of telemedicine as a diagnostic tool for eye diseases that the new technology was an affordable and reliable instrument especially when the images transmitter knows pathologies in question.

In Canada, the Northwest Territories and Alberta are developing a teleophthalmology project that will use the new technology to detect early signs of retinal complications in diabetic natives. The technicians will train the regional health service of Stanton in Yellowknife to make annual retinal screening tests using portable digital devices. They will visit remote communities in turn, will test and transmit information in Alberta for diagnostic purposes. Whereas a native of ten suffers from diabetes may lead to retinopathy and blindness often, the project will significantly affect the everyday life not only for patients but also their families and the entire community [15].

Other challenges that will face telemedicine and tele-ophthalmology include the natural resistance of many health professionals and patients to changes brought by technological developments and the current limitations of robotic telesurgery [16]. Ideally, the inevitable growth of the use of telemedicine will benefit a larger number of patients regardless of their geographical, social and economic situation.

Diabetic retinopathy is a public health problem worldwide. The latest estimates indicate that 382 million people or 8.3% of adults have diabetes worldwide and this number will exceed 592 million in less than 25 years [17]. Almost 80% of people are in developing countries and nearly half of patients have some degree of diabetic retinopathy at some point. Diabetic retinopathy (DR), microvascular complication of diabetes, remains the leading cause of blindness and visual impairment in subjects under 60 years in all industrialized countries [18]. Indeed, R & D is a silent disease for many years; it only becomes symptomatic stages of complications. Without treatment, it is causing a visual impairment up to blindness or the interests of early detection.

Diabetic retinopathy (DR) remains a cause too frequent visual impairment in industrialized countries, especially in the labor force. Its severity remains linked to its belated support because of inadequate screening [19].

Originally, the mobile phone is supposed to allow its user to call relatively anywhere. But in recent years, the market for feature phones - and smartphones - has gained ground and is necessary in our pockets or handbags. Under Android, iOS or BlackBerry, these devices are now almost forget the main function which is to them a call.

The arrival on the market of this new class of powerful and affordable terminals is the main vector of several innovations in various fields, particularly in the medical field.

The increased use of medical applications for smartphones offers new opportunities for doctors, including ophthalmologists.

The smartphone market is increasing worldwide in recent years. It is estimated that over one billion the number of outstanding smartphones in the world [20]. Recent series, 74-98% of doctors own a smartphone and this percentage is probably going to increase in the coming years [21], [22], [23].

The arrival of smartphones and tablets in our daily practice sounds by the force of circumstances, and often unconsciously on how we handle our patients. The use of medical applications in smartphones attracting more and more interest. It changes the way in which we approach the care but also opens up new management perspectives [24], [25], [26], [27].

With its exponential success, the Smartphone has quickly become one of the essential links in the chain of telemedicine. This miniature tool with many sensors that can connect to almost all networks (Internet, GSM, 3G ...), is now seen as being ideal for different types of applications [28].

One of the most important preferential application areas that can be identified is the monitoring of chronic diseases such as heart disease, Alzheimer's, diabetes, etc.

Among chronic diseases, our choice fell on the disease of diabetes, one of the most prevalent diseases in the world today which affects a very significant portion of the world population, which is described by some as "the disease of the XXI century".

The remainder of this paper is organized as follows: in Section II, basic concepts related to the image processing are presented such as filtering, morphological operators. The section III discusses all steps behind the realization of the application. Conclusion is presented in the last section.

II. Application Description

The "RETINA" application helps fulfill the principle of remote tele-ophthalmology to allow early detection of diabetic retinopathy. It consists of two parts, one dedicated for patients and the other for doctors.

The first allows through a microscopic lens, which is connected to a smartphone, capturing the fundus of diabetic patients that may have a visual defect because of diabetes including diabetic retinopathy. The application then provides multiple choice of image processing (filters and morphological operators ...) to improve the quality of the captured fundus and thus facilitate the detection of these anomalies. Personal information of patients with fundus images are then processed in a secure MySQL database.

This application also provides information on the diabetic retinopathy, and its clinical signs. The second application dedicated to doctors allows to search in the database the processed images of the fundus of patients according to their ID numbers, display them on the interface to be viewed by the doctor, who will then give its diagnostic and recommendations to their patients by email.

The following section provides more informations about the different features that characterize our application.

A. Connection Establishment

In a first step, an interconnection between the mobile terminal and the server should be performed. It will therefore agree on the same tunnel (port, address, etc.) between this device and its corresponding so they can share data. About this connection, we have exploited the http protocol to manage the inputs/outputs of the server, including the personal information of patients.

B. Image Acquisition

After recording information related to the patient, the application gives us access to the smartphone's camera to capture the fundus. For this we use a microscopic lens with a universal clip, with a 60x and 100x zoom. But the result was not as reliable as expected so we used the e_ophtha_MA database to have retinal images. It was extracted from the telemedicine network OPHDIAT for screening of DR. This part does not require a particular study, it is only the library of Android that allows such manipulation.

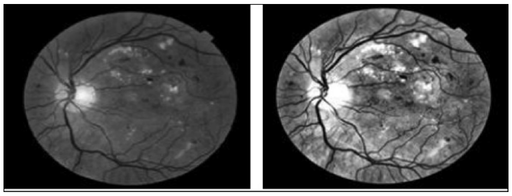


Fig. 1. Image before improvement (left) and improved by adaptive equalization (right)

C. Image Processing

Despite enormous progress in the image acquisition techniques, the retinal images taken at the clinic daily are often very noisy, they suffer from low contrast and illumination is not uniform. The reasons may be of varied origins:

- Possible diseases (such as cataracts).
- patient's movements.
- Circumstances in which the picture is taken.
- Differences in the illumination of the eye, which depend not only on technology but also on the shape of the patient's eye.

Apply image enhancement algorithms, i.e. the contrast enhancement or filtering to remove noise. These algorithms have two possible purposes:

1) On one hand, this type of algorithm is used to facilitate the work of a specialist by giving it improved images to analyze the particular algorithms to increase contrast.
2) On the other side, all methods of automatic analysis or semiautomatic begin with a pre-filtering of the image, therefore, improvement can be seen as a first step toward automatic analysis of retinal images.

In our application, with the library "OpenCV", we were able to integrate pretreatment methods, and morphological operators, which we will define later.

1) Pretreatment:

a) Adaptive histogram equalization -: To account for all the spatial distribution of gray levels in the image, a histogram adaptive equalization is used as it is depicted in Figure 1. This technology is to divide the image into non-overlapping rectangular regions.

b) Filtering -: Filtering is a basic operation in image processing, it allows the perception of certain details, reduce noise, compose the defects of the sensor. One of the filters used in image processing of our application is the Gaussian filter. This linear filter better handles cases or vessels appear darker than the background. Figure 2 shows an example of filtering with Gaussian filter.

$$g = \gamma^{\beta}\left(f * G\right) \tag{1}$$

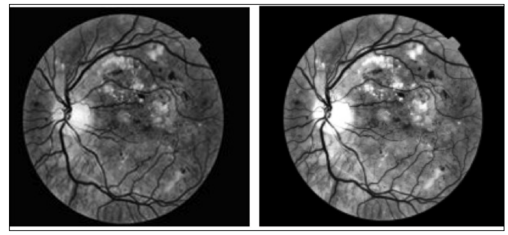


Fig. 2. Retinian image filtering (a) Improved image (b) Filtering with gaussian of the improved image

c) The Gaussian filter -: The intensity of a pixel depends on that of its neighbors: Let U (x), the gray level of a point x of the image to be processed and G the Gaussian kind given by the following formula.

$$G_{\sigma} = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{\mid x \mid}{2\sigma^2}\right) \tag{2}$$

2) Morphological operators: Mathematical morphology offers a large number of very powerful image processing and analysis tools that can be found under different headings in some image analysis software and even image editing, whose purpose is the study of objects according to their neighborhood, their texture and their grayscale or their color by the transformations it proposes. Mathematical morphology was invented in 1964 by GEORGES MATHEPON and JEAN SERRA in the laboratories of Paris Mines School. To analyze images using mathematical morphology, we rely on a certain number of operators that we specify and combine to arrive at the desired result. Morphological operators are divided into two types:

- basic morphological operators: erosion and dilation.
- compound morphological operators: they are formed by a successive application of set operators that the union, the difference.

The basic principle of mathematical morphology is to compare the image to be analyzed with a set of known geometry called structuring element.

a) Morphologic erosion: The morphological erosion of an object X by the structuring element B is defined by the principle of duality:

$$\varepsilon_E\left(\overline{X}\right) = \overline{\delta_E\left(X\right)} \tag{3}$$

The operation of the erosion of a set X by a structuring element B, is the set of points x of R^2 such that B is fully included in X when x is centered in x.

$$X \ominus B_x = \{x \in R^2, B_x \subset X\} \tag{4}$$

b) Morphologic dilation: The dilation operation denoted $\delta B(X)$ is in fact the dual operation of erosion. We obtain the dilated of a set X by a structuring element B taking the totality of the covered surface when the origin of B successively occupies all the positions $x \in X$. Since erosion and dilation are elementary morphological operators, it comes

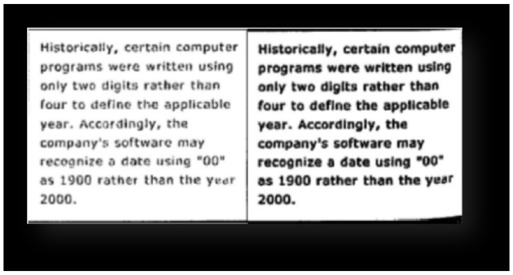


Fig. 3. Example of dilation effect

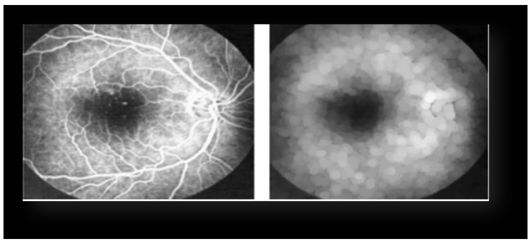


Fig. 4. Morphological opening

to mind to combine these operations; these transformations tend to smooth the contours of the particles. An example of dilation effect is shown by Figure 3.

c) Morphological opening: The morphological opening operation is obtained by the succession of erosion and dilation. The morphological opening has the effect of eliminating the detection zones of size smaller than that of the structuring element and reconstructing the shape of the object. The open set is more regular and less rich in detail than the initial set. The opening can disconnect the sets; it plays the role of a filtering.

$$\gamma_B(X) = X \circ B = \delta_{\tilde{B}}(\varepsilon_B(X)) = \left(X \ominus \tilde{B}\right) \oplus B \quad (5)$$

Figure 4 shows the morphological opening operation.

d) Morphological closing: By reversing the order of the operations used to define the opening, we obtain a new operation called closing. This operation is performed by concatenation of a dilation followed by erosion of the structuring element. The morphological closing makes it possible to fill the holes in the image, it can connect two related particles to become one.

$$\varphi_B(X) = X \cdot B = \varepsilon_{\tilde{B}}(\delta_B(X)) = \left(X \oplus \tilde{B}\right) \ominus B \quad (6)$$

Figure 5 shows an example of morphological closing operation.

e) Top hat transformation: Top hats are used to locate structures in the image that are smaller than the structuring element. They are usually applied only to grayscale images (see Figure 6). So, it allows to extract all the details of an image that have removed by filtering (opening or closing). The top hat transformation can be divided into:

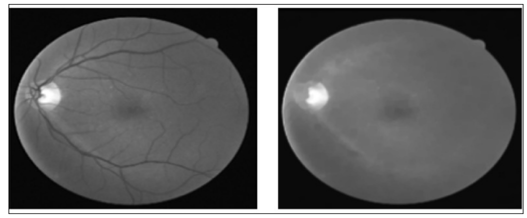


Fig. 5. Morphological closing

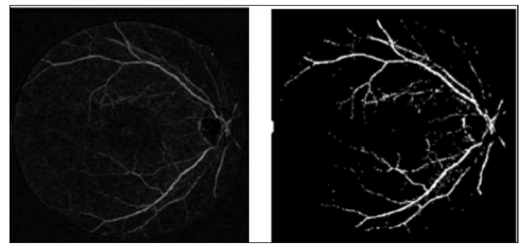


Fig. 6. Application of the top hat transformation

- "White top hat": to detect what the opening has caused, it is calculated by the difference between the image and its opening.

$$TH_B^+(f) = f - \gamma_B(f) \quad (7)$$

- "Black top hat": allows you to extract valleys or dark structures from the image. It is calculated by the difference between the closing and the image.

$$TH_B^-(f) = \phi_B(f) - f \quad (8)$$

III. Realization of the Application

The pltaform of the application is represented by the Figure 7 where the different parts and the possible transitions between them are showed. This application is done under Android environment using the Open source software library Computer Vision (OpenCV). OpenCV has more than 2500 optimized algorithms. The library includes a complete set of computer vision algorithms and learning algorithms. It has been around for a decade and is released under the Berkeley Software Distribution (BSD) license, making it easy for users to use and modify the code. OpenCV for Android supports access to its functions through its native API and also its Java wrappers API. In the case of a native API, you will define your native library using Android NDK and include the OpenCV libraries that you are using. Then, you will call your native library from the Java code using Java Native Interface (JNI). The other option is to use the OpenCV Java wrappers directly in your Java code using the usual Java imports. What will happen is that the Java wrappers will channel your calls to the native OpenCV libraries using JNI [29].

Our application consists of two parts, a part dedicated to the patients named "RETINA PATIENT" and another dedicated to the doctor named "RETINA DOCTOR".

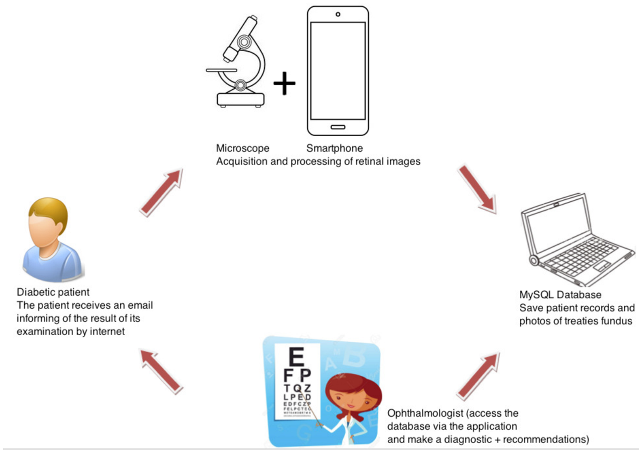


Fig. 7. System architecture

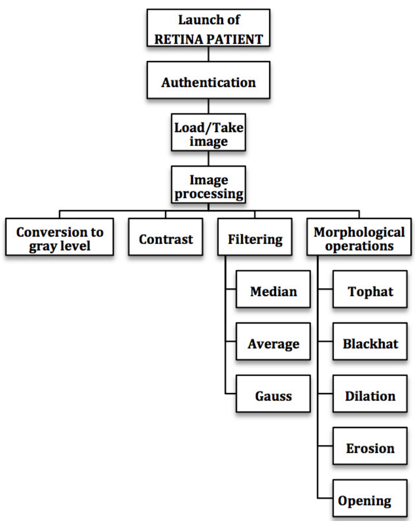


Fig. 8. RETINA PATIENT flowchart

A. RETINA PATIENT

A flowchart of RETINA PATIENT is depicted in Figure 8. When launching the application, a home page is displayed, followed by an authentication phase asking the user for his login and password (see Figure 9). This step allows a single authentication of users of the application, which allows a RETINA system protection of malicious uses. If the authentication is wrong i.e. login or the password is incorrect, an error message will be displayed to the user to try again. If authentication is successful, a window appears with four buttons labeled "NOUVEAU PATIENT", "DOSSIER PATIENT", "RETINOPATHIE DIABETIQUE" and "RETOUR" (see Figure 10). The "DOSSIER PATIENT" button leads to a personal form where the user must fill in all the necessary details of the patient to be examined.

Once all fields are filled, clicking on the "Ok" button creates a new line in the MySQL database, so a message informs us that the new patient has been successfully inserted

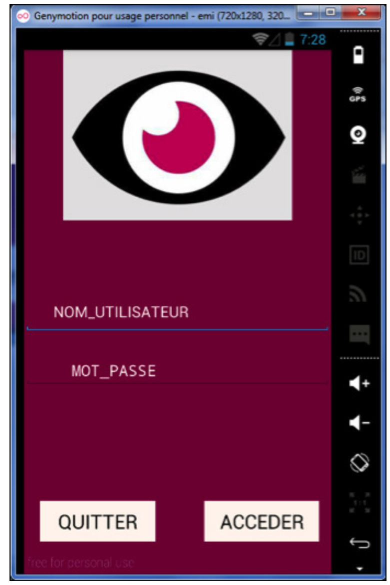


Fig. 9. Authentication page

into the database. The "RETOUR" button is used to return to the main menu, and thus cancel the registration of this patient.

After the patient registration phase, a menu is displayed, and gives us access to the smartphone camera, by pressing a button labeled "CAMERA" to take pictures of the fundus to the using the Smartphone camera or the small microscope connected to our Smartphone if the camera of the Smartphone has a low resolution. We can also load the fundus images if they are stored in the gallery (see Figure 11). After the selection, the image appears and it will be ready for processing. The latter makes it possible to: i) convert the image taken by the camera or loaded from the Smartphone gallery to gray level in order to apply other treatments, ii) improve the contrast, iii) select a filter among the three (Median, Average and Gauss), iv) or apply a morphological operator among the five (Tophat, Blackhat, Dilation, Erosion and Opening). All of these methods have been integrated through the "Opencv" library for Android.

From the button "DOSSIER PATIENT" of the main menu of RETINA, we have the possibility to change or modify the information's that are already registered in the database of our patients. These changes can be made after a file search by number to change either the patient's personal information or to update his fundus photo. A message informs us that the changes made have been saved successfully. The button "RETINOPATHIE DIABETIQUE" of the main menu, gives us some information about the disease, its definition, its symptoms and its causes. The "QUITTER" button allows us to leave the application in a secure way, and this after the confirmation message "Are you sure you want to leave?"

B. RETINA DOCTOR

The principle of RETINA DOCTOR application is described by the Figure 12. It allows the ophthalmologist to

Fig. 10. Home page application

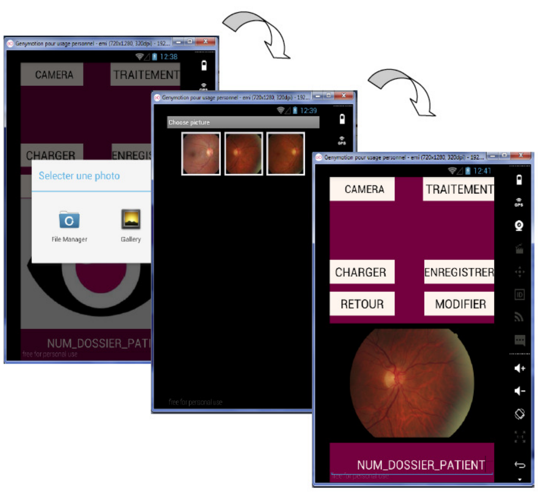


Fig. 11. Load of fundus photo

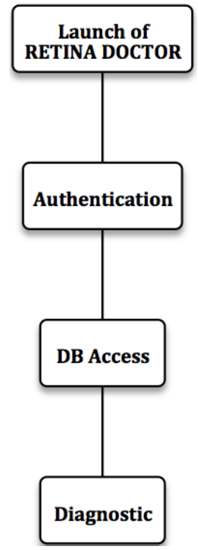


Fig. 12. RETINA DOCTOR flowchart

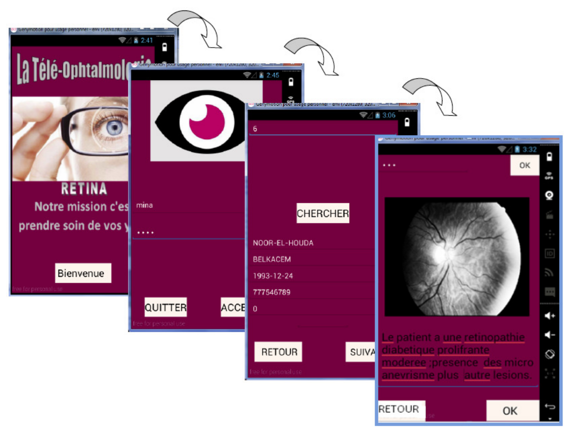


Fig. 13. Retina Doctor

access the database, after a step of personal authentication, and have the opportunity to see the coordinates and the fundus photos of his patients, thanks to the file number, so to be able to make then its diagnosis, and its recommendations (see Figure 13). If the entered folder number is not correct or does not exist in the database, an error message is displayed. If necessary, the patient will be summoned by the ophthalmologist via his mailbox. Finally, we note that the design and implementation of the application RETINA was made in collaboration with experts and professionals in the field of health including specialists in ophthalomology.

IV. Conclusion

Diabetic retinopathy is a chronic disease of the retina. As the number of patients with this condition increases each year, therefore, screening for diabetic retinopathy is currently a major problem. In this paper, we have developed an Android diagnostic aid application for scanning fundus images remotely using telemedicine tools and techniques, thus reducing the number of patients that need to be examined by ophthalmologists. In addition, improve the quality of care provided to patients by maximizing the number of diabetics examined per year and by reducing patient waiting times. The application is divided into two parts, a part dedicated to the patients allowing them to record their images of the fundus in a database after taking them or loading them from the gallery of their smartphones, and a second part dedicated to doctors who will be able to consult these images from a distance by accessing the databases for an eventual diagnosis.

References

[1] Callahan, Edward J., Hilty, Donald M., et Nesbitt, Thomas S. *Patient satisfaction with telemedicine consultation in primary care: comparison of ratings of medical and mental health applications*. Telemedicine Journal, 1998, vol. 4, no 4, p. 363-369.

[2] Potter, Andrew J., Mueller, Keith J., Mackinney, A. Clinton, et al. *Effect of tele-emergency services on recruitment and retention of US rural physicians*. Rural & Remote Health, 2014, vol. 14, no 3.

[3] Paré, Guy et Sicotte, Claude. Les technologies de l'information et la transformation de l'offre de soins. Cahier du GReSI, 2004, vol. 4, p.04-12.

[4] S. Yaya, C. Raffelini, *Technological transformations and evolution of the medical practice: current status, issues and perspectives for the development of telemedicine*, Revue medicale de Bruxelles 30 (2008), p.83-91.

[5] Johnsen, E., Breivik, E., Myrvang, R., et al. *Benefits from Telemedicine in Norway-An examination of available documentation*. HOYKOM report, 2006, vol. 1, p. 1-22.

[6] C. Kitamura, L. Zurawel-Balaura, R. Wong, *How effective is video consultationin clinical oncology? a systematic review*, Current Oncology 17 (2010) p.17-27.

[7] F. Fatehi, L. C. Gray, A. W. Russell, *Telemedicine for clinical management of diabetes: a process analysis of video consultations*, Journal of telemedicine and telecare 19 (2013) p.379-382.

[8] A. Bhattacherjee, N. Hikmet, N. Menachemi, V. O. Kayhan, R. G. Brooks, *The differential performance effects of healthcare information technology adoption*, Information Systems Management 24 (2006) p.5-14.

[9] N. Menachemi, A. Chukmaitov, C. Saunders, R. G. Brooks, *Hospital quality of care: does information technology matter? the relationship between information technology adoption and quality of care*, Health care management review 33 (2008) p.51-59.

[10] R. L. Bashshur, G. W. Shannon, E. A. Krupinski, J. Grigsby, J. C. Kvedar, R. S. Weinstein, J. H. Sanders, K. S. Rheuban, T. S. Nesbitt, D. C. Alverson, et al., *National telemedicine initiatives: essential to healthcare reform*, Telemedicine and e-Health 15 (2009) p.600-610.

[11] J.-P. Fortin, R. Landry, M.-P. Gagnon, J. Duplantie, R. Bergeron, Y. Galarneau, A. Cloutier, C. Houde, C. Poirier, *La télésanté : une stratégie pour soutenir la pratique des médecins en régions éloignées*, Rapport de recherche (2006).

[12] A. Darkins, *Patient safety considerations in developing large telehealth networks*, Clinical Risk 18 (2012) p.90-94.

[13] H. K. Li, *Telemedicine and ophthalmology*, Survey of ophthalmology 44 (1999) p.61-72.

[14] J. Nieuwenhof, A. Figueiredo, G. Lima, F. Rezende, C. Correia, M. Burnier, *The use of telemedicine as a diagnostic tool in ophthalmic pathology*., in: INVESTIGATIVE OPHTHALMOLOGY & VISUAL SCIENCE, volume 42, ASSOC RESEARCH VISION OPHTHALMOLOGY INC 9650 ROCKVILLE PIKE, BETHESDA, MD 20814-3998 USA, p.S797-S797.

[15] M. N. Burnier, *La télémédecine et la télé-ophtalmologie: 2e partie*, Canadian Journal of Ophthalmology 38 (2003) p.541-542.

[16] J. Marescaux, J. Leroy, M. Gagner, F. Rubino, D. Mutter, M. Vix, S. E. Butner, M. K. Smith, *Transatlantic robot-assisted telesurgery*, Nature 413 (2001) p.379-380.

[17] F. Aguiree, A. Brown, N. H. Cho, G. Dahlquist, S. Dodd, T. Dunning, M. Hirst, C. Hwang, D. Magliano, C. Patterson, et al., I*df diabetes atlas*, Nature (2013).

[18] J. Ding, T. Y. Wong, *Current epidemiology of diabetic retinopathy and diabetic macular edema*, Current diabetes reports 12 (2012) p.346-354.

[19] P. Massin, *Innovations thérapeutiques dans la rétinopathie diabétique*, Journal Français d'Ophtalmologie 34 (2011) p.491-497.

[20] S. Bicheno, Global smartphone installed base forecast by operating system for 88 countries: 2007 to 2017, Strategy Analytics (2012) p.1-65.

[21] K. F. B. Payne, H. Wharrad, K. Watts, *Smartphone and medical related app use among medical students and junior doctors in the united kingdom (uk): a regional survey*, BMC medical informatics and decision making 12 (2012) 1.

[22] R. J. Katz-Sidlow, A. Ludwig, S. Miller, R. Sidlow, *Smartphone use during inpatient attending rounds: prevalence, patterns and potential for distraction*, Journal of hospital medicine 7 (2012) p.595-599.

[23] N. Smart, *A survey of smartphone and tablet computer use by colorectal surgeons in the uk and continental europe*, Colorectal Disease 14 (2012) p.535-538.

[24] G. L. Warnock, *The use of apps in surgery*, Can J Surg 55 (2012) 77.

[25] B. M. Dala-Ali, M. A. Lloyd, Y. Al-Abed, *The uses of the iphone for surgeons*, The surgeon 9 (2011) p.44-48.

[26] A. S. M. Mosa, I. Yoo, L. Sheets, *A systematic review of healthcare applications for smartphones*, BMC medical informatics and decision making 12 (2012) 1.

[27] R. Wu, P. Rossos, S. Quan, S. Reeves, V. Lo, B. Wong, M. Cheung, D. Morra, *An evaluation of the use of smartphones to communicate between clinicians: a mixed-methods study*, Journal of medical Internet research 13 (2011) p.59.

[28] N. P. Terry, *Mobile health: assessing the barriers*, CHEST Journal 147 (2015) p.1429-1434.

[29] A. Muhammad, *OpenCV Android Programming By Example*, Packt Publishing Ltd, 2015.

Virtual Rehabilitation Using Sequential Learning Algorithms

Gladys Calle Condori
School of Systems Engineering
Universidad Nacional de San Agustín
Arequipa, Perú

Eveling Castro-Gutierrez
School of Systems Engineering and CiTeSoft
Universidad Nacional de San Agustín
Arequipa, Perú

Luis Alfaro Casas
School of Systems Engineering
Universidad Nacional de San Agustín
Arequipa, Perú

***Abstract*—Rehabilitation systems are becoming more important now because patients can access motor skills recovery treatment from home, reducing the limitations of time, space and cost of treatment in a medical facility. Traditional rehabilitation systems served as movement guides, later as movement mirrors, and in recent years research has sought to generate feedback messages to the patient based on the evaluation of his or her movements. Currently the most commonly used algorithms for exercise evaluation are Dynamic time warping (DTW), Hidden Markov model (HMM), Support vector machine (SVM). However, the larger the set of exercises to be evaluated, the less accurate the recognition becomes, generating confusion between exercises that have similar posture descriptors. This research paper compares two HMM classifiers and Hidden Conditional Random Fields (HCRF) plus two types of posture descriptors, based on points and based on angles. Point representation proves to be superior to angle representation, although the latter is still acceptable. Similar results are found in HCRF and HMM.**

***Keywords*—Kinect Skeletal; Sequential Learning Algoritms; Virtual Rehabilitation; Virtual Reality Therapy**

I. Introduction

Rehabilitation systems are now becoming more important because patients can access motor skills recovery treatment from home. The treatment has several components, one of the main ones being the repetition of movements, which requires the assistance of medical personnel to indicate if the movement was performed correctly to count the repetitions performed. It is necessary to go to a medical center, request an appointment and require the assistance of a therapist.

Virtual rehabilitation systems are intended to meet the needs of the mechanical part of the motor skills recovery treatment[14]. Initially these rehabilitation systems were focused only on being a guide of movements, since they showed an avatar that carried out the example of the movement to be carried out. Later, with the introduction of cheaper motion sensors such as Kinect(Kv1) in 2010, systems were built that served as a mirror, i.e. showed the user their movements on the screen so that the patient could visualize and self-correct or simply have a record that could be evaluated later by the therapist. In recent years, research has sought to generate intelligent assistants such as motion counters or motion evaluators who also send feedback messages in the form of text or voice to inform patients of the quality of their movements. One way to evaluate movements is to apply a sequential learning algorithm to position descriptors obtained by a depth sensor such as Kv1 [1], [2], [3].

The most commonly used algorithms for motion evaluation are DTW, HMM and SVM[1], [2], [3]. However, the greater the number of exercises to be evaluated, the less accurate the recognition of these movements is, which leads to the search for improvements in the performance of the algorithm used.

This research paper aims to compare two HMM and HCRF classifiers, in addition to two types of posture descriptors, based on points and based on angles.

The point based posture descriptor proves to be superior to the angle based posture descriptor, although the latter is still acceptable. While with the HCRF and HMM algorithms similar results are found. The remaining part of the paper is organized as follows. A brief review of virtual rehabilitation state-of-the-art in this paper is explained in Section 2. Details of the methodology are described in Section 3. A results with the experiment settings is introduced in Section 4. Conclusions are presented in Section 5 and some future Works are provided in Section 6.

II. State-of-the-Art

In recent years there has been an increase in the number of studies related to virtual rehabilitation with feedback [13], however, there are few studies that test new algorithms applied to exercise evaluation, as shown below.

Uttarwar et al. [4], propose a rehabilitation system for shoulder injuries. They use HMM for recognition and histograms to calculate accuracy. For the training, 10 sequences of each exercise were used, performed by 3 healthy subjects for 4 exercises. 100% accuracy is reported when the patient performs a single exercise for 30 seconds. For multiple exercises, the ranking gets a lower accuracy score . The training and testing package is very limited with respect to other jobs [5], [6], [7], [2] and [8].

M. Capecci et al. [2], propose the use of the Hidden semi-Markov model (HSMM) to model the time evolution of movements and compare them against DTW, perform tests with 5 different exercises with 33 people. As statistical models, HSMMs model the distribution of features from multiple demonstrations. They report that the proposed algorithm outperforms DTW in terms of correlation with a clinical evaluation, demonstrating the applicability of this approach. The number of subjects is larger than in other jobs [5], [6], [7], and [8]. However, the number of exercises is less than these.

Anton et al. [9], propose a system for monitoring rehabilitation exercises to track patient progress, without the need for a specialist. The system can be trained to detect exercise and compare movement with the correct form, providing feedback to the patient and tracking progress. Of 100 sequences of skeletons describing simple activities such as hand or leg movements, 97% were correctly recognized. However, evaluation errors can be expected at any time, propagated from the error of the MATCH algorithm. In addition, errors based on Kinect's schematic error may occur for the detected attachment points. Another disadvantage of the system is that it only considers the angles of the segment of the skeleton in two dimensions when projecting the segments in the XY plane, without taking into account the perspective [9]. An evaluation with Kinect v2 may yield greater results. Descriptors like in [6], [7] allow you to represent perspective.

Vemulapalli et al. [7], propose a new skeletal representation that explicitly models 3D geometric relationships between various parts of the body using rotations and translations in 3D space. Using the proposed representation, human actions can be modeled as curves in this Lie group. They then perform the classification using a combination of DTW, Fourier temporal pyramid representation and linear SVM. Experimental results in three sets of action data show that the proposed representation works better than many existing skeletal representations. The descriptors used, besides being light, obtain results higher than 90%, being necessary to use them with other algorithms such as those of [10].

Batabyal et al.[5], demonstrate that a set of active 3D skeleton coordinates can be used effectively for action recognition. The 3D joint coordinates based on Kinect suffer from low frequency noise. However, the covariance-based function could successfully eliminate the unwanted effect of noise. At the same time, mapping characteristics to the lowest dimensional variety may improve the result of the classification. The 3D descriptor they use proves to be meaningful with a large data set.

Wang et al.[10], present a discriminatory hidden state approach to gesture recognition. The proposed model combines the two main advantages of current approaches to gesture recognition: the ability of CRFs to use long-range dependencies and the ability of HMMs to model the latent structure. Results have shown that HCRFs outperform both CRFs and HMMs for certain gesture recognition tasks. For arm gestures, the multi-class HCRF model outperforms the HMM and CRF even when not using long-range dependencies, demonstrating the advantages of discriminative joint learning. HCRF and CRF are used for the recognition of gestures, while still needing to be tested with exercises.

Liu et al.[11], present an assistive physical rehabilitation system based on skeletal detection with Kinect. They construct a location of standardized three-dimensional Cartesian coordinates of correct postures in the OpenNI system. They also use the support vector machine (SVM) as a classifier to define the accuracy of the posture. Finally, the system can judge the correctness of the users' positions. Considering only 15 joints, and leaving aside the less significant ones, seems to have a favorable result when recognizing gestures.

From the previous analysis it is observed that algorithms such as HCRF still need to be tested with movements particularly of exercises, and HMMs tested with a wider set of movements, in addition there are several descriptors that could significantly represent the user's posture, which should be tested with algorithms such as HMM and HCRF.

In this paper we applied HMM described in [4] by Utarwar et al., and HCRF in [10] by Wang et al. Using the position descriptor based on points [5] and the position descriptor based on angles by Anton et al.[9], considering three dimensions XYZ, described in[7] by Vemulapalli et al. It was therefore necessary to assess whether the position descriptor based on angles is as significant as the position descriptor based on points with two different HMM and HCRF algorithms.

III. Methodology

We conducted the research in four stages, as shown in Figure 1: Stage 1: Data set collection using Kv1, which was divided into two subsets: training set and test set. In the second stage, the training set was trained in four situations: HCRF with point based descriptor, HCRF with angle based descriptor, HMM with point based descriptor and HMM with angle based descriptor. For the third stage, the test set was evaluated in all four cases, and contrasted with the expected result. In the final stage, the confusion matrix was developed for each situation and the ROC curve was plotted for the four test cases.

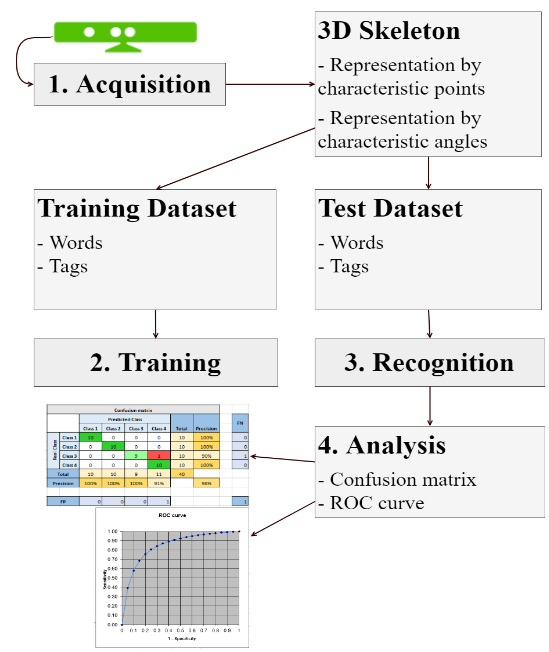


Fig. 1. Step 1: Collect the data set using the Kinect v1 (training subset and test subset). Stage 2: Trained the training subset with four algorithm cases: HCRF with points, HCRF with angles, HMM with points and HMM with angles. Stage 3: Evaluated the test subset with all four cases. Stage 4: Analysis using the confounding matrix and the ROC curve.

A. Data acquisition

The Skeletal tracking functionality of the kv1 allows the human skeleton to be tracked using an algorithm that identifies the parts of the human body of people within the sensor's field of view. It is by means of this algorithm that the points referring to the parts of the human body have been obtained. And through these reference points it has been possible to apply different algorithms for the recognition of gestures.

In other words, each sequence of movements corresponds to a sequence of postures, each posture is represented by twenty points that refer to the joints of the human body, each point has three coordinates, in total each posture has sixty values that describe it. And each movement sequence has a multiple of sixty for its representation.

There are two ways to represent these reference points, using point based descriptors and angle based descriptors.

1) Pose descriptor based on points: Each skeleton is represented by twenty joints, each joint has a pre-defined enumeration and identification, listed in table I, which are distributed as a sample in Figure 2, and each point is represented in space 3 D, by the coordinates X, Y and Z (1), as in [5], [7].

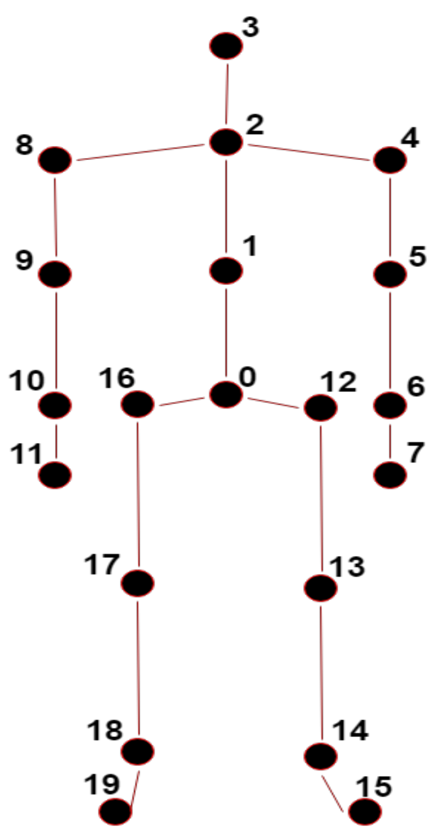


Fig. 2. Representation of the posture using characteristic points. The skeleton is made up of 20 joints, with a predefined enumeration and identification.

Source: Kinect

$$KinectA = A(x_1, y_1, z_1) \tag{1}$$

2) Pose descriptor based on angles: Eighteen angles formed by the reference points of the human body were identified, Figure 3, four angles are excluded: the angles formed by the hands and feet with the extremities, since the observations are made on movement exercises involving movements of the extremities of the body, the minimum movements of the hands and feet being of little relevance, finally fourteen angles of importance are selected, table II

TABLE I. CHARACTERISTIC POINTS

#	Joint
0	HipCenter
1	Spine
2	ShoulderCenter
3	Head
4	ShoulderLeft
5	ElbowLeft
6	WristLeft
7	HandLeft
8	ShoulderRight
9	ElbowRight
10	WristRight
11	HandRight
12	HipLeft
13	KneeLeft
14	AnkleLeft
15	FoootLeft
16	HipRight
17	KneeRight
18	AnkleRight
19	FootRight

Source: Kinect

TABLE II. CHARACTERISTIC ANGLES. ANGLE EXCLUDED.

#	Angles
0	hipCenter - spine - shoulderCenter
1	spine - shoulderCenter - head
2	head - shoulderCenter - shoulderLeft
3	shoulderCenter - shoulderLeft - elbowLeft
4	shoulderLeft - elbowLeft - wristRight
5	elbowLeft - wristLeft - handLeft *
6	head - shoulderCenter - shoulderRight
7	shoulderCenter - shoulderRight - elbowRight
8	shoulderRight - elbowRight - wristRight
9	elbowRight - wristRight - handRight *
10	spine - hipCenter - hipLeft
11	hipCenter - hipLeft - kneeLeft
12	hipLeft - kneeLeft - ankleLeft
13	kneeLeft - ankleLeft - footLeft *
14	spine - hipCenter - hipRight
15	hipCenter - hipRight - kneeRight
16	hipRight - kneeRight - ankleRight
17	kneeRight - ankleRight - footRight *

To calculate the fourteen selected angles, place the three points forming the angle, figure 4, convert these points into two vectors (2) and (3), and apply the angle formula between two vectors in three dimensions (4), as in [2].

$$\sim u = A(x_1, y_1, z_1) - B(x_2, y_2, z_2) \tag{2}$$

$$\sim v = B(x_2, y_2, z_2) - C(x_3, y_3, z_3) \tag{3}$$

$$\cos\beta = \frac{\underset{u}{\rightarrow} \cdot \underset{u}{\rightarrow}}{|\underset{u}{\rightarrow}| \cdot |\underset{u}{\rightarrow}|} \tag{4}$$

An angle of for this type of descriptors is formed from three points relating to the human posture, these points must be adjacent. The angle being the one that has as its center the joint that joins the other two joints.

TABLE III. TRAINING AND TEST DATA

Class	Exercise	Training samples	Test samples
Class 1	Diagonal	24	10
Class 2	Right Bipodal Equilibrium	24	10
Class 3	Left Bipodal Equilibrium	24	10
Class 4	Left Unipodal Equilibrium	24	10
Clase 5	Right Unipodal Equilibrium.	24	10
Clase 6	Right Elbow Flexion	24	10
Clase 7	Left Elbow Flexion	24	10
Clase 8	Right Shoulder Flexion	24	10
Clase 9	Left Shoulder Flexion	24	10
Clase 10	Squat	24	10

TABLE IV. EXPERIMENTS

Algorythm	Tolerance	Repetitions
HMM Points	0,125 %	300
HMM Angles	0,125 %	300
HCRF Points	0,125 %	300
HCRF Angles	0,125 %	300

Excluded angles are shown with an asterisk in table II.

The point based descriptor was used to capture the exercise sequences. Ten exercises from a Patient and Caregiver Guide were identified for this purpose, these being exercises to be performed at home after a stroke. Recordings of 34 repetitions were made for each exercise, 24 were added to the training set and 10 to the test set. In total there were 240 training sequences and 100 test sequences, the description of the data set in Table III.

Exercises of the upper limbs. Self-passives and mobilizations made by the patient himself, Figures 5 and 6.

Trunk exercises. Mobilizations made by the patient himself. Figures 7 and 8.

B. Training

A set of 240 sequences of 10 types of exercises were trained, with two HCRF and HMM algorithms, using two different descriptors, based on points and based on characteristic angles, table IV

1) HMM: The parameters are learned automatically by means of the Baum Welch Learning algorithm, which is created with a Forward topology with 10 states for a 60-dimensional alphabet, in addition, the classifier is created assuming a Gaussian distribution and it is defined that the training must be carried out until the tolerance is less than 0.0025 or the maximum of iterations defined as 300 have been fulfilled. In addition, the covariance matrices are prevented from degenerating adding a regularization value to the diagonal so that they remain positive.

2) After defining the parameters, the algorithm is trained with the sequences destined for this purpose, table III.HCRF: The Hidden Resilient Gradient Learning algorithm is created, which is established with a Forward topology with 10 states for a 60 dimensional alphabet, in addition, the classifier is created assuming a Gaussian distribution, it is defined that the training must be performed until the tolerance is less than 0.0025 or the maximum of iterations defined as 300 have been fulfilled.

After defining the parameters, the algorithm is trained with the sequences destined for this purpose, Table III

TABLE V. RESULTS

Algorythm	Successes (%)	Training time (ms)	Consultation time (ms)
HMM Points	100 %	88898,74	3781,51
HMM Angles	96 %	30830,35	1489,99
HCRF Points	100 %	248139,86	4371,9
HCRF Angles	95 %	621144,23	2094,39

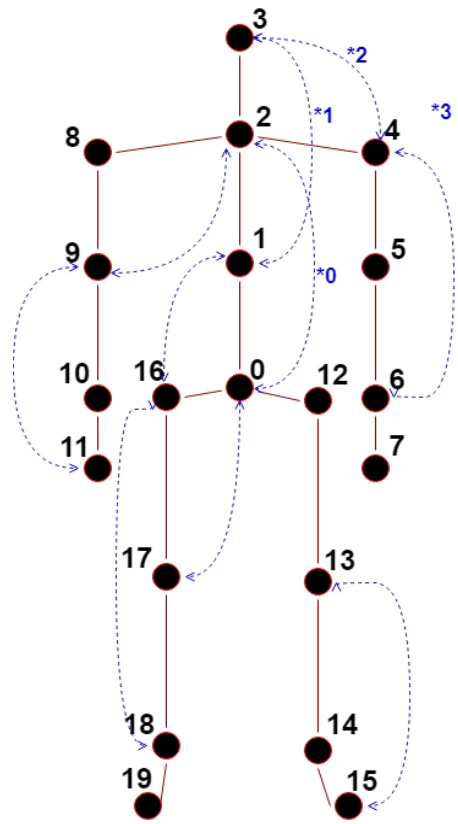


Fig. 3. Representation of the posture by characteristic angles. The skeleton is formed by 18 angles, excluding angles formed by the hands and feet. Finally, 14 angles are selected.

C. Recognition

For recognition in cases where point based descriptors are used, words and their corresponding tags are sent. In the case of angle-based recognition, the angle points are first converted and sent to the algorithm.

The set of training sequences has associated tags that identify the movement that is performed in that sequence, then 100 exercise sequences are consulted, and the answers generated are contrasted with their respective tags. In addition, the time required for the consultation is calculated.

IV. RESULTS

HCRF with point representation gets 100% accuracy, just like HMM with point representation. While the same algorithms with angle representation get 95% and 96% respectively. In the case of HCRF with angle representation, there is confusion between classes: class 1, class 4 and class 10, and in the case of HMM with angle representation the confusion is in classes: class 4 and class 5, table V.

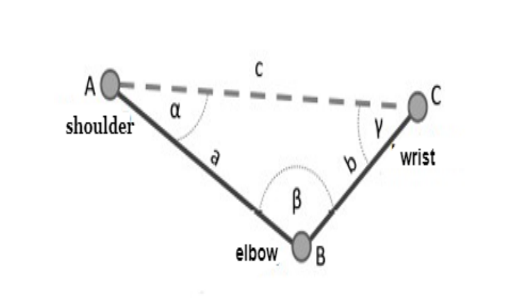


Fig. 4. Formation of a characteristic angle from three characteristic points of posture. A, B and C represent some adjacent characteristic point, beta the characteristic angle formed by A, B and C.

Source: Kinect.

Fig. 5. Exercise. Elbow flexion [12]

Fig. 6. Exercise. Shoulder flexion [12]

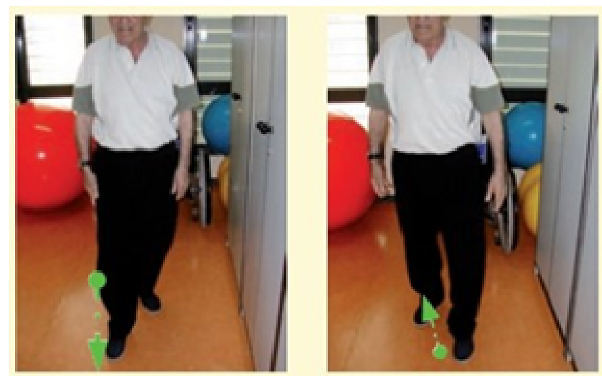


Fig. 7. Exercise. Bipodal balance[12]

HMM with point and angle representation are 88898.74 ms and 30830.35 ms respectively. While HCRF gets 248139.86 ms and 621144.23 ms for each representation respectively. Notably HMM requires less training time with either point or angle representation, table V.

The angle display with HMM gets 1489.99 ms and with HCRF 2094.39 ms respectively. While the dotted representation gets 3781.51 ms and 4371.9 ms for each algorithm respectively. The angle representation with both algorithms obtains a shorter response time than with its counterpart, Table V.

Of the 100 test sequences, 100 are correctly classified, Figure 9.

Of the 100 test sequences, 96 are correctly classified and 4 are incorrectly classified, figure 10.

Of the ten elements of class 9, 7 are correctly classified to class 9 and 3 are incorrectly classified to class 4, figure 10.

Of the ten elements of class 10, 9 are correctly classified to class 10 and 1 is incorrectly classified to class 5, figure 10.

Of the 100 test sequences, 100 are correctly classified, figure 11.

Of the 100 test sequences, 95 are correctly classified and 5 are erroneously classified, figure 12.

Of the ten elements of class 4, 8 are correctly classified to class 8 and 2 are incorrectly classified to class 1 and 10, figure 12.

Of the ten elements of class 9, 7 are correctly classified to class 9 and 3 are incorrectly classified to class 4, figure 12.

The ROC curve shows that the point based descriptor has more area under the curve with both algorithms, figure 13, so for the data set presented it is the most recommended to use.

V. Conclusion and Future Work

A total of 340 sequences of postures from 10 rehabilitation exercises were collected, 240 of which were for training and 100 for testing. These sequences were evaluated with two HCRF and HMM algorithms using two types of descriptors, based on posture points and based on posture angles. The dot-based descriptor was found to be superior to the angle-based descriptor, however, the latter is still acceptable. While HCRF and HMM have similar results, it is necessary to perform tests with a set of sequences superior to the present one.

With both HCRF and HMM, the point based descriptor proves to be more accurate, however the results with the angle based descriptor are still acceptable. Angle representation could be improved with more analysis of the relevance of angles to certain movements. HCRF and HMM show slightly similar results, being HMM the least space and time consuming, therefore it is still necessary to do experiments with a greater number of sequences and groups of exercises, to determine the most appropriate

Fig. 8. Exercise. Unipodal balance[12]

Confusion matrix													
		Predicted Class											
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10	Total	Precision
Real Class	Class 1	10	0	0	0	0	0	0	0	0	0	10	100%
	Class 2	0	10	0	0	0	0	0	0	0	0	10	100%
	Class 3	0	0	10	0	0	0	0	0	0	0	10	100%
	Class 4	0	0	0	10	0	0	0	0	0	0	10	100%
	Class 5	0	0	0	0	10	0	0	0	0	0	10	100%
	Class 6	0	0	0	0	0	10	0	0	0	0	10	100%
	Class 7	0	0	0	0	0	0	10	0	0	0	10	100%
	Class 8	0	0	0	0	0	0	0	10	0	0	10	100%
	Class 9	0	0	0	0	0	0	0	0	10	0	10	100%
	Class 10	0	0	0	0	0	0	0	0	0	10	10	100%
Total		10	10	10	10	10	10	10	10	10	10	100	
Precision		100%	100%	100%	100%	100%	100%	100%	100%	100%	100%		100%
FP		0	0	0	0	0	0	0	0	0	0		

Fig. 9. Confusion Matrix. HMM Characteristic Source Points

Confusion Matrix													
		Predicted Class											
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10	Total	Precision
Real Class	Class 1	10	0	0	0	0	0	0	0	0	0	10	100%
	Class 2	0	10	0	0	0	0	0	0	0	0	10	100%
	Class 3	0	0	10	0	0	0	0	0	0	0	10	100%
	Class 4	0	0	0	10	0	0	0	0	0	0	10	100%
	Class 5	0	0	0	0	10	0	0	0	0	0	10	100%
	Class 6	0	0	0	0	0	10	0	0	0	0	10	100%
	Class 7	0	0	0	0	0	0	10	0	0	0	10	100%
	Class 8	0	0	0	0	0	0	0	10	0	0	10	100%
	Class 9	0	0	0	3	0	0	0	0	7	0	10	70%
	Class 10	0	0	0	0	1	0	0	0	0	9	10	90%
Total		10	10	10	13	11	10	10	10	7	9	100	
Precision		100%	100%	77%	77%	91%	100%	100%	100%	100%	100%		96%
FP		0	0	0	3	1	0	0	0	0	0		

Fig. 10. Confusion Matrix. HMM Characteristic Angles

Confusion Matrix													
		Predicted Class											
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10	Total	Precision
Real Class	Class 1	10	0	0	0	0	0	0	0	0	0	10	100%
	Class 2	0	10	0	0	0	0	0	0	0	0	10	100%
	Class 3	0	0	10	0	0	0	0	0	0	0	10	100%
	Class 4	0	0	0	10	0	0	0	0	0	0	10	100%
	Class 5	0	0	0	0	10	0	0	0	0	0	10	100%
	Class 6	0	0	0	0	0	10	0	0	0	0	10	100%
	Class 7	0	0	0	0	0	0	10	0	0	0	10	100%
	Class 8	0	0	0	0	0	0	0	10	0	0	10	100%
	Class 9	0	0	0	0	0	0	0	0	10	0	10	100%
	Class 10	0	0	0	0	0	0	0	0	0	10	10	100%
Total		10	10	10	10	10	10	10	10	10	10	100	
Precision		100%	100%	100%	100%	100%	100%	100%	100%	100%	100%		100%
FP		0	0	0	0	0	0	0	0	0	0		

Fig. 11. Confusion Matrix. HMM Characteristic Angles

ACKNOWLEDGMENT

We would like to thank to the Research Center, Transfer of Technologies and Software Development R + D + i CiTeSoft-UNSA for their collaboration in the use of their equipment and facilities, for the development of this research work and Universidad Nacional de San Agustin de Arequipa - UNSA, and the collaboration of Brigitte A. Champi Paredes.

REFERENCES

[1] D. Anton, A. Goñi, and A. Illarramendi, *Exercise recognition for kinect-based telerehabilitation*, Methods of Information in Medicine, 2015.

[2] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyrki, S. Longhi, L. Romeo, and F. Verdini, *Physical rehabilitation exercises assessment based on Hidden Semi-Markov Model by Kinect v2*, 3rd ed. 3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016, pp. 256259, 2016.

[3] P. Parmar and B. T. Morris, *Measuring the quality of exercises*, 3rd ed. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2016-October, pp. 2241 2244, 2016.

[4] P. Uttarwar and D. Mishra, *Development of a kinect-based physical rehabilitation system*, Proceedings of 2015 3rd International Conference on Image Information Processing, ICIIP 2015, pp. 387392, 2016.

[5] T. Batabyal, T. Chattopadhyay, and D. P. Mukherjee, *Action recognition using joint coordinates of 3D skeleton data*, Proceedings - International Conference on Image Processing, ICIP, vol. 2015-Decem, pp. 4107 4111, 2015.

[6] O. C eliktutan, C. Akgul, C. Wolf, and B. Sankür, *Graph-based analysis of physical exercise actions*, Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Health Care, pp. 2331, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2505330

[7] R. Vemulapalli, F. Arrate, and R. Chellappa, *Human Action Recognition by Represention 3D Skeletons as Points in Lie Group*, 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 588595 , 2014.

[8] F. Deboeverie, S. Roegiers, G. Allebosch, P. Veelaert, and W. Philips, *Human gesture classification by brute-force machine learning for exergaming in physiotherapy*, IEEE Conference on Computatonal Intelligence and Games, CIG, 2017.

[9] S. Anton and F. Daniel Anton, *Monitoring rehabilitation exercises using MS Kinect*, in 2015 E-Health and Bioengineering Conference, EHB 2015, 2016.

[10] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell *Hidden Conditional Random Fields for Gesture Recognition*, 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR06, vol. 2, no. c, pp. 15211527, 2006. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1640937

[11] Tsun-Te Liu, Ching-Tang Hsieh, Ruei-Chi Chung, and Yuan -Sheng Wang, *Physical Rehabilitation Assistant System Based on Kinect*, Applied Mechanics and Materials, vol. 284-287, pp. 1686 1690, 2013. [Online]. Available: http://www.scientific.net/AMM.284- 287.1686

[12] C. Ruiz, D. Arce, *A guide for patients and caregivers. Exercises to do at home after a stroke*, —pp. 1 12, 2007.

[13] R. Adams, M. Lichter, A. Ellington, M. White, K. Armstead, J. Patrie, P. Diamond, *Virtual activities of daily living for recovery of upper extremity motor function*, —pp. 252 260, 2018.

[14] M. Chughtai, J. Newman, A. Sultan, A. Khlopas, S. Navarro, A.Bhave, MA. Mont *The Role of Virtual Rehabilitation in Total Knee and Hip Arthroplasty*, —vol. 32, 2018.

Confusion matrix													
		Predicted Class										Total	Precision
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10		
Real Class	Class 1	10	0	0	0	0	0	0	0	0	0	10	100%
	Class 2	0	10	0	0	0	0	0	0	0	0	10	100%
	Class 3	0	0	10	0	0	0	0	0	0	0	10	100%
	Class 4	1	0	0	8	0	0	0	0	0	1	10	80%
	Class 5	0	0	0	0	10	0	0	0	0	0	10	100%
	Class 6	0	0	0	0	0	10	0	0	0	0	10	100%
	Class 7	0	0	0	0	0	0	10	0	0	0	10	100%
	Class 8	0	0	0	0	0	0	0	10	0	0	10	100%
	Class 9	0	0	0	3	0	0	0	0	7	0	10	70%
	Class 10	0	0	0	0	0	0	0	0	0	10	10	100%
Total		11	10	10	11	10	10	10	10	7	11	100	
Precision		91%	100%	91%	73%	100%	100%	100%	100%	100%	91%		95%
FP		1	0	0	3	0	0	0	0	0	1		

Fig. 12. Confusion Matrix. HCRF Characteristic Angles

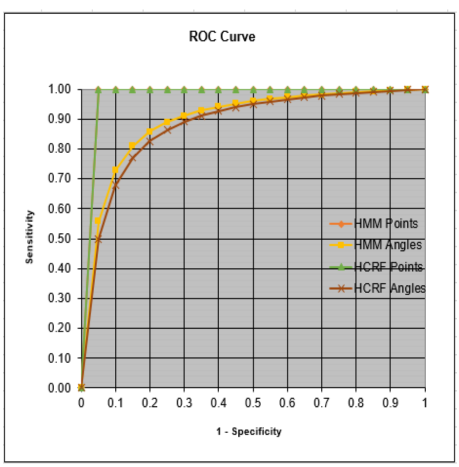


Fig. 13. ROC curve.HMM Characteristic Angles, HMM Characteristic Points, HCRF Characteristic Points, HCRF

A Machine Learning based Fine-Tuned and Stacked Model: Predictive Analysis on Cancer Dataset

Ravi Aavula
Guru Nanak Institutions Technical Campus,
Hyderabad, (T.S), India

R. Bhramaramba
GITAM Institute of Technology, GITAM,
Visakhapatnam, (A.P), India

***Abstract*—The earlier forecast and location of disease cells can be useful in curing the illness in medical applications. Knowledge discovery is having many significant roles in health sector, bioinformatics etc. Plenty of hidden information is available in the datasets present in the various domains like - medical information, textual analysis, image attributes exploration etc. Predictive analytics and modeling encompasses a variety of statistical methodologies from machine learning that can analyze the present along with historical facts to make the predictions about the future events. Breast cancer research already has involved with the good amount of progress in recent decade, but due to advancement in technologies, there is still some possibilities for an improvement.**

In this paper, the fine-tuned and stacked model procedure is presented which is experimented on standard breast cancer dataset. The obtained results show the improvement over state-of-the-art algorithms with improved performance parameters e.g. disease prediction accuracy, sensitivity and better F1 score etc.

Keywords*—*Machine learning; Cancer prediction; Data mining and Knowledge discovery; Supervised learning; Neural Networks

I. Introduction

Over past decades, the life-style of individuals has modified underneath the various factors i.e. changes in the nutrition thing, utilization of the artificial agents or varied technological enhancements. It incorporates a negative as well as positive side. On one aspect, life is easier; there are new opportunities to carry out totally diverse kind of human motions etc. On another aspect, various types of diseases have increased and that require new varieties of treatment phenomenons to enhance a high quality of health care and the all in one solutions for associated well-being. Breast cancer [6] is one in all the foremost common cancer in conjunction with respiratory organ and cartilaginous tube cancer, glandular cancer and carcinoma among others. The employment of information science and machine learning approaches [11][13] in medical fields proves to be prolific as such approaches are also thought of nice help within the higher cognitive process of medical practitioners. Conventional strategic movements are hardly tolerating while facing with the multivariate natured data. The earlier forecast and location of disease cells with better accuracy can be useful in curing the illness in various medical / healthcare applications.

A. Related Work

M. Bruijne given a machine learning approach and framework [1] for disease detection and diagnosis for medical and healthcare data. Sherafatian [2] in his work, given a concept, tree-based procedures which is indicated as the minimal length subset of miRNA for the disease diagnosis. H. Wang et al. [3] proposed an ensemble procedure based on support vector regression for the cancer disease diagnosis. C. M. Lynch et al., in their work [4], have proposed a procedure for lung cancer patient's predictive analysis and their survival scenario chances via the supervised ML methods. Sommen et al. [5] given a method for predictive features for early cancer detection scenario. Wassan et al. [6] surveyed the various machine learning methods in Bioinformatics domain. N. Khuriwal et al. [7] proposed a framework for breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. Ali et al. [8] compared two techniques i.e. SVM and Neural semantic circuit networks cancer contamination diagnosis. They did experiments with the variety of kernel functionalities for support vectors hyperplanes e.g. mpl (with 86% accuracy), radial basis function (89%), quadratic (88%) furthermore polynomial (approx 88%).

Shajahaan et al. [9] utilized "Wisconsin Breast Cancer Database" for contamination analysis within variety of learning rules like Random Forests (RF), C4.5 etc. Chaurasia et al. [10] employed WEKA environment with 10-folds cv procedure for the algorithms like - k-NN, Best First randomized nodes forests, SMO etc. Authors of [11] in their work, performed a comparative analysis among Radial Basis Function kernel circuitary, Multilayer Perceptron circuit nodes forest along with the canonical logistical marking regression algorithmic procedures. Abed et al. [12] contemplated a categorization method of hybrid nature for cancer infected cell's interpretation and analysis. Ivankov et al. [13] given an extensive comparison of some significant and practically used machine learning routines in the binate natured classification elucidation.

B. Research Contribution

This paper contributes as follows:

- First, the state of the art scenarios and developments, results in the probem domain are reviewed. The limitations in existing methods drove the development of proposed solution for prediction model on cancer dataset.
- Then the experiments are performed on Breast Cancer Wisconsin (Diagnostic) Data Set [14] and further the novelty of proposed methodology through comparison results of existing approaches is proved on the same dataset.

C. Organization of the Paper

The remaining paper is structured as - Section 2 elaborates about some significant definitions and preliminaries. The proposed idea and detailed procedure is given in section 3. Experimental results along with comparative performance analysis are given in section 4. Finally, conclusive summary is given in section 5.

II. Preliminaries and Definitions

This sections presents some definitions and preliminaries.

A. Neural Networks Architecture

The basic design and working functionality of this classifier is somewhat equal to the human brain (concluded in lot of neurological studies). In the point of view of structure, its flexible accordingly to the task user needs to perform I.e. high dimension diminishing, localization, regression, categorization etc. First, based on input training data and learning algorithm, the classifier model is trained. The structure consists of i/p layer, hidden (middle) layer and o/p layer.[15] Here, the training process is more time consuming and it can perform multiclass classification. The accuracy of this classifier sometimes degrades due to less effective preprocessing and presence of missing values etc.

B. Support Vector Regression

This falls under the category of supervised learning mechanism where training and testing both phases are present. This classifier is based on the phenomenon of n dimension planes and hyperplanes. The n value depends on the total number of classes and the features lies on the planes spaces as separate data points. Here the task is to find support vectors/hyperplanes, which separate the data points into various available categories.[16] The separating hyperplanes can be linear in nature or nonlinear depending upon input problem. Here, the functionality of chosen kernels, variance plays significant role in output performance parameters.

C. Fuzzy SVM

If the attributes values and classes in the data are not discrete in nature means they are continuous natured then fuzzy logic is utilised in the support vector procedure.[17, 18] Some data which includes noise elements into it, should also be processed through fuzziness logic.

D. Bayesian Classifiers

As the metadata collection, some statistical and probability computations are first associated here.[19] To find the certain correlations among features, Bayesian and naive independent procedures are utilised. This classifier was invented in 1950s and since then it is being used with lot of variations in terms of correlations scanning, imputed variable predictors, dynamic features induction, non-linearized learning etc.

III. Proposed Methodology

This section presents the core idea and detailed algorithmic procedure.

A. Core Idea

Here, the curse of dimensionality and curse of multivariate nature of data has been dealt. This research contribution tried to achieve a computationally efficient dimension reduction and further classification based disease prediction procedural framework with comparatively improved and high accuracy. The detailed procedures are given as Algorithmic frameworks 3.2.1 and 3.2.2 in below sub-section. In algorithm 3.2.1, dimension reduction is performed as a pre-processing phase to deal with the curse of dimensionality. The dataset is converted into lower dimensional space (projection of a feature space into a smaller subspace), which is aimed to get rid of overfitting and perform reduction in computational cost. Further in algorithm 3.2.2, fine-tuned neural network with stacking is applied for disease prediction analysis.

B. Procedure Steps

Algorithm 1 Algorithmic procedure 3.2.1

1: Compute → the mean vectors m_i (d-dimensional) for (i = 1, 2 $\cdots$, C_n); where, C_n: total no. of different classes for the dataset.

2: Compute → scatter matrices in two aspects -
Within-class matrix: Use eq. SMAT$_W$ = $\sum_{i=1}^{c} S_i$; where,

$$S_i = \sum_{x \in D_i}^{n} (x - m_i)(x - m_i)^T$$

$$m_i = \frac{1}{n_i} \sum_{x \in D_i}^{n} x_k$$

Between-class matrix: Use eq. SMAT$_B$ =$\sum_{i=1}^{c} N_i(m_i - m)(m_i - m)^T$; where m, m_i and N_i are - overall mean, sample mean and sizes of respective categories.

3: Compute → Eigenvectors $(ev_1, ev_2, \cdots, ev_d)$ along with the corresponding Eigenvalues $(\lambda_1, \lambda_2, \cdots, \lambda_d)$ for scatter matrices computed in step 2.

4: Make the tuples list and perform sorting of Eigenvectors with decrement in Eigenvalues.

5: k Eigenvectors are chosen with the largest of Eigenvalues and construct the d × k dimension matrix $\mathcal{M}$ (here, each column denotes an eigenvector).

6: Transform samples into new subspace by using d × k matrix i.e. -

$$Q_{(n\times k)} = P_{(n\times d)} \times \mathcal{M}_{(d\times k)}$$

Here, P denotes: $n \times d$ sized matrix possessing n samples, Q denotes: transformed $n \times k$ sized samples in reduced new feature-subspace.

Further, neural networks with MLP (multi-layer perceptron) architecture is used with finetuning for classification based disease prediction. Neurons are seperated in form of input layer, hidden layer and output layer. Steps involved in the procedure are given as algorithm 3.2.2 -

Learning rate is a configuration parameter, used to control the amount of weights that need to be updated in model procedure. Steps 1-4 are forward neural propagation and steps 5-11 are backward neural propagation.

Algorithm 2 Algorithmic procedure 3.2.2

1: Consider, X: as i/p matrix and Y: as o/p matrix. Initialize $\rightarrow$ assign random values to weights and biases.
Consider, w_{hl}: weight matrix for hidden layer
b_{hl}: bias matrix for hidden layer
w_{ol}: weight matrix for output layer
b_{ol}: bias matrix for output layer

2: Perform linear transformation as -
$Input_{hidden\_layer} = matrix\_dot\_product(X, w_{hl}) + b_{hl}$

3: Using Activation function (sigmoid), perform the non-linear transformation -
$$Activation_{hidden\_layer} = sigmoid(hidden\_layer\_input)$$
sigmoid function will return output as = $\frac{1}{1+e^{-x}}$

4: Perform -
$$o/p\_layer\_input = matrix\_dot\_product(hiddenlayer\_activations \times w_{ol}) + b_{ol}$$
o/p = sigmoid(o/p_layer_input)

5: Prediction is compared with actual output and calculate gradient of error.
Error(E) = Y - output

6: Compute -
$$Slope_{o/p\_layer} = derivatives\_sigmoid(output)$$
$$Slope_{hidden\_layer} = derivatives\_sigmoid(hiddenlayer\_activations)$$

7: Compute delta ($\triangle$) at output layer -
$$\triangle_{output} = E \times Slope_{o/p\_layer}$$

8: Error back propagation is done as -
$$E_{hidden\_layer} = matrix\_dot\_product(\triangle_{output}, w_{ol}^T)$$

9: Compute -
$$\triangle_{hidden\_layer} = E_{hidden\_layer} \times Slope_{hidden\_layer}$$

10: Update weights in network -
$$w_{ol} = w_{ol} + matrix\_dot\_product((hiddenlayer\_activations)^T, \triangle_{output}) \times value_{learning\_rate}$$
$$w_{hl} = w_{hl} + matrix\_dot\_product(X^T, \triangle_{hidden\_layer}) \times value_{learning\_rate}$$

11: Biases updated as -
$$b_{hl} = b_{hl} + sum(\triangle_{hidden\_layer}, axis = 0) \times value_{learning\_rate}$$
$$b_{ol} = b_{ol} + sum(\triangle_{output}, axis = 0) \times value_{learning\_rate}$$

IV. Experimental Analysis

This section presents an extensive experimental analysis. First the experimental envoirnmental setup, input dataset and its properties are discussed. Next, experimental output results are given. Later, in comparison table, the results of proposed procedure with significant existing approaches are compared.

A. Experimental setup

System specifications (Software and Hardware) used are as follows - OS: Ubuntu 16.04 LTS, 64 bit is used; hardware consists 4 GB RAM size along with Intel core i3 4030U CPU processor @1.90GHz $\times$ 4 clock speed. R language environment is used in the experimental analysis, with the specifications - R version - 3.4.4; RStudio Version - 1.0.136. R is a programming language and software framework for statistical analysis and computer graphics.

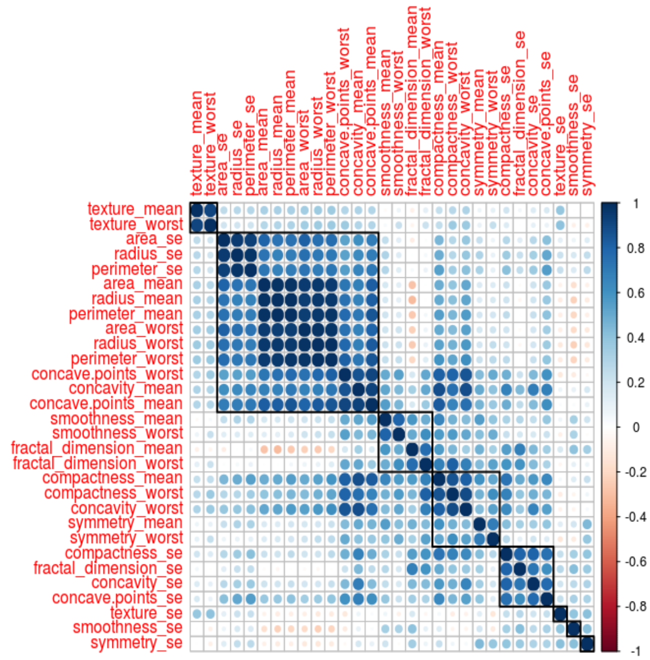


Fig. 1. Attributes correlation matrix

B. Input Dataset Details

In the experiments, *Breast Cancer Wisconsin (Diagnostic) Dataset* [14] is used. The details of this input dataset are as below:-

- Data Set Characteristics: Multivariate
- Attribute Characteristics: Real
- Number of Instances: 569
- Number of Attributes: 32
- Missing Values: N/A
- Ten real-valued features are computed for each cell nucleus
- Class distribution: 357 benign, 212 malignant.

C. Output Results Discussion

The experiments are performed on input dataset utilizing the proposed procedure. First the algorithmic procedure 3.2.1 is executed, and reduced dimensioned variables' sub-space is obtained. The attributes / variables correlation matrix is shown as Fig-1. Disease categorization for two classes of the dataset i.e. Benign(B) and Malignant(M) density distribution plot is given as Fig-2. Resultant confusion matrix after execution of the disease predictive analysis model is as follows -

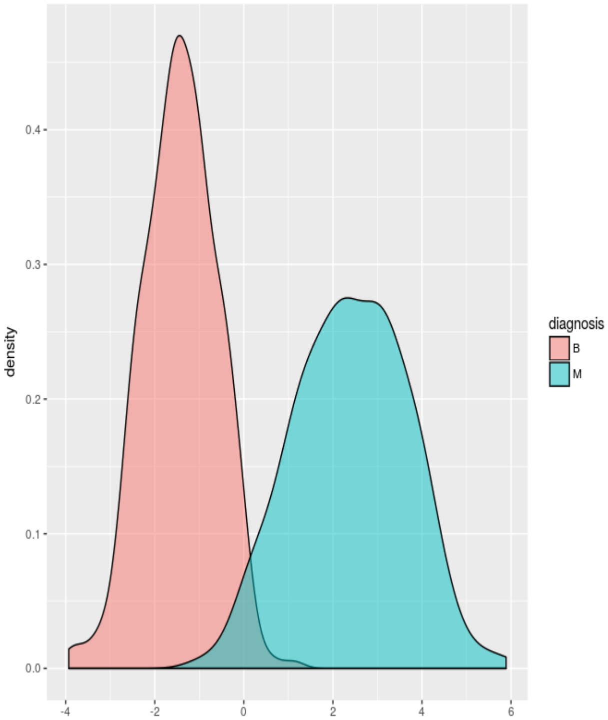


Fig. 2. Classes density distribution plot

TABLE I. EXPERIMENT RESULTS

Performance Parameters	Value
Accuracy	98.8%
P-Value	2e-16
Kappa coefficient	0.97
Sensitivity	0.98
Specificity	0.99

	Reference	
Pred	*Benign*	*Malignant*
Benign	106	1
Malignant	1	62

70:30 split ratio is used for the training and testing sample. Further, in the experimental procedure, it runs 5-folds cross validation (CV) on training set for cancer disease class prediction accuracy in test-set. Each experiment is repeated 5-times to verify the obtained accuracy of the generated models. Other statistical values / performance parameters obtained in the model (cancer disease predictive analysis model) are given in Table-1.

Graphical representation in shown as Fig-3. Better test accuracy as 98.8% is obtained along with other statistical performance parameters for disease prediction model performance i.e. P-value as 2e-16 (P-value represents, asymptotic significance for the model), Cohen's Kappa coefficient as 0.97, Sensitivity and Specificity as 0.98 and 0.99 respectively.

D. Comparative Performance Analysis

This section presents the comparative analysis where the obtained experimental results are compared with some significant existing state of the art methods on the same dataset. The comparison is given in Table-2. From comparative analysis, it is proved that the proposed procedure outperforms over other existing approaches.

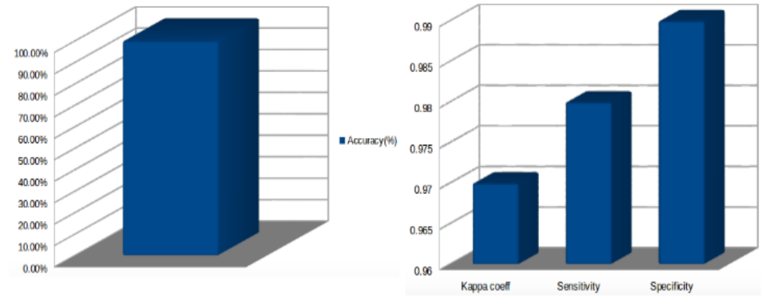


Fig. 3. Experimental analysis results: Performance parameters

TABLE II. COMPARATIVE ANALYSIS

Reference	Method	Dataset Split ratio	Prediction accuracy
E. Zafiropoulos et al. [20]	SVM with Gaussian RBF	70:30	89.28%
Ali et al. [8]	SVM with quadratic kernel	70:30	88%
Ali et al. [8]	SVM with poly kernel	70:30	88%
Ali et al. [8]	Neural Networks model	75:25	92%
S.S. Shajahaan et al. [9]	ID3 algorithm	-	92.99%
S.S. Shajahaan et al. [9]	C4.5 algorithm	-	95.57%
S.S. Shajahaan et al. [9]	CART	-	92.42%
S.S. Shajahaan et al. [9]	Naive Bayes	-	97.42%
J. Ivancakova et al. [13]	SVM	70:30	97.66%
J. Ivancakova et al. [13]	Random Forests	70:30	97.37%
J. Ivancakova et al. [13]	C4.5	70:30	95.61%
Proposed approach	fine-tuned and stacked DR-NN model	70:30	98.8%

V. CONCLUSIVE SUMMARY

Efficient preprocessing mechanisms to make the intelligent learning and prediction systems capable of dealing with multivariate data and effective learning technologies to find out the rules to describe the data are still of urgent need. Some limitations in the existing methods motivated us to propose a machine learning based fine-tuned and stacked model, utilizing which the experimental analysis is performed on cancer dataset in an computationally efficient manner with improved disease prediction accuracy.

REFERENCES

[1] Marleende Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis, Medical Image Analysis, Volume 33, October 2016, Pages 94-97.

[2] Masih Sherafatian. Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping, Gene, Volume 677, 30 November 2018, Pages 111-118.

[3] Haifeng Wang, BichenZheng, Sang Won Yoon, Hoo Sang Ko. A support vector machine-based ensemble algorithm for breast cancer diagnosis, European journal of Operational Research, Volume 267, Issue 2, 1 June 2018, Pages 687-699.

[4] Chip M.Lynch, Behnaz Abdollahi, Joshua D.Fuquac, Alexandra R.de Carlo, James A.Bartholomai, Rayeanne N.Balgemann, Victor H.van Berkel, Hermann B.Frieboes. Prediction of lung cancer patient survival via supervised machine learning classification techniques, International Journal of Medical Informatics, Volume 108, December 2017, Pages 1-8.

[5] Fonsvan der Sommen, Sander R.Klomp, Anne-FrSwager, Svitlana Zinger, Wouter L.Curvers, Jacques J.G.H.M.Bergman, Erik J.Schoon, Peter H.N. de. Predictive features for early cancer detection in Barrett's esophagus using Volumetric Laser Endomicroscopy, Computerized Medical Imaging and Graphics, Volume 67, July 2018, Pages 9-20.

[6] Jyotsna T. Wassan, Haiying Wang, Huiru Zheng. Machine Learning in Bioinformatics, Reference Module in Life Sciences, 2018.

[7] Naresh Khuriwal, Nidhi Mishra. Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm, IEEMA Engineer Infinite Conference (eTechNxT), IEEE, 13-14 March 2018.

[8] E. E. E. Ali, W. Z. Feng. Breast Cancer Classification using Support Vector Machine and Neural Network, International Journal of Science and Research, vol. 5, issue 3, 2016, pp. 1-6.

[9] S. S. Shajahaan, S. Shanthi, V. ManoChitra. Application of Data Mining Techniques to Model Breast Cancer Data, International Journal of Emerging Technology and Advanced Engineering. vol. 3, issue 11, 2013, pp. 362-369.

[10] V. Chaurasia, S. Pal. A Novel Approach for Breast Cancer Detection using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, issue 1, 2014, pp. 2456-2465.

[11] J. Padmavathi. A Comparative study on Breast Cancer Prediction Using RBF and MLP, International Journal of Scientific & Engineering Research, vol. 2, issue 1, 2011, pp. 1-5.

[12] B. M. Abed, K. Shaker, Hamid A. Jalab. A hybrid classification algorithm approach for breast cancer diagnosis, Proceedings of IEEE Industrial Electronics and Applications Conference (IEACon 2016), Malaysia, 2016, pp. 264-269.

[13] Juliana Ivancakova, Frantiek Babi, Peter Butka. Comparison of Different Machine Learning Methods on Wisconsin Dataset, IEEE 16th World Symposium on Applied Machine Intelligence and Informatics, Slovakia, February 7-10, 2018.

[14] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+ Wisconsin+%28 Diagnostic%29

[15] S Agatonovic-Kustrin, R Beresford. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, Journal of Pharmaceutical and Biomedical Analysis, Volume 22, Issue 5, June (2000), Pages 717-727.

[16] Cortes, C., Vapnik, V. (1995). Support-vector networks, Machine Learning, 20(3), 273-297.

[17] Abonyi, J., Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers, Pattern Recognition Letters, 24(14), 2195-2207.

[18] C.-F. Lin and S.-D. Wang, Fuzzy support vector machines, IEEE Transactions on Neural Networks, vol. 13, no. 2, March 2002.

[19] Nir Friedman, Ron Kohavi. Bayesian classification, Stanford Artificial Intelligence Laboratory, (1999).

[20] E. Zafiropoulos, I. Maglogiannis, and I. Anagnostopoulos. 2006. A support vector machine approach to breast cancer diagnosis and prognosis, Artificial Intelligence Applications and Innovations (2006), 500-507.

Video Streaming Analytics for Traffic Monitoring Systems

Muhammad Arslan Amin*, Muhammad Kashif Hanif†, Muhammad Umer Sarwar‡, Muhammad Kamran Sarwar§, Ayesha Kanwal¶, and Muhammad Azeem‖
Department of Computer Science,
Government College University, Faisalabad, Pakistan

***Abstract*—It is considered a difficult task to have check on traffic during rush hours. Traditional applications are manual, costly, time consuming, and the human factors involved. Large scale data is being generated from different resources. Advancement in technology make it possible to store, process, analyze, and communicate with large scale of video data. The manual applications are wiped out with the invention of automatic applications. Automatic video streaming analytics applications helps to reduce computational resources. The reason is cost efficient and accurate predictions while monitoring traffic on roads. This study reviews previously developed application of video streaming analytics for traffic monitoring systems using Hadoop that are able to efficiently analyze video streams.**

***Keywords*—Video streaming analytics; Traffic monitoring system; Video streams; Hadoop; GPU; CNN; Deep learning**

I. Introduction

Traffic monitoring is considered a complicated task on heavy roads. In traditional traffic monitoring systems, traffic control rooms monitor the live video streams obtained from live-feed cameras or stored on a bank of digital video recorders or computer hard disk drives. Manual registration of auto mobiles, counting of vehicles by magnetic loops and road side movement cameras are commonly used approaches by traffic police for traffic monitoring. If the video analysis are not done automatically then it requires a number of human operators and a control room with resources to analyze the footage manually. These approaches are costly, time consuming, and inaccurate. The most frequently asked questions during occurrence of an event are about color, size, and type of vehicle. It is difficult to accurately detect, classify, and track objects.

Video analytics is also known as video content analysis. It involves different techniques to analyze, monitor and fetch meaningful information from video streams. There exists different techniques for the analysis of real-time and pre-recorded videos. One of the major applications of video stream analysis is automated security and surveillance systems. Video analytic techniques can be used for public safety and security by law enforcement agencies, work stations, and organizations. Automatic video analysis requires less cost and human efforts to analyze video stream than manual applications [1].

Video analytics are helpful in different fields like business intelligence, healthcare, law enforcement, industries, and traffic monitoring systems. The study discusses application of video analytics for traffic monitoring system. Automated video analytics can help to reduce the human efforts and computational cost [1]. Video analytics can help to detect objects like face, vehicle color and vehicle number plate. Object detection is a process to recognize real-world objects and object classification is to categorize or classify the real-world objects such as determining that an object is a person, vehicle or animal. Classification is different from detection or recognition. The reason is object classification is unable to identify objects. Object classification and detection can be used in simple and complicated applications. Object classification can classify the vehicle into car, bus, van and truck by their size, detected as they pass through the region of interest and represented in a square rectangle [2].

Object detection is a useful technique for traffic monitoring systems and law enforcement agencies. Object detection can be done using different algorithms. The most common algorithms are template matching [3], background separation using Gaussian Mixture Models (GMM) [4], and cascade classifiers [5]. Template matching is suitable only for objects defined in templates. This approach will not work for the new objects. Background frame differencing approach only detects the objects which are moving in background [4]. Background separation and frame differencing are expensive techniques for performing computations [2], [6]. A cascade classifier is an object detection approach which uses a real AdaBoost algorithm [5]. Cascade classifier increases the detection performance and it is comparatively affordable and useful [2], [6].

The displacement of centroids can help to calculate the speed of a moving object. The camera calibration parameters helps in the problem of perspective distortion that utilizes to transform the pixel distance measured from the video frames into real world distance. The accuracy of the estimated speed is calculated by comparing with average traffic speed. The reference background image is updated when there is a change in the illumination. Traffic congestion can be detected when vehicles are traveling below the average speed for a period of time. Similarly, road accidents were detected when there is a difference between centroids of vehicles and the speed falls close to zero. It may not give an accurate results for road accidents. The collision can be detected when the centroid of the vehicles collides with another. For producing rerouting messages to subscribers if there is an occurrence of congestion, current places of clients are computed utilizing the geolocation API and an alternate route estimation is given utilizing Google Maps API [7].

II. Video Streaming Analytics using Hadoop

Cloud computing can be used for automatic video stream analysis. This section discusses different techniques for video stream analysis.

A. Video Streaming Analytic using Hadoop based GPU Cluster

A large scale of data is being generated from different resources like GPS, social media, internet of things, tablets, web services and videos. It is difficult task to analyze, process, and store video data. To handle large scale of video data, processing can be distributed among different nodes. There can be hundreds or even more cameras over a wide area that captures the stream and process at the local server. For such a large scale, the cloud based storage can be used [6].

Video streaming analytics requires computations at a very large scale. This can be done using cloud computing. Cloud computing provides high scale computing and efficiency in real time monitoring of traffic video data. Commercial vendors also provide applications of video analytics. In Vi-System[1], there is an intelligent surveillance system which tracks and monitor objects with the help of analytical protocols. It provides updates to different users based on defined parameters. The Vi-System does not work well for recorded videos and limited rules of analysis that needs to be defined in advance. Another project BESAFE[2] can helps to automatic surveillance and tracks the abnormal behavior. The downside of BESAFE project is scalability. An intelligent vision tool[3] can be used to perform video analysis intelligently with the help of automatic video monitoring. However, this tool is also less scalable and sometimes does not fulfill the requirements.

Abdullah et al., proposed a framework which utilizes a cloud based storage to store and capture video streams and a GPU cluster for video stream analysis (Figure 1). Stream acquisition component gets the videos from surveillance cameras and forward it to the requesting client in traffic control room to store these streams. These video streams are stored on cloud server as mp4 files. APS server is responsible for video streaming analytics. APS server contains cloud storage to store recorded video streams and a processing server having compute nodes based GPU cluster. The analysis is started after client submits job. Then video streams are downloads from cloud storage. The meta-data of video stream is stored on the analytical database. APS client is responsible for interaction between end-user and APS server. The user can specify the duration of video, criteria for analysis, i.e., vehicle detection, classification etc. These video streams are fetched automatically from Hadoop based cloud storage, decoded and analyzed on Hadoop based GPU cluster without intervention of operator. The user is notified after completion of analysis [6]. This work was extended by incorporation of features like face and vehical detecion [2]. GPU enabled video analytics process can perform faster than the CPU and human operators.

[1]https://www.agentvi.com/
[2]http://imagelab.ing.unimore.it/besafe/
[3]https://www.intelli-vision.com/

B. Video streaming analytics using Hadoop MapReduce

The previous framework of video streaming analytics in clouds using Hadoop based GPU cluster is not appropriate for live video data analysis, e.g., detection of road accidents and congestion etc. Public safety on roads and helping them in crowded situations is a very difficult and demanding task for traffic monitoring systems in developing countries. Natarajan et al., presented a video analytics framework using Hadoop MapReduce that can work better for live video data analysis. This framework uses video processing based on MapReduce. Hadoop is used to split video frames into smaller unit and convert them in a sequence of image frames in parallel mode. HDFS stores large amount of data over a number of physical disk clusters. The sequence file after splitting into smaller units is distributed over the nodes of HDFS to increase the processing speed. Raw video stream is split into 64 MB (default) size blocks [7].

This framework is divided into two phases. First, mapping phase assigns the data to a cluster node from the single video frames. Every frame processed in map phase produces a corresponding output which is indexed for better access. The index consists of a key-value pairs which connects a frame identifier with corresponding data. In second phase, index maps are reduced into a single result. The information about vehicle accidents, speed and congestion are kept in Hive for further analysis. The system architecture consists of media server, IP wireless transmission, video streams, and supervising/control station (Figure 2). Video recording module streams the video to supervising station and stores the video in its local storage. The video streams are captured from different places and transferred by IP wireless transmission module to the supervising station via media server. The live video streams are processed by supervising station and results are stored for further analysis. The information about vehicle speed, accident and congestion are stored in Hive. When road congestion occurs, Hive tables can suggest the alternative routes to the travelers. The important places around the city are stored in Hive tables. When road accidents is detected, the system sent alerts to the nearby hospitals and highway rescue teams [7].

C. Video Analytics Using Hadoop Video Processing Interface

Hadoop does not provide video read/write interface. Moreover, Hadoop framework is implemented in java. However, many video analysis applications are implemented in C/C++. In these situations, Hadoop framework is not compatible.

To overcome these issues, Zhao et al., have design Hadoop Video Processing Interface (HVPI) (Figure 3). HVPI helps to port existing video analytic applications into Hadoop platform written in C or C++ and make easy use of video read/write interface. Video read/write interface is responsible to read video input and convert it into key-value pair for mapper to process. After completing the process on video read/write interface, we can analyze the frame sequence by implementing video analytics applications [8].

D. Vehicle Recognition and Detection using Deep Learning

Chen et al., have presented convolutional neural network (CNN) and MapReduce technique for video vehicle recognition and detection. They have utilized an efficient deep learning

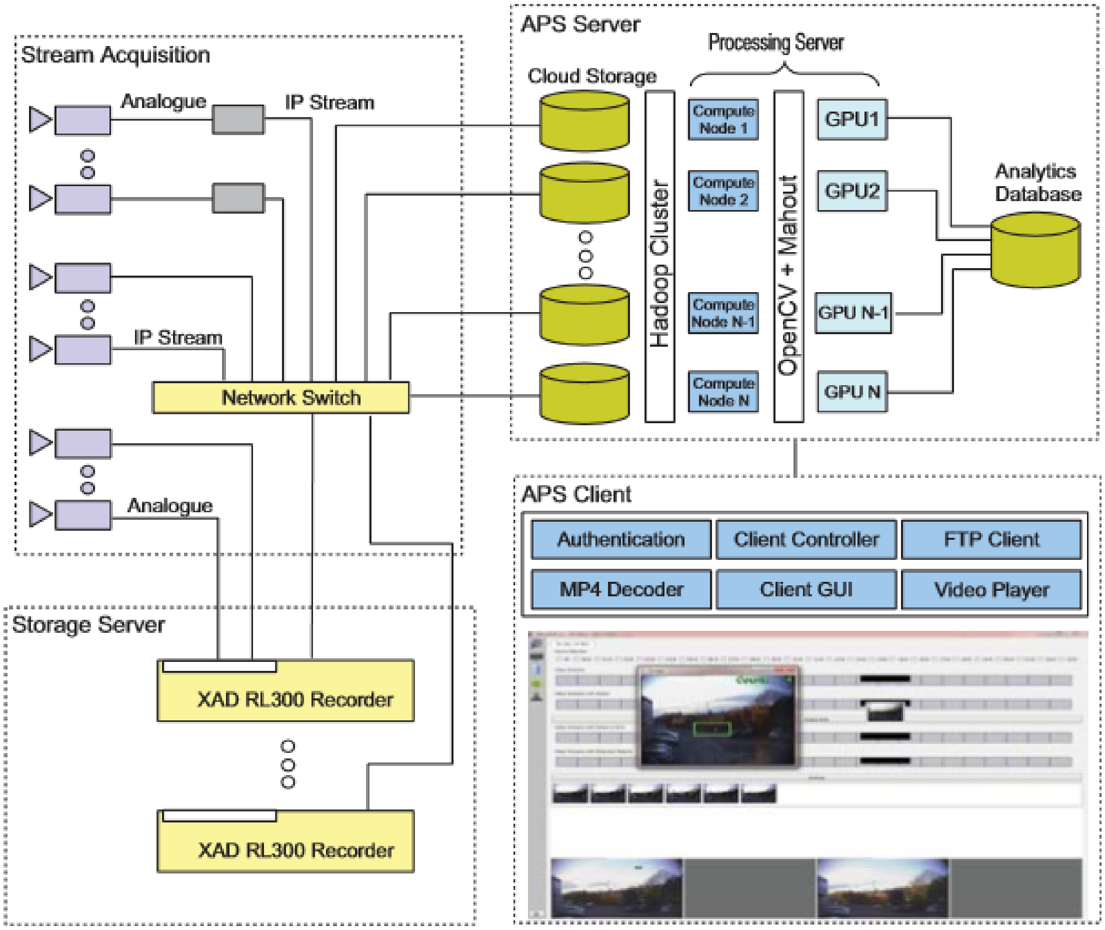


Fig. 1. System architecture of stream cloud [2], [6].

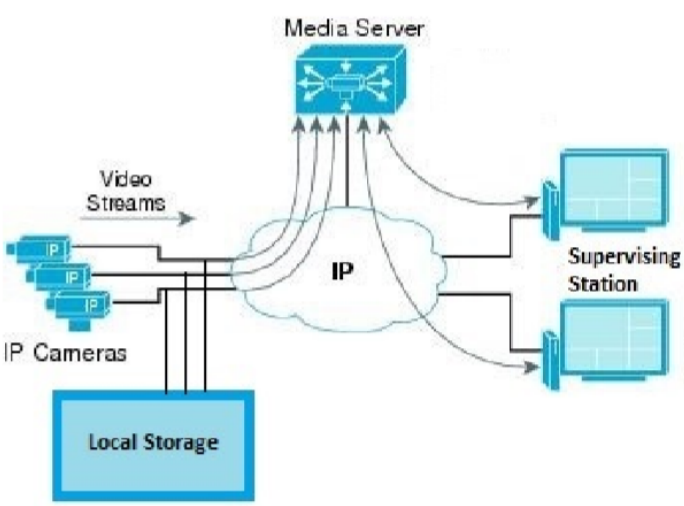


Fig. 2. System architecture of Hadoop MapReduce based video analytics framework [7]

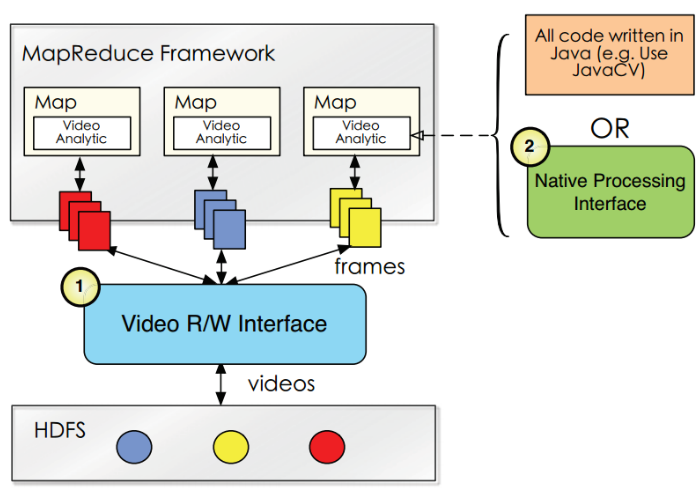


Fig. 3. HVPI architecture [8]

(DL) algorithm based on YOLOv2 [9] to detect vehicles in real-time. Moreover, they presented an improved CNN based license plate recognition algorithm. They have combined the HVPI and MapReduce framework for parallel processing of video vehicle detection and recognition. The results indicate that their method is less time-consuming and provides an high accuracy for detection. This method can process large scale video data [10].

III. Conclusion

Video analytics in traffic monitoring system can help in different forms like public safety on roads and violation of traffic rules. Video analytics is an analysis of video streams by using computational resources. Video streaming analytics can works efficiently with the help of automatic systems. The video analytics in traffic monitoring systems can provide information of object detection and object classification. The framework for video streaming analytics in clouds using Hadoop based GPU cluster is capable to analyze recorded video streams. This study discussed different Hadoop based techniques for automatic traffic video analysis. However, Hadoop does not provide video read/write interface and many analytical applications does not work well with Hadoop framework. There exists some frameworks like HVPI which can overcome these issues.

References

[1] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.

[2] A. Anjum, T. Abdullah, M. Tariq, Y. Baltaci, and N. Antonopoulos, "Video stream analysis in clouds: An object detection and classification framework for high performance video analytics," *IEEE Transactions on Cloud Computing*, vol. PP, no. 99, pp. 1–1, 2016.

[3] J. S. Bae and T. L. Song, "Image tracking algorithm using template matching and psnf-m," *International Journal of Control Automation and Systems*, vol. 6, no. 3, p. 413, 2008.

[4] S. C. Sen-Ching and C. Kamath, "Robust techniques for background subtraction in urban traffic video," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 881–892.

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[6] T. Abdullah, A. Anjum, M. F. Tariq, Y. Baltaci, and N. Antonopoulos, "Traffic monitoring using video analytics in clouds," in *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, Dec 2014, pp. 39–48.

[7] V. A. Natarajan, S. Jothilakshmi, and V. N. Gudivada, "Scalable traffic video analytics using hadoop mapreduce," *ALLDATA 2015*, p. 18, 2015.

[8] X. Zhao, H. Ma, H. Zhang, Y. Tang, and Y. Kou, "Hvpi: Extending hadoop to support video analytic applications," in *2015 IEEE 8th International Conference on Cloud Computing*, June 2015, pp. 789–796.

[9] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint*, 2017.

[10] M. Chen, W. Wang, S. Dong, and X. Zhou, "Video vehicle detection and recognition based on mapreduce and convolutional neural network," in *International Conference on Sensing and Imaging*. Springer, 2018, pp. 552–562.

Role of Bloom Filter in Big Data Research: A Survey

Ripon Patgiri, Sabuzima Nayak, Samir Kumar Borgohain
Dept of Computer Science & Engineering
National Institute of Technology Silchar
Assam-788010, India

***Abstract*—Big Data is the most popular emerging trends that becomes a blessing for human kinds and it is the necessity of day-to-day life. For example, Facebook. Every person involves with producing data either directly or indirectly. Thus, Big Data is a high volume of data with exponential growth rate that consists of a variety of data. Big Data touches all fields, including Government sector, IT industry, Business, Economy, Engineering, Bioinformatics, and other basic sciences. Thus, Big Data forms a data silo. Most of the data are duplicates and unstructured. To deal with such kind of data silo, Bloom Filter is a precious resource to filter out the duplicate data. Also, Bloom Filter is inevitable in a Big Data storage system to optimize the memory consumption. Undoubtedly, Bloom Filter uses a tiny amount of memory space to filter a very large data size and it stores information of a large set of data. However, functionality of the Bloom Filter is limited to membership filter, but it can be adapted in various applications. Besides, the Bloom Filter is deployed in diverse field, and also used in the interdisciplinary research area. Bioinformatics, for instance. In this article, we expose the usefulness of Bloom Filter in Big Data research.**

Keywords*—Bloom Filter; Big Data; Database; Membership Filter; Deduplication; Big Data Storage; Flash memory; Cloud Computing

I. INTRODUCTION

Big Data has been revolutionizing the IT industry since its inception, and it is proven as a game changer paradigm. Nowadays, Big Data is continuously dominating IT industry, because data are oils of the modern era. Interestingly, Big Data touches almost all fields, including Government sector, Business, Politics, Science, Economy, and Healthcare [1]. Therefore, datacenter deals with a variety of data from various sources. The varieties of data are typically classified into three category, namely, structured, semi-structured, and unstructured data which compose a large sized data silo. Revenue and investments are growing, and it is expected to grow till 2030. Big Data is a data silo of variety of data that grows exponentially. From the year 2003, Big Data grows exceptional pace and continue to grow with the same pace. Moreover, Big Data gears up the speed speed of growth in 2009. In 2012, IDC predicted that Universal Data size should touch 44 ZB in 2020. In most of the IT industries, petabytes of data are a dime a dozen. As we know that data are generated through various sources, for instance, IoT. Various organizations are gathering data to generate revenue. Because, data are an asset to the organizations. Most of the companies have surpassed petabytes of data, and they have already landed in exabytes of data. For example, Google and NSA reach 15 EB and 10 EB data. Undoubtedly, it is an onerous task to deal with this kind of data sizes. Thus, Big Data is defined by $V_3^{11} + C$ [1].

Undoubtedly, managing a large data silo is a grand challenge. There are dozens of issues and challenges in Big Data [2]. Therefore, there are numerous research papers have been published to address various issues of Big Data by various publishers, for example, Elsevier, ACM, IEEE, and Springer. Similarly, Bloom Filter [3] also applied in numerous filed and published numbers of papers by various publishers. Figure 1 extrapolate the number of patents awarded based on Bloom Filter according to Google Patent Search Engine. In other word, Bloom Filter is very popular hashing technique. In addition, Bloom Filter is a very simple yet powerful data structure to handle large scale data by sacrificing a small amount of memory space. Unlike a conventional hash data structure, Bloom Filter is more space efficient and Bloom Filter is able to reduce memory space consumption by order of magnitude. The Bloom Filter is an approximate membership query. Therefore, the Bloom Filter is unable stand itself, but it is just like a cog in a machine. Big Data can be enhanced in the presence of Bloom Filter.

In this article, we present the impact of Bloom Filter in Big Data. We also outline future challenge of Bloom Filter in Big Data. In Section III, we explore the employment of Bloom Filter in Big Data storage system. Section II outlines Bloom Filter. Section III reviews the employment of Bloom Filter in Big Data storage system. In addition, comparative study is presented in Section III. Section IV highlights the future direction of Bloom Filter in Big Data storage system. And finally, article is concluded by Section V.

II. BLOOM FILTER

Bloom Filter is a probabilistic data structure with some errors. Bloom Filter is a variant of hashing algorithm that uses a tiny amount of extra memory space which is introduced by Burton Howard Bloom [3]. Bloom Filter has been applying various research area to boost up the performance of a system. Bloom Filter is a membership filtering algorithm that returns either “true” or “false”. However, the “true” can be either false positive or true positive. Similarly, the “false” can also be either false negative or true negative. The false positive and false negative are the error of Bloom Filter. Nevertheless, the error is negligible and tolerable. But, Bloom Filter does not suit for many systems too. Real-time system, for instance.

Let us assume, P and Q are two elements to be inserted into the Bloom Filter. Bloom Filter depends on the k value

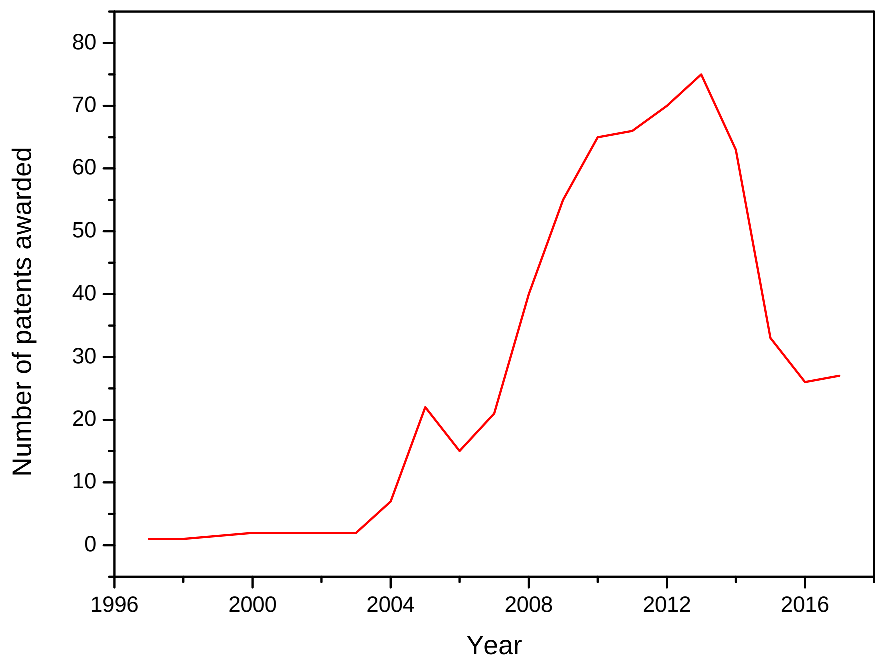


Fig. 1. Number of patent awarded in Bloom Filter/using Bloom Filter according to Google Patent Search Engine.

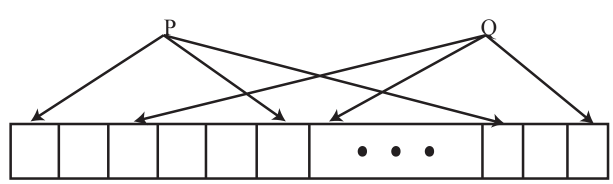


Fig. 2. Conventional Bloom Filter.

which is the number of hash functions of a key or an element. If $k = 3$, then Figure 2 depicts the number of cells occupied by P and Q. P is hashed into three different places in the array. Some variants of Bloom Filter store fingerprints of the keys, and some variants just put '1' or '0'. Similarly, Q is also hashed into three different locations in the array. Those three positions must be set (true) while checking the membership of P or Q and it is a necessary condition for positive answer of the Bloom Filter.

Let, S be a set and K be the elements of S where $K \in S$, $K = K_1, K_2, K_3, \ldots, K_n$, and n is the total number of elements. Let, B be the Bloom Filter and insert all elements of S into B. Let us assume, K_j be a random query element where $j = 1, 2, 3, \ldots$. Article [4] defines false positive, true positive, false negative, and true negative as follows-

- False positive: If $K_j \notin S$ and Bloom Filter returns $K_j \in B$, then the result of Bloom Filter is a false positive.
- True positive: If $K_j \in S$ and Bloom Filter returns $K_j \in B$, then the result of Bloom Filter is a true positive.
- False negative: If $K_j \in S$ and Bloom Filter returns $K_j \notin B$, then the result of Bloom Filter is a false negative.
- True negative: If $K_j \notin S$ and Bloom Filter returns $K_j \notin B$, then the result of Bloom Filter is a true negative.

The key issue in Bloom Filter is false positive and false negative. Most of the Bloom Filter variants suffer from false positive, but not false negative. However, the false negative issue arises in the deletion of an element. The counting Bloom Filters are introduced to address the issue of scalability. But, false negative is a serious issue with the counting Bloom Filter upon deleting an item. Therefore, most of the Bloom Filters avoid deletion of a key. On the other hand, Cuckoo Filter [5] reduces the false positive, however, it suffers from performance issue. Also, Rottenstreich et al. [6] implements variable counting Bloom Filter, called CBF. CBF avoids increment by one on insertion to reduce false positive. CBF uses B_h sequence to increase upon insertion and decrease upon deletion [6]. However, memory consumption per item is very high.

Let us outline the false positive of conventional Bloom Filter. Let, m be the total memory occupied by the Bloom Filter, n be the total element entry into the Bloom Filter and k be the total number of hash functions, then the probability of a bit to be '0' after n element insertion in the given filter is [7]

$$\left(1-\frac{1}{m}\right)^{nk}$$

The probability of the given bit to be '1' is

$$\left(1-\left(1-\frac{1}{m}\right)^{nk}\right) \approx \left(1-e^{-kn/m}\right)$$

The probability of all bits to be '1' is

$$\left(1-\left(1-\frac{1}{m}\right)^{nk}\right)^{k} \approx \left(1-e^{-kn/m}\right)^{k}$$

On the other hand, F. Grandi [8] present another approach to analyze the false positive probability. Let, X be the random variable representing the total number of set bits (i.e., '1'), then

$$E[X] = m\left(1-\left(1-\frac{1}{m}\right)^{nk}\right)$$

Let us conditioned the false positive to a number $X = x$, then

$$Pr(FP|X=x) = \left(\frac{x}{m}\right)^{k}$$

where x is set bits. Then, F. Grandi [8] calculates the false positive probability (FPP) as follows

$$\begin{aligned} FPP &= \sum_{x=0}^{m} Pr(FP|X=x)\ Pr(X=x) \\ &= \sum_{x=0}^{m} f(x)\left(\frac{x}{m}\right)^{k} \end{aligned}$$

where $f(x)$ is probability mass function. F. Grandi [8] applies $\gamma - tranformation$ to obtain the value of probability mass function, and the false positive probability of standard Bloom Filter is calculated as follows

$$FPP = \sum_{x=0}^{m}\left(\frac{x}{m}\right)^{k}\binom{m}{x}\sum_{j=0}^{x}(-1)^{j}\binom{x}{j}\left(\frac{x-j}{m}\right)^{kn}$$

Thus, the probability of false positive also depends on m and n. The false positive becomes overhead of Bloom Filter. If a system can tolerate negligible overhead, then Bloom Filter can enhance a performance of a system. Therefore, Bloom Filter is deployed to various domains, namely, Big Data [9], Deduplication [10], [11], [12], Network Security [4], [13], Network Traffic control [14], Name Lookup [15], [16], IP address lookup [17], [18], Biometric [19], [20], Bioinformatics [21], [22], File System [23], Indexing, and many more. However, Bloom Filter is not suitable in case of correct query-answer requirements. For instance, password database. Because, Bloom Filters response a query by approximating either true or false.

A. *Classifications of Bloom Filter*

Bloom Filter is classified into six major categories as sketched in Figure 3, namely, Standard Bloom Filter, Counting Bloom Filter, Hierarchical Bloom Filter, Multidimensional Bloom Filter, Fingerprint-based Bloom Filter, and Compressed Bloom Filter.

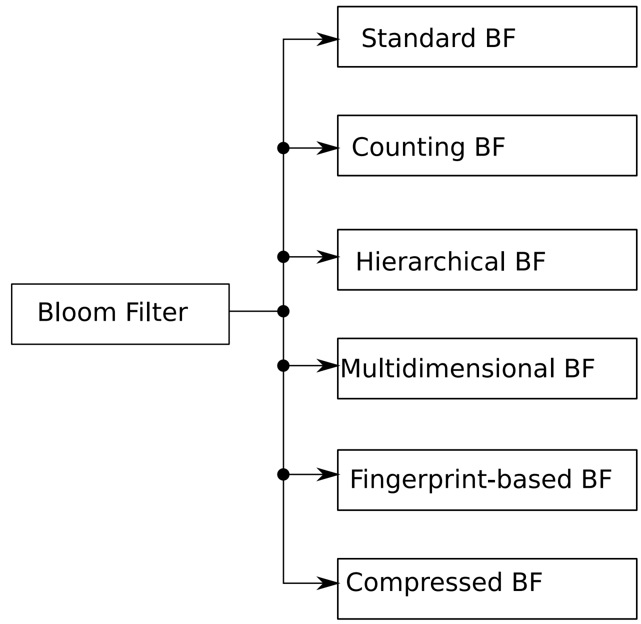


Fig. 3. Taxonomy of Bloom Filter

1) Standard Bloom Filter: Standard Bloom Filter (SBF) is known as conventional Bloom Filter. There are false positive and false negative in SBF. False negative occurs in case of deleting some elements, since, Bloom Filter is a stateless filter. Therefore, SBF cannot guarantee the false negative. Also, false positive of SBF depends on number of hash functions which is termed as k. The value of k cannot be too large or too small. The optimal value of k is defined by $k = \frac{m}{n}\ ln2$.

2) Counting Bloom Filter: Counting Bloom Filter (CBF) emerges due to lack of scalability in SBF. CBF is able to scale very high number of inputs. CBF comprises of many counters in the array, and the counters count the number of inputs. Thus, CBF is able to scale very large amount of input as compared to SBF. The counters are incremented in the insertion of an input item and decremented in the deletion of an input item. Therefore, there is very high false positive probability as compared to SBF. Also, CBF causes false negative due to deletion.

3) Fingerprint Bloom Filter: CBF has a high false positive probability. Fingerprint-based CBF has been introduced to reduce the false positive probability. For instance, Cuckoo Filter [5]. Fingerprint-based CBF stores a small sized hash value instead of dealing with binary bits. Thus, a fingerprint can directly be matched with a stored fingerprint which reduces false positive probability. However, the space consumption is more than CBF.

4) Hierarchical Bloom Filter: The hierarchical Bloom Filters (HBF) are known as tree or forest-structured Bloom Filter. The HBF is a Bloom Filter by augmenting several Bloom Filter to achieve high accuracy, scalability, and low false positive probability. For instance, Forest Structured Bloom Filter [24] and BloomStore[25]. HBF is deployed in deduplication of very large scale datasets. Moreover, HBF uses Flash-memory or HDD to increase scalability.

TABLE I. BLOOM FILTER IN A LARGE SCALE MEMBERSHIP FILTERING

Author or Name	Name of Bloom Filter	False Positive	False negative	Scala-bility	Delete	Type	Platform	Purposes
BigTable[9]	Conventional	✓	×	×	×	Conventional	RAM & HDD	Database
BloomFlash [10]	BloomFlash	✓	×	✓	×	Hierarchical	RAM & Flash	Database
Lu et al [24]	Forest-structured BF design (FBF)	✓	×	✓	✓	Forest Structured	RAM & Flash	Database
Lee et al. [26]	Bloomjoin	✓	×	×	×	Conventional	RAM & HDD	MapReduce Join
Wang et al. [27]	Bloom Filter Tree	✓	×	✓	×	Hierarchical	RAM & HDD	Auditing
BloomStore [25]	BloomStore	✓	×	✓	✓	Chain	RAM & Flash	Key/Value deduplication
DBLK [28]	Multilayer Bloom filter	✓	×	✓	×	Hierarchical	RAM & HDD	Deduplication & Compression
Blasco et al. [29]	Conventional	✓	×	✓	×	*Conventional*	Client-Server	Proof of Ownership
Dong et al. [30]	Counting Bloom Filter	✓	✓	✓	×	Counting	Data Center	Storage Cluster
DBA [23]	Dynamic Bloom filter Array	✓	✓	✓	✓	Dynamic	General	Large scale variable membership filter
Guo and Efstathopoulos [31]	Conventional	✓	×	×	✓	Conventional	RAM & SSD	Deduplication
Li et al. [12]	Conventional	✓	×	×	×	Conventional	Cloud	Storage deduplication
Zhang et al. [11]	Conventional	✓	×	✓	×	Conventional	Distributed System	Storage deduplication
I-sieve [32]	Conventional	✓	×	✓	×	Conventional	Cloud	Storage deduplication
FFBF [33]	Fuzzy Folded Bloom Filter	✓	×	✓	×	Conventional	Cloud Computing	Bloom Filter-as-a-Service for Big Data.
Khare et al. [34]	Standard Bloom Filter	✓	×	✓	×	Conventional	Distributed System	Frequency count of sub-graph and pruning.

5) *Multidimensional Bloom Filter:* Unlike conventional Bloom Filter, Multidimensional Bloom Filter (MDBF) uses a multidimensional Bloom Filter array. The array may be 2D, 3D, or many dimensions to reduces the false positive probability. Less work has been done in this MDBF.

6) *Compressed Bloom Filter:* Compressed Bloom Filter [35] is very space efficient Bloom Filter. However, there is a trade-off between space consumption and false positive on compressed Bloom Filter. The trade-off is- "more compressing results high false positive probability".

III. DATA STORAGE

Big Data consists of a variety of data, namely, structured, semi-structured and unstructured data. About 90% of the data are unstructured in Big Data. Moreover, the data are humongous in size. It is very tough task to store, process and manage. The data size is increasing daily basis due to involvement of people in new technology. Smart Phone, for instance. People are generating data through various ways, namely, IoT, Modern Healthcare, Web, Social Media, and business. Thus, deduplication plays a vital role in preventing duplicate data in Big Data paradigm. Bloom Filter is a tiny data structure to manage the mammoth size of data to filter out the duplicate data.

Table I extrapolates and compares the various features of the recent deployment of Bloom Filter in Big Data storage system. In the comparative study, we have found that Bloom Filter has been deployed for various purposes.

A. Flash memory

Interestingly, tape drives are used to backup a bulk size of data. Tape drives are cheap and high storage capacity. However, tape drives are very slow. Therefore, non-working set of data are dumped in data warehouse using the tape drive. A working set of data is stored in HDD nowadays. The active working set of data is stored in SSD/PICe-PCM/NVMe and passive working set of data are stored in HDD. Processing speed of RAM is 1000 times faster than HDD and 100 times faster than SSD. Cost of RAM is higher than SSD and HDD. SSD is costlier than HDD. Thus, SSD/PCIe-PCM/NVMe replace the HDD in Big Data era. Most of the datacenters are upgrading their storage to SSD/Flash memory. However, HDD is also used for various purposes.

BloomFlash [10] is a design of a NAND Flash memory based storage system for very large scale data size. BloomFlash consider the case of very large Bloom Filter that does not fit in RAM. BloomFlash is cleverly designed to maximize the random writes and random read throughput. In-place bit update is costlier in SSD than RAM. Hence, BloomFlash employs a lazy update policy of Bloom Filter in Flash memory. In addition, BloomFlash augment tiny Bloom Filter in every flash page to localize the reads and writes on Flash memory. Thus, BloomFlash is able to scale massive amount of data.

Moreover, Lu et al. [24] devises a forest structured Bloom Filter (FBF) for Flash memory storage. FBF deals with RAM and Flash memory. Similar to BloomFlash [10], Bloom Filter is stored in RAM and Flash memory. Write and read operations are instantly done in RAM-based Bloom Filter. On the contrary, write operation request into Flash memory is delayed and it is buffered in memory to update later on. The update to Bloom Filter is performed periodically to maximize the write performance. Random write operation takes more time than random read operation. Because, write operation requires erasing the data present in the Flash memory. Hence, it is a clever way to wait for some time to write information into the Flash memory. However, read operations are instantly accessed in the Flash memory. The FBF is forest-structured Bloom Filter that initially resides in RAM and expands the Bloom Filter to Flash memory. The Bloom Filter sizes are small and fixed, and thus, it is very easy to grow. The lookup time complexity of FBF is $O(\log_b m)$ due to forest-structured, where b is the total branches (degree of the tree) and m is the overall FBF size. The FBF is very useful in very large scale deduplication of data.

Similarly, BloomStore[25] is a key/value storage system using NAND Flash memory. BloomStore is similar to FBF [24]. Each physical node spawns several BloomStore within its key ranges. The key ranges are disjoint to each other. The BloomStore addresses an efficient key/value storage in Flash memory. Conventional hash consumes more storage space as compared to Bloom Filter variant. Thus, BloomStore implements the key/value Flash memory storage with a chain architecture of Bloom Filters.

B. Cloud storage

BigTable [9] deploys Bloom Filter to avoid unnecessary disk accesses. The disk access takes huge times. Therefore, Bloom Filter is checked for membership for a query before being accessed to the disk. If Bloom Filter denies, then no hard disk access is performed. In addition, keys are stored in a Sorted String Table (SST). SST stores data in RAM and HDD. However, the access request to SST also a time consuming process. If a data does not exist, then lookup time in SST is wasted. Thus, Bloom Filter enhances drastically the Big Data performance. Similarly, Cassandra also deploys Bloom Filter to avoid unnecessary disk accesses [36]. Moreover, Lahiri et al. [37] deploys Bloom Filter to filter the duplicate data in querying oracle databases.

Kaur and Sood [38] proposes a resource management system in Big Data streaming. Bloom Filter is used to determine the variety of data. The author considers four types of streaming data, namely, text data, audio data, video data, and image data. Four Bloom Filters are deployed to separate the types of data in the Big Data streaming which greatly boosted up the streaming performance, and reduces the memory consumption.

Gessert et al. [39] implements Orestes, Database-as-a-Service, by deploying Bloom Filter to achieve low latency. Bloom Filter is extremely useful in the web browser cache filtering. Hence, Orestes [39] uses Bloom Filter web browser cache filtering to prevent from reading the stale data. Gessert et al. [39] devises a new variant of Bloom Filter, called Bloom-Filter-Bound (BFB) to provide cache consistency to Orestes. The key focus of BFB is to prevent the reading stale data from cache. BFB guarantees to provide recent object versions. BFB enhances the performance of Orestes and ensure cache consistency.

Moreover, Li et al. [12] constructs secure Cloud storage where Bloom Filter is used for trapdoor indexing. The trapdoors are encrypted using secret keys of a user. Keywords are encrypted and then inserted into the indexing. A user computes hash value of a file to be uploaded, and sends it to the Cloud storage server. The user must proof the ownership of the file by answering the challenge given by the server. The Bloom Filter is examined for the existence of the user request.

In addition, Droplet [11] is distributed storage system that employs Bloom Filter for fingerprint indexing. Droplet avoids unnecessary disk lookup by Bloom Filter which boost up the performance. Zhang et al. [11] claims that Droplet is able to avoid 90% of unnecessary disk lookup by deploying Bloom Filter.

Moreover, Bloom Filter is also deployed in Bioinformatics, for instance, Genome project. Melsted and Pritchard [40] deploys Bloom Filter for k-mer DNA sequencing. A single DNA data are very large to store, process and manage. Sequencing DNA is challenging task. $de\ Bruijn$ graph represents k-mer as a node and billions of such k-mer nodes are interconnected to form the $de\ Bruijn$ graph. Bloom Filter is deployed to discover the existence of k-mer in $de\ Bruijn$ graph. All k-mers are inserted into Bloom Filter which is greatly compacted for storing information.

Similarly, Chikhi and Rizk [41] also construct $de\ Bruijn$ graph through deploying Bloom Filter. ABySS 2.0 [22] uses the Bloom Filter for resource efficient assemble of large genomes. Use of Bloom Filter reduces the overall memory requirements and enables to assemble large genomes on a single machine. It is used to represent the de Brujin Graph. Likewise, BLESS [21] is also DNA sequencing using Bloom Filter. However, BLESS re-correct the error caused by false positive for sequencing. However, the false positive has been an open problem for Bloom Filter since its inception.

Moreover, demands of Bloom Filters is increasing for Big Data research. Therefore, Singh et al. [33] designs a Bloom Filter for hosting in Cloud Computing environment, and enables Bloom Filter-as-a-Service. Singh et al. [33] devices a novel fuzzy Bloom Filter, called Fuzzy-folded Bloom Filter (FFBF). FFBF is deployed to design Bloom Filter-as-a-Service. It incorporates a very large amount of the data which enables ease of computing for the Big Data using Bloom Filter.

IV. Future Direction

As we have experienced, SSD/Flash memory is replacing HDD, and thus, HDD is used for backup purposes. The HDD will be completely replaced by SSD/Flash memory in near future. Therefore, the direction of Bloom Filter is designing an in-memory Bloom Filter and Flash-based Bloom Filter. The Flash memory is faster than HDD, however, slower than RAM. But, the RAM space is precious, and a Bloom Filter alone cannot consume entire spaces of RAM. Thus, designing RAM and Flash-based Bloom Filter is a challenge. In addition, Bloom Filter requires few bit update upon insertion of an element. The Flash-based Bloom Filter suffers from such update, because random write consume more time in Flash memory. Thus, it requires a clever way to design such kind of Bloom Filter. As we have discussed, the lookup time also depends on the structure of Bloom Filter, for example, forest structured and chained structured Bloom Filter.

Moreover, Bloom Filter is an unintelligent filter, but it is time to bring in the smart Bloom Filter that can learn. Incorporating machine learning and AI can make Bloom Filter even better than what we expect. The future of Bloom Filter will act as a smart filter, that can identify patterns. On the other hand, Bloom Filter is also deployed interdisciplinary computing area. Moreover, $de\ Bruijn$ graph also be constructed in Flash memory using Bloom Filter which will greatly impact the DNA sequencing performance.

Bloom Filter is inapplicable in exact query requirements. For instance, real-time system. Moreover, Bloom Filter can also be deployed in a password management system which will greatly compact the storage space. However, false positive is the key barrier of the password management system. Moreover, Bloom Filter is also deployed in IoT, Fog Computing, and recent emerging technology to reduce the space consumption. On the other hand, scalability of Bloom Filter is also a big challenge. However, it is addressed using Flash memory. It is observed that there is a challenge in designing scalable in-memory Bloom Filter that is more compact to accommodate a larger number input. Besides, forest structured and chain structured Bloom Filter take high lookup cost. Distributed and parallel approach of Bloom Filter designing is also called for to reduce lookup cost.

V. Conclusion

In this article, we feature the applicability of Bloom Filter in Big Data storage system. We have found that Flash memory is most promising research area for integrating Bloom Filter to build Big Data storage system. We also demonstrated that scalability of Bloom Filter is a serious issue which can be achieved by implementing Bloom Filter in Flash memory. The random write of Flash memory is faster than writing into disk. However, writing a bit to Flash memory requires erasing the entire pages and write it again. Hence, random write is slower in Flash memory than RAM. On the contrary, the RAM size is limited and deduplication requires huge memory space. Therefore, implementing Bloom Filter in Flash memory. Nevertheless, there is a tradeoff between performance and space.

In this review, we find that most of the Big Data storage system deploy Bloom Filter to avoid unnecessary disk access which boost up the system performance dramatically. Moreover, Bloom Filter also enhances the performance of DNA sequencing. The billions of nodes in $de\ Bruijn$ graph is really arduous task to process without Bloom Filter. Finally, we conclude that a tiny data structure takes a lion's share in performance improvement in Big Data.

References

[1] R. Patgiri and A. Ahmed, "Big data: The v's of the game changer paradigm," in *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2016, pp. 17–24.

[2] R. Patgiri, "Issues and challenges in big data: A survey," in *Distributed Computing and Internet Technology*, A. Negi, R. Bhatnagar, and L. Parida, Eds. Cham: Springer International Publishing, 2018, pp. 295–300.

[3] B. H. Bloom, "Space/time trade-o s in hash coding with allowable errors," *Comm. of the ACM*, vol. 13, no. 7, pp. 422–426, 1970.

[4] R. Patgiri, S. Nayak, and S. K. Borgohain, "Preventing ddos using bloom filter: A survey," *EAI Endorsed Transaction on Scalable Information Systems*, 2018, accepted paper.

[5] B. Fan, D. G. Andersen, M. Kaminsky, and M. D. Mitzenmacher, "Cuckoo filter: Practically better than bloom," in *Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies*. New York, NY, USA: ACM, 2014, pp. 75–88.

[6] O. Rottenstreich, Y. Kanizo, and I. Keslassy, "The variable-increment counting bloom filter," *IEEE/ACM Trans. Netw.*, vol. 22, no. 4, pp. 1092–1105, Aug. 2014.

[7] P. S. Almeida, C. Baquero, N. Preguia, and D. Hutchison, "Scalable bloom filters," *Information Processing Letters*, vol. 101, no. 6, pp. 255 – 261, 2007.

[8] F. Grandi, "On the analysis of bloom filters," *Information Processing Letters*, vol. 129, pp. 35 – 39, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020019017301606

[9] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *ACM Trans. Comput. Syst.*, vol. 26, no. 2, pp. 4:1–4:26, Jun. 2008.

[10] B. Debnath, S. Sengupta, J. Li, D. J. Lilja, and D. H. C. Du, "Bloomflash: Bloom filter on flash-based storage," in *2011 31st International Conference on Distributed Computing Systems*, 2011, pp. 635–644.

[11] Y. Zhang, Y. Wu, and G. Yang, "Droplet: A distributed solution of data deduplication," in *Proceedings of the 2012 ACM/IEEE 13th International Conference on Grid Computing*, ser. GRID '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 114–121.

[12] J. Li, X. Chen, F. Xhafa, and L. Barolli, "Secure deduplication storage systems with keyword search," in *2014 IEEE 28th International Conference on Advanced Information Networking and Applications*, 2014, pp. 971–977.

[13] T. Halagan, T. Kováčik, P. Trúchly, and A. Binder, "Syn flood attack detection and type distinguishing mechanism based on counting bloom filter," in *Information and Communication Technology*. Springer, 2015, pp. 30–39.

[14] M. Al-hisnawi and M. Ahmadi, "Qcf for deep packet inspection," *IET Networks*, vol. 7, no. 5, pp. 346–352, 2018.

[15] D. He, D. Zhang, K. Xu, K. Huang, and Y. Li, "A fast and memory-efficient approach to ndn name lookup," *China Communications*, vol. 14, no. 10, pp. 61–69, 2017.

[16] H. Dai, J. Lu, Y. Wang, T. Pan, and B. Liu, "BFAST: High-speed and memory-efficient approach for ndn forwarding engine," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1235–1248, 2017.

[17] F. Lin, G. Wang, J. Zhou, S. Zhang, and X. Yao, "High-performance ipv6 address lookup in gpu-accelerated software routers," *Journal of Network and Computer Applications*, vol. 74, pp. 1 – 10, 2016.

[18] J. H. Mun and H. Lim, "New approach for efficient ip address lookup using a bloom filter in trie-based algorithms," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1558–1565, 2016.

[19] D. Sadhya and S. K. Singh, "Providing robust security measures to bloom filter based biometric template protection schemes," *Computers & Security*, vol. 67, no. Supplement C, pp. 59 – 72, 2017.

[20] J. Bringer, C. Morel, and C. Rathgeb, "Security analysis and improvement of some biometric protected templates based on bloom filters," *Image and Vision Computing*, vol. 58, no. Supplement C, pp. 239 – 253, 2017.

[21] Y. Heo, X.-L. Wu, D. Chen, J. Ma, and W.-M. Hwu, "Bless: Bloom filter-based error correction solution for high-throughput sequencing reads," *Bioinformatics*, vol. 30, no. 10, pp. 1354–1362, 2014.

[22] S. D. Jackman, B. P. Vandervalk, H. Mohamadi, J. Chu, S. Yeo, S. A. Hammond, G. Jahesh, H. Khan, L. Coombe, R. L. Warren, and I. Birol, "Abyss 2.0: resource-efficient assembly of large genomes using a bloom filter," *Genome Research*, vol. 27, 05 2017.

[23] J. Wei, H. Jiang, K. Zhou, and D. Feng, "Dba: A dynamic bloom filter array for scalable membership representation of variable large data sets," in *2011 IEEE 19th Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems*, 2011, pp. 466–468.

[24] G. Lu, B. Debnath, and D. H. C. Du, "A forest-structured bloom filter with flash memory," in *2011 IEEE 27th Symposium on Mass Storage Systems and Technologies (MSST)*, 2011, pp. 1–6.

[25] G. Lu, Y. J. Nam, and D. H. C. Du, "Bloomstore: Bloom-filter based memory-efficient key-value store for indexing of data deduplication on flash," in *2012 IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST)*, 2012, pp. 1–11.

[26] T. Lee, K. Kim, and H.-J. Kim, "Join processing using bloom filter in mapreduce," in *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, ser. RACS '12. New York, NY, USA: ACM, 2012, pp. 100–105.

[27] J. Wang, X. Chen, X. Huang, I. You, and Y. Xiang, "Verifiable auditing for outsourced database in cloud computing," *IEEE Transactions on Computers*, vol. 64, no. 11, pp. 3293–3303, 2015.

[28] Y. Tsuchiya and T. Watanabe, "Dblk: Deduplication for primary block storage," in *2011 IEEE 27th Symposium on Mass Storage Systems and Technologies (MSST)*, 2011, pp. 1–5.

[29] J. Blasco, R. D. Pietro, A. Orfila, and A. Sorniotti, "A tunable proof of ownership scheme for deduplication using bloom filters," in *2014 IEEE Conference on Communications and Network Security*, 2014, pp. 481–489.

[30] W. Dong, F. Douglis, K. Li, H. Patterson, S. Reddy, and P. Shilane, "Tradeoffs in scalable data routing for deduplication clusters," in *Proceedings of the 9th USENIX Conference on File and Stroage Technologies*. Berkeley, CA, USA: USENIX Association, 2011, pp. 2–2.

[31] F. Guo and P. Efstathopoulos, "Building a high-performance deduplication system," in *Proceedings of the 2011 USENIX Conference on USENIX Annual Technical Conference*, ser. USENIXATC'11. Berkeley, CA, USA: USENIX Association, 2011, pp. 25–25.

[32] J. Wang, Z. Zhao, Z. Xu, H. Zhang, L. Li, and Y. Guo, "I-sieve: An inline high performance deduplication system used in cloud storage," *Tsinghua Science and Technology*, vol. 20, no. 1, pp. 17–27, 2015.

[33] A. Singh, S. Garg, K. Kaur, S. Batra, N. Kumar, and K. R. Choo, "Fuzzy-folded bloom filter-as-a-service for big data storage on cloud," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2018.

[34] A. Khare, V. Goyal, S. Baride, S. K. Prasad, M. McDermott, and D. Shah, "Distributed algorithm for high-utility subgraph pattern mining over big data platforms," in *2017 IEEE 24th International Conference on High Performance Computing (HiPC)*, 2017, pp. 263–272.

[35] M. Mitzenmacher, "Compressed bloom filters," *IEEE/ACM Transactions on Networking*, vol. 10, no. 5, pp. 604–612, 2002.

[36] A. Lakshman and P. Malik, "Cassandra: A decentralized structured storage system," *SIGOPS Oper. Syst. Rev.*, vol. 44, no. 2, pp. 35–40, 2010.

[37] T. Lahiri, S. Chavan, M. Colgan, D. Das, A. Ganesh, M. Gleeson, S. Hase, A. Holloway, J. Kamp, T. H. Lee, J. Loaiza, N. Macnaughton, V. Marwah, N. Mukherjee, A. Mullick, S. Muthulingam, V. Raja, M. Roth, E. Soylemez, and M. Zait, "Oracle database in-memory: A dual format in-memory database," in *2015 IEEE 31st International Conference on Data Engineering*, 2015, pp. 1253–1258.

[38] N. Kaur and S. K. Sood, "Efficient resource management system based on 4vs of big data streams," *Big Data Research*, vol. 9, pp. 98 – 106, 2017.

[39] F. Gessert, F. Bcklers, and N. Ritter, "Orestes: A scalable database-as-a-service architecture for low latency," in *2014 IEEE 30th International Conference on Data Engineering Workshops*, 2014, pp. 215–222.

[40] P. Melsted and J. K. Pritchard, "Efficient counting of k-mers in dna sequences using a bloom filter," *BMC Bioinformatics*, vol. 12, no. 1, p. 333, 2011.

[41] R. Chikhi and G. Rizk, "Space-efficient and exact de bruijn graph representation based on a bloom filter," *Algorithms for Molecular Biology*, vol. 8, no. 1, p. 22, 2013.

Efficient Iris Pattern Recognition Method by using Adaptive Hamming Distance and 1D Log-Gabor Filter

Rachida Tobji [1], Wu Di [2]*, Naeem Ayoub [3], Samia Haouassi [4]
School of Computer Science and Technology
Dalian University of Technology, Dalian 116000, China

***Abstract*—Iris recognition is one of the highly reliable security methods as compared to the other bio-metric security techniques. The iris is an internal organ whose texture is randomly determined during embryonic gestation and is amenable with a computerized machine vision system for the remote examination. Previously, researchers utilized different approaches like Hamming Distance in their iris recognition algorithms. In this paper, we propose a new method to improve the performance of the iris recognition matching system. Firstly, 1D Log-Gabor Filter is used to encode the unique features of iris into the binary template. The efficiency of the algorithm can be increased by taking into account the coincidence fragile bit's location with 1D Log-Gabor filter. Secondly, Adaptive Hamming Distance is used to examine the affinity of two templates. The main steps of proposed iris recognition algorithm are segmentation by using the Hough's circular transformation method, normalization by Daugman's rubber sheet model that provides a high percentage of accuracy, feature encoding and matching. Simulation studies are made to test the validity of the proposed algorithm. The results obtained ensure the superior performance of our algorithm against several state-of-the-art iris matching algorithms. Experiments are performed on the CASIA V1.0 iris database, the success of the proposed method with a genuine acceptance rate is 99.92%.**

Keywords*—Iris recognition; bio-metric; Hamming Distance; iris recognition matching; Adaptive Hamming Distance; 1D Log-Gabor Filter; segmentation; normalization; feature encoding; genuine acceptance rate

I. Introduction

Bio-metric security techniques are comprehensive techniques that establish a secure identity of a person by measuring one or more types of physical characteristics. Bio-metric based techniques are currently favored by a fashion phenomenon, mainly film and television reported in [1]. Thus, it is not uncommon to see retinal scanners with beautiful red lasers, readers fingerprints with very nice flashing lights, etc. All this becoming the ultimate in access control technology. There are some bio-metric techniques that have been used from the last few decades but they have some limitations. In this study we focus on iris bio-metric security technique which is the most reliable and accurate among all bio-metric traits at present. It has been proved that the probability of finding two identical irises is less than the reverse of the number of humans who have lived on earth. Fig. 1 shows structure of human eye.

Iris recognition is based on interpretation and comparison information between two irises of the same person, belonging to two individuals who may be the same person or not. In

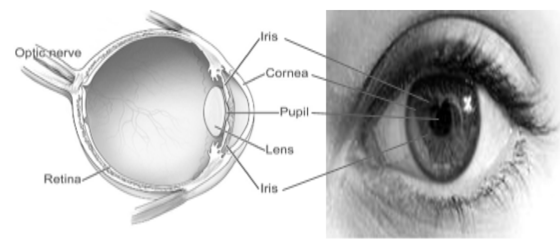


Fig. 1. Structure of human eye.

the first case, there must be authentication, but authentication should not be performed in the second case. Selection of iris images may have different size and orientation between several photos, it is therefore necessary to use information that is independent of the parameters of the image. It is necessary to eliminate the pupil, the white of the eye, perhaps the reflections and the area of the eyelid in the case of significant encroachment on the iris. This operation will make subsequent processing more reliable, as compares iris information only, and in addition gain memory occupancy when saving code vectors.

In an eye, iris is the colored area between the pupil and the white of eye. The iris has diameter smaller than a hair which consists of a network of the fine tubes. Iris formation in the human eye begins in the third month and the structure of the distinctive elements finishes in the eighth month of gestation. The pigmentation process continues in the first years of birth. The formation of the iris is chaotic which generates patterns with high variability.

There are two main stages of iris recognition, first, the features of are recorded in the enrollment process. Second, in matching process, the authentication system attempts the confirmation of an individual's identity by comparing a submitted sample to previously enrolled templates.

The remaining sections of this paper are organized as follows: Background is described in section 2. In section 3 related work is briefly discussed regarding iris bio-metric systems based on Hamming Distance with feature encoding schemes adopted by different researchers. Section 4 proposed method and provides a detailed description of iris recognition system components with pre-processing i.e., image acquisition, segmentation and normalization, feature extraction and encoding, matching. Experimental results and discussion are presented in section 5. Finally, conclusion is made in section

6.

II. Background

The American ophthalmologist Frank Burch was the first person who had officially realized the possibilities of the texture of the iris as an identification tool. He proposed this method in a lecture that was delivered in American Academy of Ophthalmology in 1936. Before that, the iris was mainly considered only for its color. Burch's idea has been reproduced in the ophthalmology manuals, but there was very little research done, due to particular lack in the field of engineering [2]. Professor John Daugman developed a mathematical approach to the analysis of patterns of the iris. The collaboration between Daugman, Sarin and Flom was resulted in a functional prototype patented in 1994. This system has been greatly improved. Several companies used licenses of iris recognition algorithms in their applications. The patent is now the property of the company iridian. Daugman's work was based on the non-orthogonal Gabor complex wavelets, the complex-valued filters were applied to the texture of the two-dimensional iris and the phase information was extracted to form the signature [3], [4]. Wildes [5] presented a system based on a pyramidal representation multi-resolution iris textures. The idea behind the image pyramid is as follows: an image can be represented as a sub-sampled approximation and one or more images residuals at different resolutions. In Wildes's system four sub-images were used from a multi-resolution pyramid as a signature and a standardized correlation was measured for classification. Boles et al. [6] proposed an algorithm based on "Zero-Crossings" of a one-dimensional orthogonal wavelet transform. First, the one-dimensional signal was transformed and acquired by recording the gray scale values of one or several concentric circles of the iris. Then the radius of the circle was normalized with respect to the radius of the iris, ensuring that the same points of the texture of the iris sampled according to size of the iris in the picture. Zhu et al. [7] have done comparative studies between wavelets non-orthogonal Gabor and orthogonal wavelet of Daubechies of order 4 ("db4"). In this method, Gabor wavelets multi-directional and different frequencies were used to separate the texture of the iris in different frequency sub-bands. The average and standard deviation were used as a signature. In the case of wavelets, the mean and the standard deviation of the sub-bands were generated from a transformation on five levels and used as signature. The result was very much in favor of the wavelet analysis of Gabor. Lim et al [8] used an orthogonal wavelet transform with a simple Haar Wavelet. Four levels of transformations were calculated first and the signature was composed based on the coefficients of the details diagonals at the lowest scale. Each average value of the diagonal detailed the coefficients to the other three sub-bands at higher. The signature coefficients were all of zero mean and truncated to only one bit as a function of their sign. Tisse et al. [9] used the Hilbert transform for iris detection. In which frequency emergent and the instantaneous phases (combination of the original signal and the Hilbert transform) were used to generate the signature. Then Hamming Distance was calculated to compare two signatures. Rydgren et al. [10] used a wavelet packet transform on three levels and applied it to the unrolled images. The sub images in the wavelet packet tree which holds information about both frequency and location were calculated as the suitable candidates for iris signatures. An energy measure was used to identify the particular packets that carries discriminating information about the iris texture. Noh et al. [11] used a method called multi-resolution ICA by calculating the value in average gray levels as a function. They construct a one-dimensional iris signature which served as input to the multi-resolution ICA algorithm. In addition, a presentation of some current methods as well as a very recent global study that covers development history and current state-of-the art for understanding the recognition of the iris was also presented in literature [12]. Nabti and Bouridane [13] introduced a new approach to Multi-Scale contour detection for iris segmentation. Lionel [14] proposed a new iris recognition system to improve safety by using a technique of portable object. He studied a human-machine based on gesture recognition. In a first portion, he evaluated grid photo sensors from modeling up to the characterization of pixels and then tested the vehicle in CMOS technology that allowed them to compare 24 pixels to photo grid. He also addressed research on the integration of video stabilization on a sensor of images, key points on portable objects and on near infrared filtering. Wang et al. [15] introduced an eyelash detection method, in which a combination of Expectation Maximization (EM) and Gaussian Mixture Model(GMM) was used to extract the eyelash and eyelids information. Wang et al. [16] used random transform with polynomial curve fitting and least squares fitting to detect the eyelid and eyelashes. They used diagonal gradient and thresholding method to get the information of eyelashes .

Several iris recognition techniques have been developed by employing the traditional iris recognition systems [17], [18], [19], [20], feature value extraction [3], [21], [22], [23], [24] and iris matching [25], [26]. Most of state-of-the-art methods used simple binarization based on the sign of feature values [3],[21], [22],[27], [28].

III. Related Work

In iris localization methods, an integro-differential operator was used by Daugman [21], which is as follows:

$$max_{(r,x_0,y_0)} \mid G_\sigma(r) * \frac{\partial}{\partial_r} \oint_{r,x_0,y_0} \frac{I(x,y)}{2\Pi r} ds \mid \qquad (1)$$

Where $I(x, y)$ is an image containing an eye. By equation (1), maximum angular integral of radial derivative over the image domain, the outer (limbus) boundary of the iris and pupillary boundary location can be realized. Firstly, the outer limbus boundary location can be estimated, to avoid the eyelids occlusion, the limit of angular arc of contour is bounded in range to $90°$ centered on the horizontal meridian. Secondly, a finer convolution scale σ find pupillary boundary. Wildes [5] proposed a method to estimat the iris boundaries by adopting the hough transform. Masek [29] also used hough transform method for iris boundary estimation. These two methods can efficiently estimate the iris boundaries. In [12] and [30], a different method was introduced by using the active contours to find the non-circular shape of the iris which gives better representation of circle. This method does not perform well in some cases because if the edges details are weak then the desired iris boundary can not be evaluated according to the iris pattern. A threshold method has also been adopted by Somnath

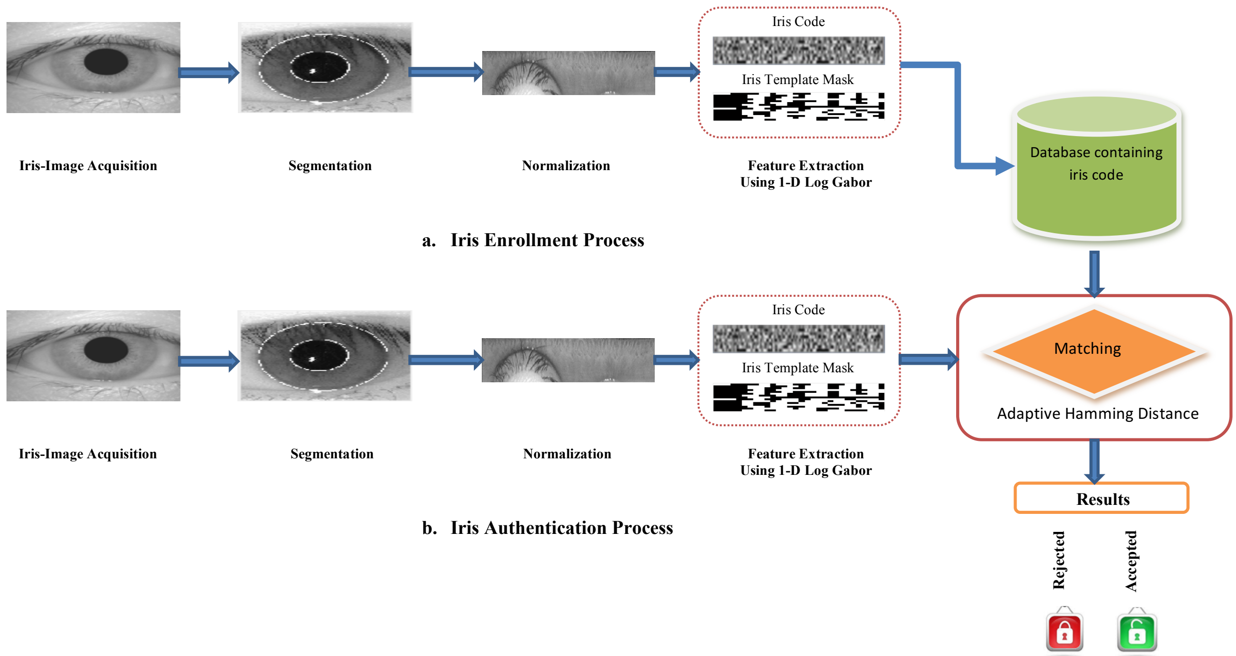


Fig. 2. Flow diagram proposed iris biometric system.

et al. [31] to detect boundary of the pupil in the eye image areas as it is darker than the other. Different state-of-the art techniques used fourier methods to account eyelid occlusions and iris segmentation. Avila [32] presented a similar type of system, in which he used a zero-crossing discrete dyadic wavelet transform which showed a high level of accuracy. Li Ma et al. [33] proposed to use circular symmetry filters to capture local-texture information of the iris. Then they utilized this information to construct feature vector the fixed length. For instance, Tisse et al. [34] proposed new approach for extraction and localization of iris by implementing C language and tested individually the performance of the different processing. Rai et al. [35] used Support Vector Machine (SVM), Hamming Distance, parabola detection and trimmed median filter to detect and remove the eyelid, and eyelashes from the iris image. Soliman et al. [36] used a coarse-to-fine algorithm to address the computational cost problem and the integro-differential operator to get information aboutpupil centers and radii. In [37] Dehkordi et al. used Adaptive Hamming Distance to improve the performance of iris code matching stage. The expanding and adjoining behavior of Hamming subsets to the right or left neighboring bits increased the accuracy of Hamming distance computation.

IV. Proposed Method

In this paper, we prpose an approach for iris pattern recognition by Adaptive Hamming Distance with improving the speed and accuracy of the process. we use a statistical method to account eyelashes presented in litirature [12] and 1D Log-Gabor Filter features to uniquely identify iris. We have studied various well known algorithms for iris recognition [29],[32], [33], [34], [35], [36], [37] and compared the results with state-of-the art algorithms. Fig. 2 shows flow diagram of iris biometric system which is described in detail in the following subsection.

A. *Iris Recognition System*

The proposed iris recognition system based on the five main steps. In the first step, the person's eye enrollment is performed. In the second step, iris is segmented from the image to locate and study the iris pattern. In the third step, we normalize and scaled the iris pattern to a predefined size. In forth step, we extract the features of iris code from filtered iris details, obtained from generated iris templates. In the fifth step, we compute the similarity score of two iris codes by matching them with Adaptive Hamming Distance scheme. For generating the base templates, we use the same methods as introduced in [29] with the Adaptive Hamming Distance for handling the images of CASIA V1.0 [38]. The same methodology adopted by Daugman [21] in the matching by metric employed and to extract textural information of iris image. Fig. 3 shows the steps of an iris recognition system.

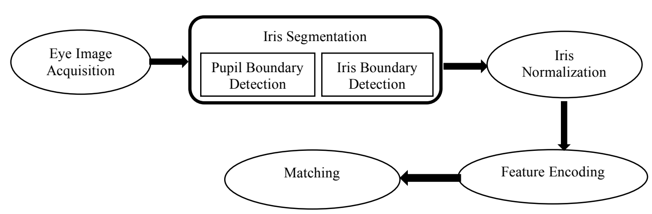
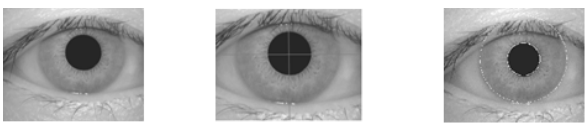


Fig. 3. Typical main steps of an iris recognition system.

The detailed mathodology is given in the subsubsections below.

1) Image Acquisition: Iris image acquisition is an important step in the recognition system. In this step images are acquired using CCD camera. Images having good resolution and sharpness need to maintain with adequate intensity. For the purpose of fair evaluation of our proposed algorithm, we use CASIA iris image database (version 1.0).

2) Segmentation and Normalization : For the purpose of segmentation, we isolate the zone of the iris of the eye (Fig. 4-a), then horizontal direction and vertical direction are estimation for segmentation. We use the description of the function which locates the center of the image. Once we get both directions, it becomes easy to detect the center (Fig. 4-b). To detect the centers, Hough transform is applied to the circular forms to detect the border of iris / pupil. For effective and precise circle, first Hough transform was executed for the border of iris / sclera then for the border of iris / pupil (Fig. 4-c). Fig. 4 shows steps of Segmentation.



(a) Iris base. (b) Horizontal and Vertical Diameters. (c) Located Iris.

Fig. 4. example of steps of Segmentation: (a) Iris base, (b) Horizontal and Vertical Diameters, (c) Located Iris.

Once the iris region is successfully segmented, we normalize the iris segmented area by transforming the representation space of the iris with rectangular space to get fixed dimensions for authentication of the data that has the same dimension as shown in Fig. 5.

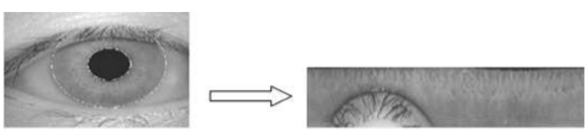


Fig. 5. Normalization of the localized iris.

In the normalization process, we unwrap the iris image and convert it to its polar equivalent by using Daugman's Rubber sheet model. The points on the Cartesian scale to the polar scale of center of the pupil are converted through a remapping formula. The remapping of iris image I(x, y) from raw Cartesian coordinates to polar coordinates (r, θ) can be calculated as follows

$$I(x(r,\theta), y(r,\theta)) \longrightarrow I(r,\theta) \quad (2)$$

Where r is on the interval $[\theta, 1]$ and θ is angle $[\theta, 2\pi]$, with:

$$x(r,\theta) = (1-r)x_p(\theta) + r x_l(\theta) \quad (3)$$

$$y(r,\theta) = (1-r)y_p(\theta) + r y_l(\theta) \quad (4)$$

Where $x_p(\theta)$,$y_p(\theta)$ and $x_l(\theta)$, $y_l(\theta)$ are the coordinates of the pupil and iris boundaries along the direction θ.

Daugman's Rubber sheet model is shown in Fig. 6.

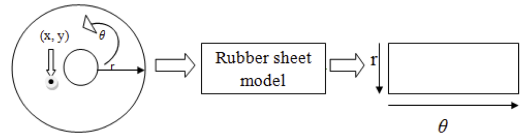


Fig. 6. Daugman's Rubber sheet model for the normalization of the iris.

3) Feature Extraction and Encoding: In this step, a template that represent the iris pattern information is generated by using a 1D log-Gabor filter. The error between the two different images occur when the pixels' intensity of two different iris images directly compared. To diminish this difficulty, 1D Log-Gabor filter is used same as [39], [40]. 1D Log-Gabor filter can be calculated by equation (5) as follows :

$$G_{(X,Y)} = e^{-\frac{log(\frac{radius}{fo})^2}{2*log(sigmaOnf)^2}} \quad (5)$$

To extract and encode image texture information, the filters are multiplied by the raw image pixel data and coefficients are generated by integrating them over their supported domain. A noise mask is then generated with the feature template to mark the corrupted bits in the template. Fig. 7 shows feature encoding of iris.

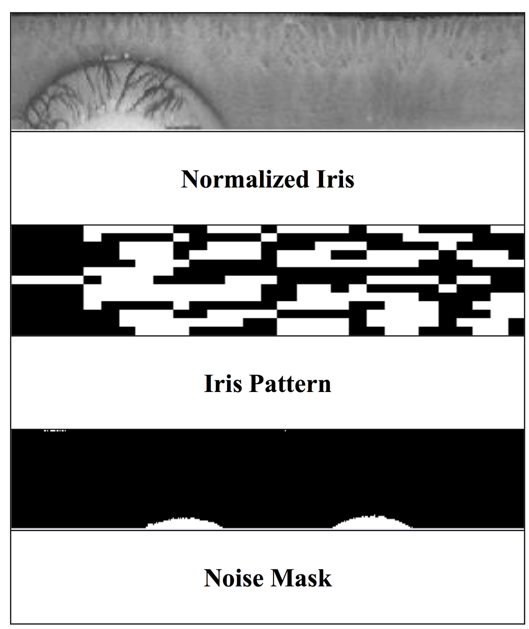


Fig. 7. Feature encoding.

4) Matching: Matching step consists of all the previous steps that are performed through out the enrollment of the encoded iris template. By extracting the binary form of the encrypted bit pattern, a simple boolean "XOR" operation can be used to match the stored encrypted bit patterns [41]. The dissimilarity measure is computed between any two iris bit patterns by utilizing the Adaptive Hamming Distance. The

Hamming distance can be defined as follows

$$HD = \frac{1}{N}\sum_{j=1}^{N} X_j \oplus Y_j \tag{6}$$

Where, N represents the total number of the bits per bit pattern. HD represents the fractional measure of dissimilarity and 0 will indicate the perfect match. The Adaptive Hamming Distance can be calculated by the following formula

$$HD = \frac{\|(Code_A \cap Mask_A) \oplus (Code_B \cap Mask_B)\|}{\|Mask_A \cup Mask_B\|} \tag{7}$$

In this equation (7), $Code_A$ and $Mask_A$ represent the iris code of unidentified individual *A*. $Code_B$ and $Mask_B$ represents kits of template *B* of bitwise iris code and mask vector, which contains 20*480=9600 template bits and 20*480=9600 mask bits, respectively, that are calculated by using "XOR" and "AND" boolean operators. $\|Mask_A \cup Mask_B\|$ denotes the total number of comparisons. Algorithm 1 shows the general steps of matching method.

Algorithm 1: Iris Matching Algorithm

```
Input: Normalized iris image
Initialization: while Iris Code matching do
    for calculate Bitwise Mask A do
        for Ckeck bitwise Mask B stored in database
         do
            check iris code B stored in database
        produce iris code A
    set the initial size of Hamming Divide the iris
     code into Hamming subsets (subset A, subset B)
    L = 8|16|32;
    i = 0;
    k = search engine;
    if density >= 0 then
        HD = subset A & subset B
        i = = i + L;
        break;
        k=search engine;
        for density = 1 + i;
         do
            if density >= 0 then
                HD = neighbore subset A + L
                subset = subset+1;
                break;
                k = search engine;
                i = i + L;
Output: Authentication
```

In order to calculated Adaptie Hamming Distance, First, bitwise iris code based Hamming space is divided into Hamming subsets. The initial constant length of Hamming subsets can be the size of 8,16 or 32 bits [42], [43]. Then both iris codes is divided into Hamming subsets with initialization size of *L*. The density value of the masked bit in the Hamming subset can be calculated calculate which is calculated as follows

$$Mask(i,j) = \begin{cases} 1 & \text{correct bit} \\ 0 & \text{trival bit} \end{cases}$$

Here the vector *Mask* represents the trivial bit location in the both iris codes. Where *Mask(i,j)* = 0 represents *Mask* bit, which is excluded from computation and *Mask(i,j)* = 1 correct bit to be compared.

The number of trivial bits are presented in hamming subset as:

$$P_{(i,j)} = L - \sum_{m=1}^{L} Mask(i+m,j) \tag{8}$$

Where hamming subset is equal to $\frac{L}{2}$ or less than adaptive hamming subset. The HD is calculated from the one subset to the neighbor subset, it will indicate that the trivial bits in Hamming subset are larger and the HD is not accurate if it is more than $\frac{L}{2}$. In this step, the search engine verify the density of mask bits in the neighbor hamming subset. If density is low that means the subsets are not affected by noise. Hence, the hamming subsets can expand to the correct bits and these bits will be added to left or right neighbor subset. The location of masked bits can be estimated as follows

$$Left_{(i,j)} = L - \sum_{m=1}^{\frac{L}{2}} Mask(i+m,j) \tag{9}$$

$$Right_{(i,j)} = L - \sum_{m=\frac{L}{2}}^{L} Mask(i+m,j) \tag{10}$$

The correct bits will be added to the subset at the left or right to next hamming subset and will be increased as

$$Left_{(i,j)} \Longrightarrow L \leftarrow L + Left(i,j) \tag{11}$$

$$Right_{(i,j)} \Longrightarrow L \leftarrow L + Right(i,j) \tag{12}$$

Hence, the computed Hamming Distance at correspondent bit of subset will be updated by the addition of the correct bits to left neighbor hamming subset, .

V. Experimental Results and Discussion

A. Dataset

Experiments are performed on the CASIA Iris database v1, which contains 756 iris images of 108 different people. Totally, 756*755/2 = 285390 pairs of comparisons for each algorithm, 2268 for Intra-Class comparisons and 283122 for Inter-Class comparisons. Which contains 9600 template bits and also for mask bits, respectively.

For the evaluation of our algorithm, we employ False Acceptance Rate (FAR), False Rejection Rate (FRR), Equal Error Rate (EER) and Receiver Operating Curves (ROC).

In contrast to the degree of match, two error rates FRR and FAR are calculated for selecting a proper point as a threshold for different Adaptive HD using the formula as in equation (13) and equation (14).

$$FAR = \frac{No.\ of\ time\ of\ different\ imposter}{No.\ of\ comparisons\ total} \tag{13}$$

$$FRR = \frac{No.\ of\ time\ of\ different\ genuine}{No.\ of\ comparisons\ total} \tag{14}$$

The error rate can be minimized by selecting the threshold value on the intersection of these two error rates curves as shown in Fig. 8 with EER =0.004. If Hamming Distance is smaller then FAR will reduces and FRR increase, FAR will be increased and FRR will be decreased if HD increases.

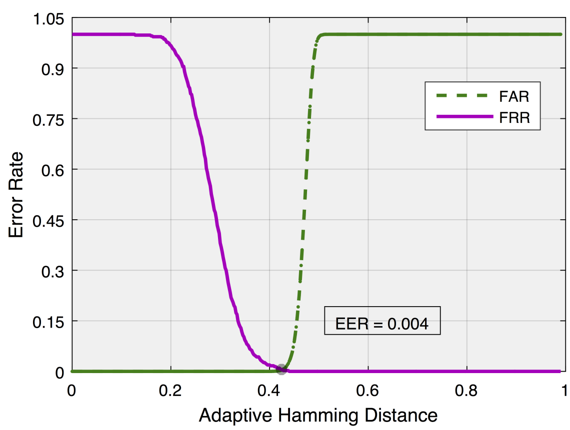


Fig. 8. The Equal Error Rate Value where FAR and FRR are equal.

Fig. 9 shows comparisons of Receiver Operating Curves (ROC) and recognition rate.

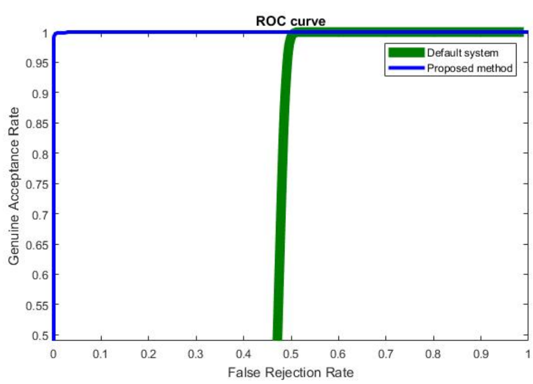


Fig. 9. ROC curves for the system for different recognition rate.

The Table I shows the comparisons performance of the proposed method in terms of recognition rate, False Reject Rate (FRR). The recognition rate of 99.92% and FRR = 0.00 shows the success of method.

B. Intra-Class Comparisons

In our work 2268 Intra-Class comparisons of iris templates were successfully performed and their Adaptive Hamming Distance distribution is shown in Fig. 10.

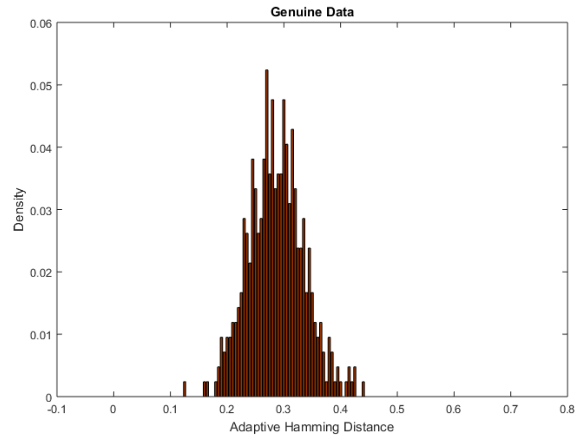


Fig. 10. Intra-Class Adaptive Hamming Distance distribution.

C. Inter-Class Comparisons

On our work 283122 Inter-Class comparisons of iris templates were executed successfully and the histogram of distribution is shown in Fig. 11.

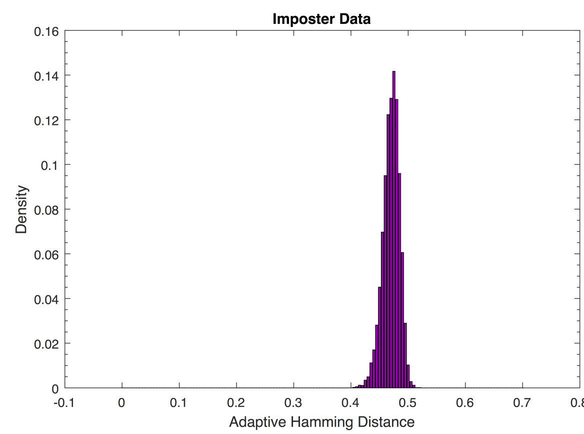


Fig. 11. Inter-Class Adaptive Hamming Distance distribution.

D. Intra-Class and Inter-Class comparisons

In the Intra-Class and Inter-Class comparisons very few values were seen that overlap each others. Fig. 12 shows the combined Adaptive HD distribution comparison for 756*755/2=285390 pairs.

VI. Conclusion

In our proposed work, Adaptive Hamming Distance is used with the 1D Log-Gabor Filter. Adaptive HD can expand left or right based on the masked bits density. Search engine process the neighboring subsets if the Hamming subset reaches to threshold. Adaptive behavior can give more efficient results than simple HD. To produce a bit-wise biometric template, the normalized iris regions are convolved with the 1D Log-Gabor filter. Results of the proposed method on CASIA V1.0 show 99.92% best performance methods with 0% the state of the art iris recognition FRR, an EER down to 0.004%

TABLE I. QUANTITATIVE COMPARISION OF RESULTS

Methods	Database	Matching Process	Feature Extraction	FAR	FRR	EER	Overall Accuracy	Correct Regnition Rate
Masek [29]	Dataset	HD with XOR	2D Gabor	0.005	0.238	0.35	99.75	-
Avila [32]	Database	HD	Zero-Crossing	0.03	2.08	0.21	97.89	-
LI Ma et al. [33]	Database	Expanded binary Feature vector & Exclusive OR operations	Classes of 1D Wavelets i.e, 1-D intensity signals	0.02	1.98	0.29	98.00	99.60
Tisse et al. [34]	Database	HD	2D Gabor	1.84	8.78	0.41	89.37	-
Rai et al., [35]	CASIA Database	SVM and HD with XOR	1-D Log-Gabor wavelets	0.07	0.33	-	99.60	-
Soliman et al., [36]	CASIA-V3 CASIA-V1	HD with XOR	1-D Log-Gaborwavelets	-	-	-	98.80	-
Dehkordi et al., [37]	CASIA-V3	Adaptive HD with XOR	2-D Log-Gabor wavelets	-	0.06	-	-	99.96
Proposed	CASIA-V1	Adaptive HD with XOR	1-D Log-Gabor Filter	0.08	0.00	0.004	99.92	99.996

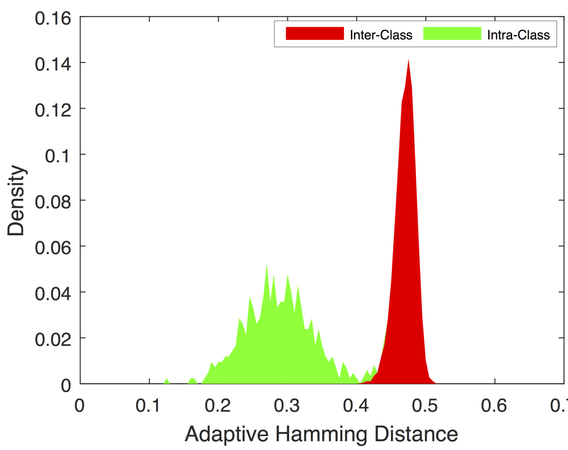


Fig. 12. Intra-Class and Inter-Class Adaptive Hamming Distance distribution.

and with different class of variations i.e. Inter-Class Adaptive HD distribution and Intra-Class Adaptive HD distribution. In future, PCA and FLDA can be incorporated with this model to get better results.

ACKNOWLEDGMENT

The work presented in this paper was supported by the National Natural Science Foundation of China under Grant no. 61370201

REFERENCES

[1] N.P. Joseph, "Challenges and Opportunities. Washington," In Biometric Recognition; I.M. Lynette, Eds.; DC: The National Academies Press: Washington, 2010. [CrossRef]

[2] L. Ma, Y. Wang and T. Tan, "Iris Recognition Based on Multichannel Gabor Filtering," ACCV2002: The 5th Asian Conference on Computer Vision, pp. 23-25, January 2002.

[3] J. Daugman, "How Iris Recognition Works," IEEE Transactions on Circuits and Systems for Video Technology, pp. 21-30, 2004. [CrossRef]

[4] J. Daugman, "Probing the Uniqueness and Randomness of IrisCodes: Results From 200 Billion Iris Pair Comparisons," Proceedings of the IEEE, November 2006. [CrossRef]

[5] R. P. Wildes, "Iris recognition: an emerging biometric technology," Proceedings of the IEEE, pp. 1348-1363, September 1997. [CrossRef]

[6] W. W. Boles, B. A. Boashash, "Human identification technique using images of the iris and wavelet transform," IEEE Transactions on Signal Processing, pp. 1185-1188, 1998. [CrossRef]

[7] Z. Yong, T. Tieniu and W. YunHong, "Biometric personal identification based on iris patterns," Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, Espagne, 2000. [CrossRef]

[8] S. L. Lim, K.L. Lee, O. B. Byeon, et al., "Efficient Iris Recognition through Improvement of Feature Vector and Classifier," ETRI Journal, pp. 61-70, 2001. [CrossRef]

[9] T. Christel-Loic, M. Lionel, T. Lionel, et al., "Person Identification Technique Using Human Iris Recognition," The 15th International Conference on Vision Interface, pp. 294-299, 2002.

[10] E. Rydgren, T. Ea, F. Amiel, et al., "Person Identification Technique Using Human Iris Recognition," IEEE International Conference on Image Processing (ICIP'04), Singapore, 2004.

[11] S. Noh, K. Bae, K. Park, et al., "A New Iris Recognition Method Using Independent Component Analysis," IEICE Transactions on Information and Systems, pp. 2573-2581, 2005.

[12] J. Daugman, "New Methods in Iris Recognition," IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics), pp. 1167-1175, 2007. [CrossRef]

[13] M. Nabti, L. Ghouti and A. Bouridane, "New An effective and fast iris recognition system based on a combined multiscale feature extraction technique," Pattern Recognition, pp. 868-879 2008. [CrossRef]

[14] M. Lionel, "from the pixel to the integrated iris recognition system," University Paul Cezanne (Aix-Marseille III), Marseille, 2008.

[15] T. Wang, M. Han, H. Wan, et al., "A robust and fast eyelash detection basted on expectation maximization and Gaussian mixture model," Springer, pp. 9397, 2011. [CrossRef]

[16] A. Walid, K. Lotfi and M. Nouri, "A fast and accurate eyelids and eyelashes detection approach for iris segmentation," International conference on IEEE, Sousse, Tunisia, June 2013. [CrossRef]

[17] Z. He, T. Tan, Z. Sun, et al., "Toward accurate and fast iris segmentation for iris biometrics," IEEE Transactions, pp. 1670-1684, 2009. [CrossRef]

[18] Y. Hu, K. Sirlantzis and G. Howells, "Improving color iris segmentation using a model selection technique," Pattern Recognition, pp.24-32, 2015. [CrossRef]

[19] D. S. Swati, C. RajaBhushnam, "Review of IRIS recognition techniques" 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, India, Feb 2017. [CrossRef]

[20] A. M. A. Musab, M. T. Nooritawati, "Cancelable biometrics technique for iris recognition," 2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Malaysia, April 2018. [CrossRef]

[21] J. Daugman, "High confidence visual recognition of persons by a test of statistical independence," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1148-1161, November 1993. [CrossRef]

[22] Z. Sun, T. Tan, "Ordinal measures for iris recognition," IEEE Transaction, pp. 2211-2226, 2009. [CrossRef]

[23] L. Ma, T. Tan, Y. Wang, et al., "Efficient Iris Recognition by Characterizing Key Local Variations," IEEE Transactions on Image Processing, pp. 739-750, 2004. [CrossRef]

[24] C. W. Tan, A. Kumar, "Accurate iris recognition at a distance using stabilized iris encoding and Zernike moments phase features," IEEE Transaction, pp. 3962-3974, 2014. [CrossRef]

[25] K. P. Hollingsworth, K. W. Bowyer and P. J. Flynn, "The best bits in an iris code," IEEE Transaction, pp. 964-973, 2009. [CrossRef]

[26] Y. Hu, K. Sirlantzis and G. Howells, "Exploiting stable and discriminative iris weight map for iris recognition under less constrained environment," IEEE International Conference on Biometrics, Theory, Arlington, VA, USA, September 2015. [CrossRef]

[27] C. W. Tan, A. Kumar, "Efficient and accurate at a distance iris recognition using geometric key-based iris encoding," IEEE Transaction, pp. 1518-1526, 2014. [CrossRef]

[28] P. Li, G. Wu, "Iris recognition using ordinal encoding of log-Euclidean covariance matrices," IEEE International Conference on Pattern Recognition, Tsukuba, Japan, November 2012.

[29] L. Masek, "Recognition of Human Iris Patterns for Biometric Identification," University of Western Australia, Australia, 2002.

[30] S. Shah, A. Ross, "Iris segmentation using geodesic active contours," IEEE Signal Processing Society, pp. 824-836, 2009. [CrossRef]

[31] S.Dey, D. Samanta, "A novel approach to iris localization for iris biometric processing," International Journal of Biological and Medical Sciences, 2008.

[32] C. Sanchez-Avila, R. Sanchez-Reillo and D. De Martin-Roche, "Iris recognition for biometric identification using dyadic wavelet transform zero-crossing," IEEE 35th International Carnahan Conference on Security Technology, Quebec City, London, England, UK, October 2001.[CrossRef]

[33] L. Ma, Y. Wang and T. Tan, "Iris Recognition Based on Multichannel Gabor Filtering," Proc. The 5th Asian Conference on Computer Vision, January 2002.

[34] T. Christel-Loic, M. Lionel, T. Lionel, et al., "Person Identification Technique Using Human Iris Recognition," The 15th International Conference on Vision Interface, 2002.

[35] H. Rai, A. Yadav, "Iris recognition using combined support vector machine and Hamming distance approach," Expert Systems with Applications, pp. 588-593, 2014. [CrossRef]

[36] N. F. Soliman, E. Mohamed and F. Magdi, "Efficient iris localization and recognition," Optik-International Journal for Light and Electron Optics, pp. 469-475, 2017. [CrossRef]

[37] A. B. Dehkordi, S. A. R. Abu-Bakar, "Iris code matching using adaptive Hamming distance," 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Malaysia, 2015. [CrossRef]

[38] CASIA V.1.0 iris image database version one. http://www.cbsr.ia.ac.cn

[39] P. Yao, J. Li, X. Ye, et al., "Iris recognition algorithm using modified log-Gabor filters," International Conferenece on Pattern Recognition, Hong Kong, China, August 2006. [CrossRef]

[40] L. Masek, v. Kovesi, "MATLAB Source Code for a Biometric Identification System Based on Iris Patterns," School of Computer Science and Software Engineering, University of Western Australia, 2003. http://www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html6

[41] A. Ross, D. Nandakumar and A. K. Jain, "Biometric personal identification system based on iris analysis," in Handbook of Multibiometrics, Springer, Heidelberg, 2006.

[42] K. W. Bowyer, K. P. Hollingsworth and P. J. Flynn, "Handbook of iris recognition," A survey of iris biometrics research, 2008-2010, Springer, pp. 15-54, 2013.

[43] K. W. Bowyer, "The results of the NICE.II Iris biometrics competition," Pattern Recognition Letters, pp. 965-969, 2012. [CrossRef]

Empirical Evaluation of SVM for Facial Expression Recognition

Saeeda Saeed[1], Junaid Baber[1], Maheen Bakhtyar[1], Ihsan Ullah[1], Naveed Sheikh[2], Imam Dad[1], Anwar Ali Sanjrani[1]
[1] Department of Computer Science & IT, University of Balochistan
[2] Department of Mathematic, University of Balochistan

***Abstract*—Support Vector Machines (SVMs) have shown better generalization and classification capabilities in different applications of computer vision; SVM classifies underlying data by a hyperplane that can separate the two classes by maintaining the maximum margin between the support vectors of the respective classes. An empirical analysis of SVMs on the facial expression recognition task is reported with high intra and low inter class variations by conducting an extensive set of experiments on a large-scale Fer 2013 dataset. Three different kernel functions of SVM are used; linear kernel, quadratic kernel and cubic kernel, whereas, Histogram of Oriented Gradient (HoG) is used as a feature descriptor. Cubic Kernel achieves highest accuracy on Fer 2013 dataset using HoG.**

Keywords*—*Facial Expression Recognition; Support Vector Machine (SVM); Histogram of Oriented Gradients (HoG)

I. Introduction

Facial expression is one of the most important non-verbal means of communication, enabling human beings to exchange the social information. There is a vast range of applications of facial expression recognition (FER) such as human and man-machine interaction [1], smart healthcare [2], biometrics, medical diagnosis [3], surveillance [4] and mental state identification [5].

However, high intra and low inter-class variations make human FER more challenging task in the field of machine learning and computer vision. Based on emotions and social interactions, facial expressions are usually categorized into six basic emotions, namely, happiness, sadness, anger, fear, surprise, and disgust [6]. The expression *neutral* is also added as another facial expression and widely accepted by the researchers [7]. Figure 1 represents some sample images of these emotions/facial expressions.

In recent years, several interesting solutions have been proposed for FER, relying on the diversified set of techniques and strategies to better represent and classify facial expressions in various application domains [5], [8]. Most of the existing FER frameworks focus on algorithms to extract better features, such as Gabor filter and RBF network [7], CNN [9], hybrid CNNSIFT aggregator [10], and SVM [11].

In this paper, empirical study of SVMs with different kernels and features on FER is reported. In experiments, the analysis of false positives for each expression are also discussed. The rest of the paper is organized as follows: section II discusses the related work. The detailed description of the methodology adopted for this empirical study is provided in the section III. The section IV reports the experiments, the dataset used for the evaluations along with the detailed analysis of the obtained results. Finally, section V provides some concluding remarks.

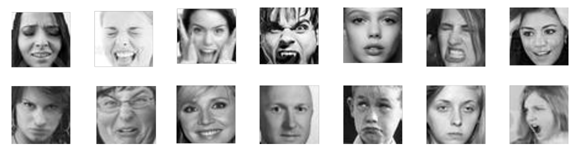


Fig. 1. Sample images from the Fer 2013 dataset. Images in the same column show identical expressions. The labels of expressions from left to right are: *Anger, Disgust, Fear, Happy, Sad, Surprise* and *Neutral*.

II. Related Work

In recent years, the research community has shown great interest in FER. As a result, several interesting solutions have been proposed for FER. For instance, Usman et al. in [12] proposed a three step solution to FER. In the first step, a state-of-the-art Viola-Jones face detection method has been employed to detect faces in the images. Subsequently, HoG features are extracted from the identified regions of interest followed by autoencoder and PCA based features dimensional reduction techniques. Finally, SVMs are used for the classification purposes. A Similar technique has been adopted in [13], where, initially, faces are detected using a Viola-Jones face detector followed by a feature extraction phase where Local Binary Pattern (LBP) features are used for representation purposes.

On the other hand, Anurag De et al. in [14] modeled FER system using an eigenface based approach where hue-saturation values are used for face detection. By calculating Euclidean distance between the test image and mean of training dataset, the expressions are then classified. For FER, Lekdioui et al. in [15] rely on a novel facial decomposition technique. First, the regions of a face are extracted using facial landmarks by using the algorithm IntraFace. For feature extraction, different techniques, such as LBP, CSLBP, LTP and Dynamic LTP, are used. SVMs are then used for classification purposes. Suzan Anwar et al. in [16] proposed an Active Shape Model (ASM) tracker, which takes input from a webcam and tracks 116 facial landmarks. These tracked landmark points are used for extracting the feature of expression from face and Support Vector Machine (SVM) is used for classification.

III. Proposed Methodology

Figure 2 provides a block diagram of the methodology adopted in this study. The method is mainly composed of two components, namely (i) feature extraction and (ii) classification. For feature extraction, state-of-the-art feature descriptors, namely HoG, is used, whereas, three different kernels are used of SVM for the classification.

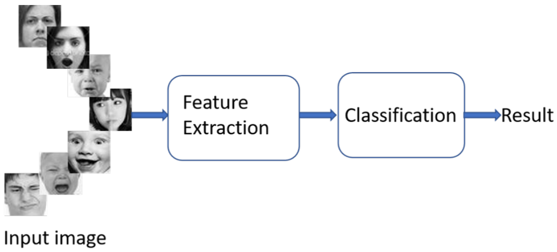


Fig. 2. The flow chart of the proposed model

The Histogram of Oriented Gradient (HoG) features are widely used for classification in different fields. The steps of HoG for extracting features are explained as follows:

1) The input image is divided into 8×8 pixel grid.
2) Then, 2×2 cells are created as block by 50% of overlapping, the total number of blocks are $5 \times 5 = 25$ for 48×48 images.
3) The direction of the gradient orientation range $(0 - 180^o)$ is divided into 9 bins. By the equation 1 gradient magnitude and orientation are calculated as follow

$$\begin{aligned} \mathrm{d}x &= f(x+1,y) - f(x-1,y) \\ \mathrm{d}y &= f(x,y+1) - f(x,y-1) \qquad (1) \\ \mathrm{m}(x,y) &= \sqrt{dx^2 + dy^2} \end{aligned}$$

In Equation 1, $f(x,y)$ describes the pixel value at location (x,y) in the given face patch I, and m describes the magnitude.
4) The length of the HoG feature vector for given I is $N \times C \times P$, where N denotes the number of the blocks, C denotes the number of cells in each block, and P denotes the number of orientation bins. The orientation are quantized into 9 bins, where makes the HoG features of length $25 \times 4 \times 9 = 900$ [17], [18]. An example of the HoG process for feature extraction is shown in Figure 3.

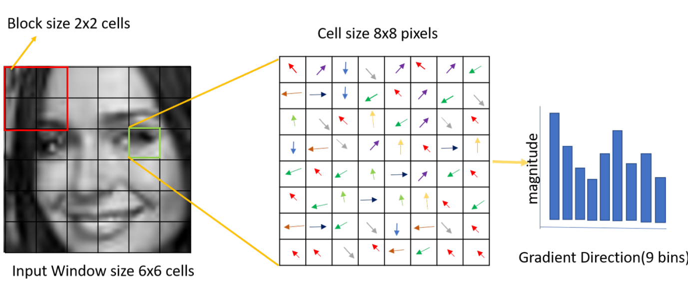


Fig. 3. HoG process for extracting features.

The SVM has shown outstanding performances in different application domains [19], [20]. The core goal of SVM is to find the best hyperplane to separate binary classes at maximum distance with the minimum number of support vectors. Consider a set of l training examples $(r_i, t_i), i = 1, ...l$, where each example is n-dimensional, $r_i \in \mathbb{R}^n$, a class label $t_i \in \{+1, -1\}$. A function ϕ is learned that maps given unknown instance r_j to t'_j, $t'_j = \phi(r_j)$.

Inherently, SVM is a binary classifier. To extend SVM for multi-class classification, either one-vs-one (OVO) or one-vs-all (OVA) approach is used. In case of OVA, for q different classes where $q > 2$, q different classifiers are trained, for each class i, it assumes i as positive and rest all other as negative. The OVA often leads to imbalanced training dataset. In Fer 2013 dataset, it also create huge imbalance for few classes such as facial expression *Disgust*. If binary classifier for *Disgust* expression is trained using OVA, then there will be only 436 positive instances and 28273 negative instances; the positive instances will only be 1.5% of given training instances.

For many applications, OVO has shown better accuracy compared to OVA. The OVO approach trains $q(q-1)/2$ binary classifiers. In case of FER, there will be 21 different binary classifiers. The ensemble/voting approach is used to decide the label of given instance r_j, the unknown instance r_j is given to 21 different binary classifiers, and the label with highest frequency is decided as the expression; OVO is used in this paper.

IV. Experiments and Results

A. Dataset

The SVMs are evaluated on a large-scale benchmark dataset, namely, Fer 2013 dataset [21]. The imbalanced nature of the dataset makes it one of the most challenging datasets for FER. In total, the dataset covers 7 different facial expressions, namely *Angry, Disgust, Fear, Happy, Sad, Surprise* and *Neutral*. The dataset is divided into training and test sets. The test set is further divided into private and public testing. The training set contains 28709 instances against all expressions, whereas, the test set contains 7180 instances. The number of instances of each facial expression in test and training set are described in Table I.

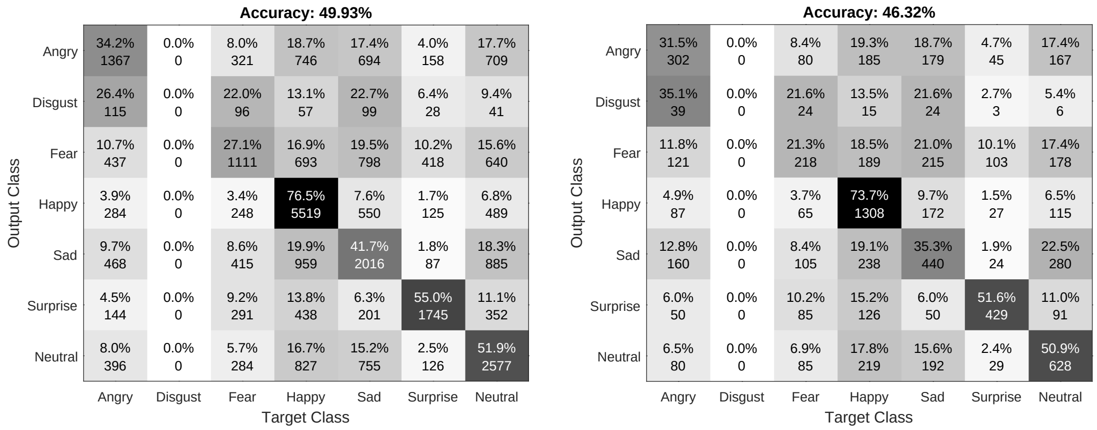
It can be observed that some expressions have very few training instances. For instance, *Disgust* has only 436 training instances, which makes dataset imbalanced for the disgust expression. Since classification models are data greedy models, therefore, the recognition heavily depends on the number of training instances per class.

TABLE I. Fer 2013 Dataset Description

Facial Expression	Training Set Instances	Test Set Instances
Angry	3995	958
Disgust	436	111
Fear	4097	1024
Happy	7215	1774
Sad	4830	1247
Surprise	3171	831
Neutral	4965	1233

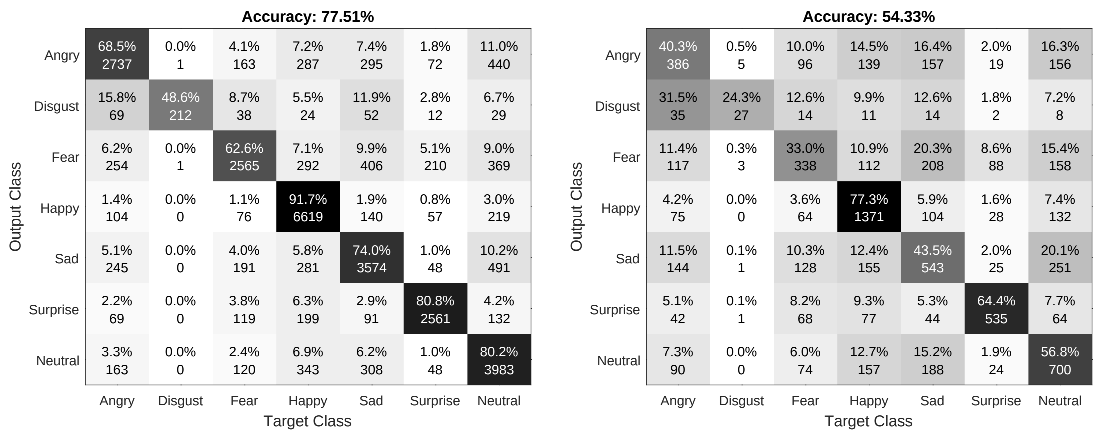
B. Experimental Results

Table II summarizes the recognition accuracies. It can be seen that linear SVM has limited performance compared to



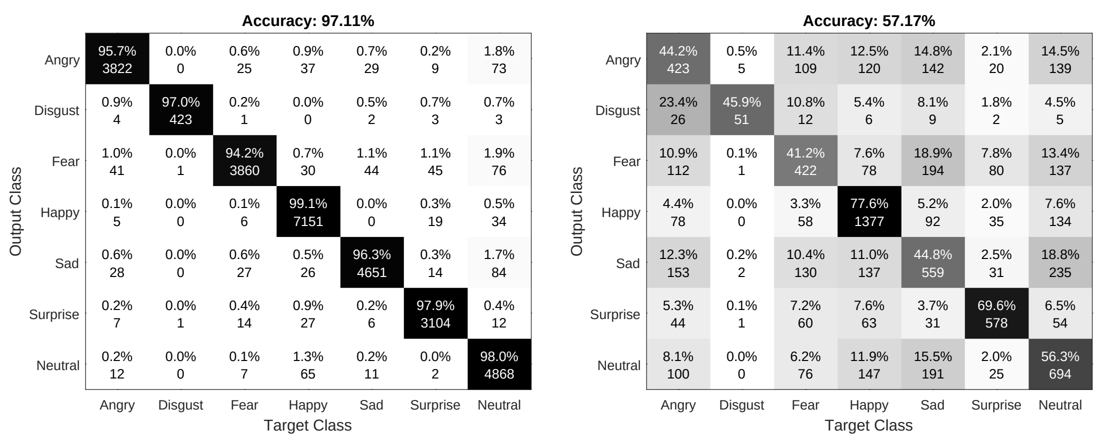
(a) (b)

Fig. 4. FER on linear SVM using HoG features. a) Shows the results on the training set, and b) Shows the results on the test set.



(a) (b)

Fig. 5. FER on quadratic SVM using HoG features. a) Shows the results on the training set, and b) Shows the results on the test set.



(a) (b)

Fig. 6. FER on cubic SVM using HoG features. a) Shows the results on the training set, and b) Shows the results on the test set.

the other kernels. The quadratic SVM gives 77.51% accuracy on training set but 54.33% accuracy on the test set which gives an impression that feature space is quite challenging and HoG feature representation is limited for FER application. Even cubic SVM, which has better performance for non-linear feature space, has only 57.17% accuracy on test set. Interestingly, cubic SVM over-fits on the training set and achieves 97.11% accuracy.

Moreover, in order to provide more detailed analysis of the results; Figure 4, Figure 5 and Figure 6 show the confusion matrices of linear, quadratic and cubic kernels, respectively, on HoG features.

TABLE II. COMPARISON BETWEEN THE RESULTS OF LINEAR SVM, QUADRATIC SVM AND CUBIC SVM ON HOG FEATURES

Feature Vector	Training Set Accuracy	Test Set Accuracy
Linear SVM	49.93%	46.32%
Quadratic SVM	77.51%	54.33%
Cubic SVM	**97.11%**	**57.17%**

It is interesting to observe that linear SVM fails to learn hyperplane for *Disgust* expression in training set and also in test set which indicates that *Disgust* expression is not linearly separable. It is also evident from Figure 4 that expression *Happy* achieves better recognition compared to the all other expressions. The happy expression also have similar recognition in quadratic and cubic SVMs test sets. On a leader-board, top three positions are the same for all SVM kernels whereas few expressions keep changing their positions. In case of disgust expression, it is on the last position (7th) on linear and quadratic SVMs, whereas, it improves the position on Cubic SVM that is 4th which indicates that cubic SVM gives better results in imbalance training as well. Almost all expressions improve their accuracies from linear to cubic SVMS except *Neutral* expression; for *Neutral* expression quadratic SVM gives the highest accuracy.

V. CONCLUSION

In this paper, an empirical study is conducted on SVM for investigating the impact of different kernel functions on the performance of FER. There are seven different facial expressions; the expressions include *Angry, Disgust, Fear, Happy, Sad, Surprise* and *Neutral*. Cubic SVM achieves highest accuracy on the test set, whereas, it over-fits on the training set. The cubic SVM also learns better hyperplanes under imbalance training instances. In case of *Disgust* expression, the training instances are only 1.52%. The linear SVM fails to learn the hyperplane for *Disgust* expression, the accuracy of *Disgust* is 0%; it neither appeared as false positive for other expressions as well. The cubic SVM has better accuracies for all the expressions except *Neutral*; quadratic SVM gives highest accuracy for the *Neutral*. The linear SVM give poor performance for all the expressions except *Happy* expression; linear SVM gives competitive accuracy for *Happy* expression, that is 73.7%, whereas, the best is 77.6% using cubic SVM.

REFERENCES

[1] C. Shan and R. Braspenning, *Recognizing Facial Expressions Automatically from Video*. Springer US, 2010, pp. 479–509.

[2] M. S. Hossain, *Patient State Recognition System for Healthcare Using Speech and Facial Expressions*. Plenum Press, 12 2016, pp. 251–287.

[3] L. C. Jain, U. Halici, I. Hayashi, S. Lee, and S. Tsutsui, *Intelligent biometric techniques in fingerprint and face recognition*. CRC press, 1999, vol. 10.

[4] C. Ilyas, K. Nasrollahi, T. Moeslund, M. Rehm, and M. Haque, "Facial expression recognition for traumatic brain injured patients," *Imaging and Computer Graphics Theory and Applications*, pp. 259–275, 2018.

[5] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *PATTERN RECOGNITION*, pp. 259–275, 1999.

[6] F. Ekman, Paul and W. V., "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, pp. 124–129, 1971.

[7] J. K. Shruti Karkra, "Compound facial expression recognition through gabor filter and rbf network," *International Journal of Computer Science and Mobile Computing*, pp. 576–583, 2016.

[8] D. Astler, H. Chau, K. Hsu, A. Hua, A. Kannan, L. Lei, M. Nathanson, E. Paryavi, M. Rosen, H. Unno, C. Wang, K. Zaidi, X. Zhang, and C.-M. Tang, "Increased accessibility to nonverbal communication through facial and expression recognition technologies for blind/visually impaired subjects," in *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 2011, pp. 259–260.

[9] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *ICMI*, 2015.

[10] M. Al-Shabi, W. P. Cheah, and T. Connie, "Facial expression recognition using a hybrid CNN-SIFT aggregator," *CoRR*, vol. abs/1608.02833, 2016.

[11] M. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *CoRR*, 2018.

[12] M. U. Ilyas, S. Latif, and J. Qadir, "Using deep autoencoders for facial expression recognition," *CoRR*, 2018.

[13] N. Kauser and J. Sharma, "Facial expression recognition using lbp template of facial parts and multilayer neural network," *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 445–449, 02 2017.

[14] A. S. Anurag Dea and D. M. Palb, "A human facial expression recognition model based on eigen face approach," *International Conference on Advanced Computing Technologies and Applications (ICACTA2015)*, 2015.

[15] K. Lekdioui, Y. Ruichek, R. Messoussi, Y. Chaabi, and R. Touahni, "Facial expression recognition using face-regions," in *Proceedings of the 3rd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2017, pp. 1–6.

[16] M. M. Suzan Anwar, "Real time intention recognition," *IEEE International Conference on Advances in Computer Applications*, 2015.

[17] A. Alsubari and R. J. Ramteke, "Extraction of face and palmprint features based on lbp, hog and zernike moments," *International Journal of Computer Applications*, vol. 172, pp. 31–34, 08 2017.

[18] S. Kim and K. Cho, "Fast calculation of histogram of oriented gradient feature by removing redundancy in overlapping block," *Journal of Information Science and Engineering*, vol. 30, pp. 1719–1731, 11 2014.

[19] H. Byun and S.-W. Lee, "Applications of support vector machines for pattern recognition: A survey," in *Pattern recognition with support vector machines*. Springer, 2002, pp. 213–236.

[20] K. Ahmad, M. L. Mekhalfi, N. Conci, F. Melgani, and F. D. Natale, "Ensemble of deep models for event recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2, p. 51, 2018.

[21] D. Goodfellow, Ian J.and Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*, 2013, pp. 117–124.

Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study

Hend Sayed, Manal A. Abdel-Fattah, Sherif Kholief
Information Systems Department
Faculty of Computers and Information
Helwan University
Cairo, Egypt

***Abstract*—This study was conducted based on an assumption that Spark ML package has much better performance and accuracy than Spark MLlib package in dealing with big data. The used dataset in the comparison is for bank customers transactions. The Decision tree algorithm was used with both packages to generate a model for predicting the churn probability for bank customers depending on their transactions data. Detailed comparison results were recorded and conducted that the ML package and its new DataFrame-based APIs have better-evaluating performance and predicting accuracy.**

***Keywords*—Churn prediction; Big data; Machine learning; Apache Spark; ML package; MLlib package; Decision tree**

I. Introduction

Recently, big data [1] technologies became more popular and are being used in many fields. It is a critical issue for most business owners to find the optimal solutions to automate their work and process their huge amount of data. A deluge of data is flooded all the time from many resources and there is a real competition in how to deal with it efficiently and with high performance. One of the most common needs is to predict customers churn depending on their data and activities. This need increases for businesses which are dealing with numerous clients such as telecommunications and banking[2], in this paper, the comparative study was conducted on transactions data of bank customers.

Churn prediction[3][4] is the process of predicting the intention of customers to leave. It is one of the most debated researches in last years. This study is conducted on a dataset of transactions of bank customers to predict their probability to leave.

Apache Spark has added solutions for MapReduce limitations and now it is widely used due to its high performance and efficiency in processing a huge amount of data that is 100 x times faster than Apache Hadoop [5]. Spark's machine learning APIs were based on Resilient Distributed datasets (RDD) in MLlib package, and now the primary API is the ML package which is a higher level API that is based on DataFrames that facilitate practical ML pipelines, especially feature transformations. AS mentioned in Apache Spark ML guide, DataFrames provide a more user-friendly API than RDDs. The many benefits of DataFrames include Spark Datasources, SQL/DataFrame queries, Tungsten and Catalyst optimizations, and uniform APIs across languages. MLlib package is in the maintenance mode with Spark 2.0. In this paper, a comparative study between the two packages is conducting in terms of accuracy, model training and model evaluation.

This research aims to highlight the practical differences between the two packages and list the pros and cons of each package depending on the real results of processing the same dataset using the same algorithm.

This paper is organized as follows, the 2nd section presents an overview of the related work. Then in the 3rd section, the used dataset and the steps of processing the data are discussed. Also, it is discussed in details the used algorithm and why it is chosen for this study. In 4th section evaluation and results are outlined. The conclusion of this research is summarizing the results of the comparative study.

II. Literature Review

A. Churn prediction

Customer churn [6] is the term used in the banking sector tries to denote the movement of customers from one Bank to another.

The Importance of Predicting Customer Churn [7]

- Avoiding losing revenue that results from a customer abandoning the bank.
- The cost of acquiring a new customer is 5x higher (Lee Resources 2010).
- Intensifies the competition among commercial banks.

According to these reasons, it is urgent for commercial banks to improve the capabilities to predict customer churn, thereby using good solutions for churn predicting to retain customers.

B. Churn prediction with big data

A large amount of data is being generated daily from different sources, which is much more expensive and much slower to be processed and analyzed[8]. Suitable and efficient solutions for storing, processing and analyzing a massive volume of data are critically needed to be able for churn prediction efficiently and accurately.

C. ML in Apache Spark

Applying machine learning on a big dataset needs a very large amount of physical resources for processing data, the need of having a platform that can efficiently perform very complex processing and analysis operations on enormous datasets increases daily[9]. Apache Spark is an efficient one of these platforms, it offers a set of modules for different machine learning tasks.

Gauri D et al., (2015)[10] proposed a churn prediction model for the telecommunications field that predicts churn probability for customers using their records data, which can help the company know which customer has an intention to leave or move to another service provider in the near future. Due to the increasing volume of customers' data, they are using Hadoop framework for data storing and processing. They applied the prediction model using the decision Tree C4.5 algorithm. The model generates the rules from training data and applies these rules on testing data to determine which customer may leave. The process of rules generation involves calculation of entropy for every attribute of each record along with the information gain. After applying rules on the test data, a list of the predicted churners and non-churners customers is returned as a result and added to a text file in HDFS.

Avishkar D. et al., (2016) [11] proposed an application for predicting customers attrition in the telecommunication field and enable business owners to take all needed procedures to retain the existing customers rather than increasing the number of customers. They take customer records as an input and give a prediction of customer churn as output. They applied prediction through analyzing customer behavior based on several attributes such as calls per day and using provided services. Because of the overall growth rate for over 35 percent over the past decade in terms of subscribers, they applied prediction using the Hadoop framework that made it much easier. They applied the decision tree approach using C4.5 data mining algorithm, which provides great accuracy in predicting churners. The result was obtained after applying the algorithm is a list of churners and non-churners.

Wei Z et al., (2016) [12] conducted a comparison of decision tree based churn prediction model between SPSS and Spark ML package and used a customers' data of insurance company as an example. The comparison was conducted on the execution flow of each, run-time, model evaluation, and model precision. Their results show that Spark ML is easier and more efficient than SPSS in applying a churn prediction model, especially for insurance companies.

D. Decision tree

The decision tree is a supervised learning algorithm that is used for regression or classification. Decision tree model has a tree-like structure that consists of nodes. Each node in this tree refers to a test on a single attribute, each branch represents an outcome of this test, and each leaf node holds a class label. The splitting of nodes is decided by /algorithms like information gain, chi square, gini index[13]. There are different algorithms used in the decision trees: ID3, C4.5, CART, C5.0, CHAID, QUEST, and CRUISE.

III. Methodology

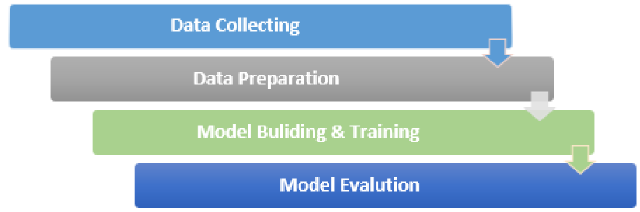


Fig. 1. Building ML classification model workflow

The workflow for building a machine learning classification model shown in Figure 1. After collecting data, it is prepared, transformed and processed to be ready for the training phase. A classification algorithm is applied to the prepared data for generating the classification model, then the model is being tested and evaluated using testing data. Methodology is divided for subsections, section 3.1 discusses the used dataset, subsection 3.2 presents the pre-processing steps required to prepare the dataset for the training phase. Then in section 3.3 Spark MLlib is illustrated with its components and work-flow, and the same discussion for spark ML package in section 3.4. In section 3.5 the results are outlined.

TABLE I. Dataset Attributes

Attribute	Description
Row Number	
Customer ID	
Surname	
Credit Score	
Geography	The location of the custormers of three countries where the bank is operating
Gender	
Age	
Tenure	The period of having the account in monthes
Balance	
NumOfProducts	
HasCrCard	If the customer has a credit card
IsActiveMember	
Estimated Salary	According to the different factors such as credit score, transactions company has used the data to calculate the estimated salary of the customer.
Churn	Indicates the customer has leaved or not

A. Dataset

A dataset of bank customers transactions is used in this study for predicting bank customers churn. The dataset is freely available online on Kaggle[1]. It contains 10k row and 14 columns, where each row represents a customer data and each column represents a single attribute. Table I illustrates the attributes of the used dataset and a description for non-descriptive attributes names.

B. Correlations Data Preparation

First, the irrelevant attributes (Row Number, CustomerID and Surname) were dropped. Then the strings or categorical attributes were mapped to numeric values. The mapped values are for Geography and Gender attributes, for Geography, there

[1]Kaggle: a platform for predictive modeling and analytics competitions, www.kaggle.com.

are three categories or values (France: 0.0, Spain: 1.0, Germany: 2.0) and there are two categories for Gender (Female: 0.0, Male: 1.0).

Then, a statistical analysis was performed on a part of the dataset to examine correlations between the numeric columns and generate scatter plots of them. As shown in Figure 2, it resulted that no high correlated pairs found, no more attributes were dropped.

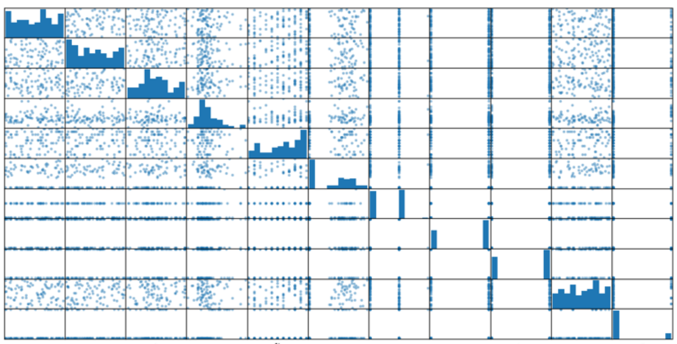


Fig. 2. Correlations scatter plots

After applying the training algorithms, despite the high accuracy, the sensitivity to the non-churners was much more than the sensitivity to the churner customers. By checking the count of each class, the count of churners was 2037, and for non-churners, the count was 7963. so the non-churners was down-sampled to a fraction of 2037/7963, that resulted in 2048 non-churners.

C. MLlib package

Spark MLLib package (spark.mllib) is the older package for machine learning on spark, it contains the original API that is built on top of Resilient Distributed Datasets (RDDs) which are immutable and partitioned collections of elements that can be operated on in parallel. Other advantages of RDDs are Persistence, fault-tolerance, lazy evaluation and typing. RDD-based API requires transforming the data into rows of type LabeledPoint before processing, each LabeledPoint consists of a label and a vector of features which represent the attributes.

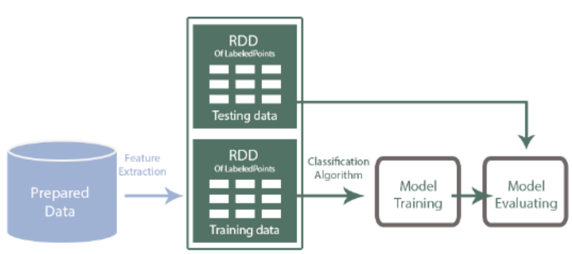


Fig. 3. MLlib Workflow

As shown in Figure 3, after transforming the data to RDD of LabeledPoints, each LabeledPoint consists of the label which represents the Churn column, and features column which is a vector of all features values. The RDD was divided into two sets (Training data & Testing data). Training data are used in applying the decision tree classification algorithm to generate the model using trainClassifier() function in MLlib DecisionTree module. The decision tree algorithm was applied with Gini impurity. After generating a model, the testing data was used for evaluating the model.

D. ML package

Spark ML (spark.ml) is a newer package that was introduced in Spark 1.2, it contains a newer machine learning APIs that are built on top of DataFrames, and it is currently the primary APIs for machine learning on Spark. As shown in Figure 4, The ML package workflow is represented by **ML Pipeline** that consists of some chained PipelineStages. Each **PipelineStage** can be either a **Transformer** or an **Estimator**.

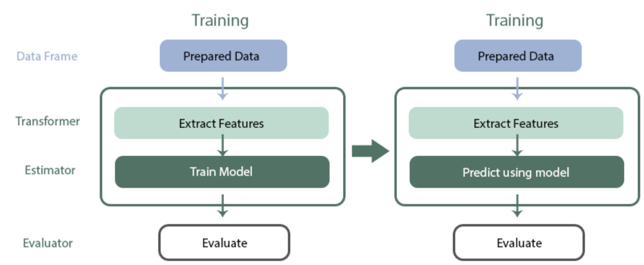


Fig. 4. General ML Workflow

The transformer is an algorithm that is transforming a dataFrame into another dataFrame while the estimator is an algorithm which can be fitted on a dataFrame to produce a model, which is a transformer. The pipeline itself is an estimator that is generating the classifier model. As shown in Figure 5, StringIndexer estimator is used for indexing the label column and VectorIndexer estimator to automatically check categorical features depending on a provided **maxCategories** value and generate a new column that contains a vector if features indices. The Spark ML decision tree classifier DecisionTreeClassifier() is an estimator that is added to the pipeline chain and used to generate the decision tree model. The researchers followed Spark python API docs [2] in implementation.

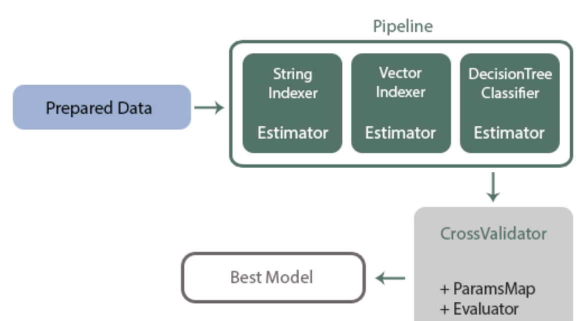


Fig. 5. Implemented ML Workflow

CrossValidator is used for best model selection using the generated pipeline as an estimator, and provided parameters

[2] Spark python API docs, https://spark.apache.org/docs/latest/api/python.

map that is generated using ParamGridBuilder() and, identifying the decision tree maxDepth values to search through. An evaluator MulticlassClassificationEvaluator() is used for testing the generated model, it works on a dataset with two attributes (Label and Prediction) and is used to get the predicting accuracy. CrossValidator takes a number to split the dataset into a set of folds that are used as separate training and testing datasets.

E. Results

Result 1, the resulted accuracy values for the two models were slightly close, however, the accuracy of the ML package model was better. As shown in Figure 6, the accuracy value for ML package model was 0.79 and the MLlib package model's accuracy was 0.73.

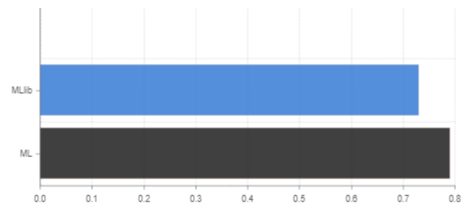


Fig. 6. Compared accuracy

Result 2, MLlib package needed less time for data transformations, applying the classification algorithm and training the model. The results are shown in Figure 7, it took only 6 seconds, and the model that was generated using ML package took 25 seconds.

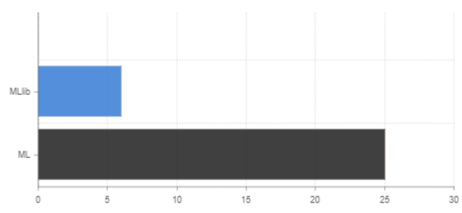


Fig. 7. Compared training time

Result 3, In contrast to the previous result, and as shown in Figure 8, the time needed for evaluating the model using the same testing data was much fewer in ML package model than the MLlib package model. It took only 5 seconds, and the model of MLlib needed 14 seconds.

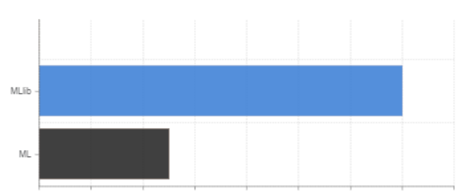


Fig. 8. Compared evaluation time

IV. Conclusion

In this paper, a comparative study between Apache Spark ML and MLlib packages was conducted, in terms of accuracy, model training and model evaluation. The researchers performed the comparison on bank customers transactions dataset and using decision tree algorithm for predicting customers churn probability. MLlib package with its RDD-based API has a better result for training time, which can be caused by the internal transformations in both packages. ML with its DataFrames-based API has better results for testing time and accuracy. So, the results indicate that ML churn prediction models can perform better and help in getting more accurate results faster. These results are useful for banks and any businesses that are dealing with numerous clients and records in predicting their churn probability. In the future, more detailed comparative studies will be done for more packages and platforms with different types of data and using different algorithms to check the best and most accurate models in different situations.

References

[1] N. K. Gyamfi, "Big Data Analytics : Survey Paper Scanned by CamScanner," no. February, 2017.

[2] N. Elgendy and A. Elragal, "Advances in Data Mining. Applications and Theoretical Aspects," vol. 7987, no. August, 2013. [Online]. Available: http://link.springer.com/10.1007/978-3-642-39736-3

[3] T. Y. Fei, L. H. Shuan, L. J. Yan, G. Xiaoning, and S. W. King, "Prediction on customer churn in the telecommunications sector using discretization and Naïve Bayes classifier," *International Journal of Advances in Soft Computing and its Applications*, vol. 9, no. 3, pp. 23–35, 2017.

[4] A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining," *Financial Innovation*, vol. 2, no. 1, p. 10, 2016. [Online]. Available: http://jfin-swufe.springeropen.com/articles/10.1186/s40854-016-0029-6

[5] T. Lin, Y. Li, and R. Mai, "Modeling and Control Method of Inductive Power Transfer System Based on LCL-S Topology," *Diangong Jishu Xuebao/Transactions of China Electrotechnical Society*, vol. 33, no. 1, pp. 104–111, 2018.

[6] K. Chitra and B. Subashini, "Customer Retention in Banking Sector using Predictive Data Mining Technique," p. 4, 2011.

[7] B. He, Y. Shi, Q. Wan, and X. Zhao, "Prediction of customer attrition of commercial banks based on SVM model," pp. 423–430, 2014.

[8] A. Bhattacharya and S. Bhatnagar, "Big Data and Apache Spark : A Review," no. 5, pp. 206–210, 2016.

[9] M. Assefi, E. Behravesh, G. Liu, and A. P. Tafti, "Big data machine learning using apache spark MLlib," *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, vol. 2018-Janua, no. November, pp. 3492–3498, 2018.

[10] G. D. Limaye, J. P. Chaudhary, and N. Mumbai, "Churn Prediction Using MapReduce and HBase," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, no. 3, pp. 1699–1703, 2015.

[11] A. Dalvi, B. Bhor, H. Chauhan, and P. Sawant, "Churn Prediction Using Hadoop," Tech. Rep., 2016. [Online]. Available: www.irjet.net

[12] W. Zhang, T. Mo, W. Li, H. Huang, and X. Tian, "The Comparison of Decision Tree Based Insurance Churn Prediction between Spark ML and SPSS," *Proceedings of International Conference on Service Science, ICSS*, pp. 134–139, 2017.

[13] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2011.08.024

Multimodal Automatic Image Annotation Method using Association Rules Mining and Clustering

Mounira Taileb, Eman Alahmadi
Faculty of Computing and Information Technology
King Abdulaziz University
Saudi Arabia, Jeddah

***Abstract*—Effective and fast retrieval of images from image datasets is not an easy task, especially with the continuous and fast growth of digital images added everyday by used to the web. Automatic image annotation is an approach that has been proposed to facilitate the retrieval of images semantically related to a query image. A multimodal image annotation method is proposed in this paper. The goal is to benefit from the visual features extracted from images and their associated user tags. The proposed method relies on clustering to regroup the text and visual features into clusters and on association rules mining to generate the rules that associate text clusters to visual clusters. In the experimental evaluation, two datasets of the photo annotation tasks are considered; ImageCLEF 2011 and ImageCLEF 2012. Results achieved by the proposed method are better than all the multimodal methods of participants in ImageCLEF 2011 photo annotation task and state-of-the-art methods. Moreover, the MiAP of the proposed method is better than the MiAP of 7 participants out of 11 when using ImageCLEF 2012 in the evaluation.**

***Keywords*—Automatic image annotation; association rules mining; clustering**

I. Introduction

Nowadays, we are witnessing an enormous increase in the number of images available on the web which makes image retrieval (IR) a challenging task. In literature, content-based image retrieval (CBIR) and text-based image retrieval (TBIR) are the main two approaches to achieve the IR. In the TBIR systems, the retrieval of images relies on the text or keywords (called also labels or concepts) entered by users; such systems depend mainly on the image tags typed manually by human and/or the text accompanying and describing the image on the Web page. The main disadvantages of TBIR systems are: (1) the inaccuracy of the manual annotation, (2) manual annotation is not always available and impossible for large image database. Whereas, in the CBIR systems, the retrieval relies only on the visual content of images which is represented by visual features such as texture, edge, color, etc. CBIR systems suffer from a well known problem called the semantic gap problem [1], [2] which refers to the difference between the low level visual features of the image and the high-level user semantic concept. The automatic image annotation (AIA) is a way to address this issue; it has received an increasing attention from researchers and has become an important field of research for image retrieval [3], [4], [5]. Indeed, AIA facilitates the search in huge datasets of images and improve the retrieval of images that are semantically similar to a query image. AIA is the process of automatically assigning keywords from a predefined vocabulary to an image, keywords that characterize its semantic visual content. Several approaches exist in the AIA, they are classified into three categories [6] as illustrated in Fig. 1.

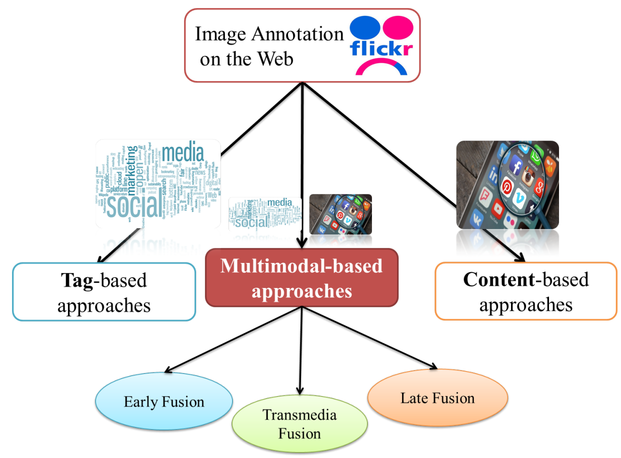


Fig. 1. Automatic image annotation approaches

In tag-based image annotation [7], [8], the tags used to annotate an image are collected from the predefined user tags and the articles around the image. However, in content-based image annotation, only the visual features are used [9], [10]. And in multimodal-based approaches both modalities (tag and visual features) are utilized to annotate an image [6], [11], [12]. In this paper, a new multimodal image annotation method is proposed using association rules mining (ARM) and clustering techniques. The ARM is used to explore the semantic relations that might exist between text clusters and visual feature clusters in order to associate them. The goal is to combine visual and text modalities to improve the image annotation performance. The proposed method consists of two phases; the training phase and the annotation phase. In the training phase, the input is a set of tagged or labeled images. The visual features of images are clustered as well as the tags of images which are considered as text features. Then the ARM is used to fuse the visual clusters and text clusters and find relations between them. The training phase provides at the end a list of association rules. In the annotation phase, a new image is annotated using its visual features which are used to extract the related association rules to provide a list of tags to the image.

The remaining part of the paper is structured as follows.

In Section 2, the different approaches in AIA are presented. Then Section 3 describes the association rules mining. Section 4 provides a detailed description of the proposed method with its training and testing phases. The experimental evaluation is detailed in Section 5. The paper conclusion is provided in section 6.

II. Related Work

As mentioned above, several approaches can be used to solve the problem of AIA, they are categorized into three categories: tag-based approaches, content-based approaches and multimodal-based approaches. In the tag-based approach, images are retrieved and annotated based on the text or keywords entered by users; such systems depend mainly on the image tags typed manually by humans and/or the text provided with the image on the web page. The goal of the methods proposed in the tag-based approach is whether the refinement or enrichment of the image annotation. In some methods [8], [13] a list of tags is refined and reduced to a smaller list of reliable tags. And in other methods [6], [7], [12], [14] a short list of tags are enriched after performing AIA.

In content-based annotation approach, several methods have been proposed [5], [10], [15], [16], [17], [18]. In this approach, a set of training images is used with the image annotation keywords. The images are segmented into regions [19] to create a high-quality segmentation, then the visual features of regions are extracted and clustered to obtain the blobs. A blob refers to the label associated to a region. After that, blobs are linked with the keywords Fig. 2, and this is the key process of the content-based annotation approach. The linking process can be performed using several methods, such as: Expected Maximum (EM) in the translation model [15], the latent Dirichlet allocation (LDA) model [16], the co-occurrence model [17] , the Cross-Media Relevance Model (CMRM) [10] and Multiple Bernoulli Relevance Model (MBRM) [5]. To be annotated, a new image is segmented to extract regions, and the features of each region are utilized to determine its corresponding blob. Finally, keywords are predicted from the corresponding blobs to annotate the image. The methods proposed in multimodal-based annotation approach focus on leveraging both modalities (tag-based and content-based) to improve the automatic image annotation performance. Many studies have been conducted in this area, falling into three categories: the early fusion, transmedia fusion and late fusion [20], [21], [22]. The early fusion consists of combining the visual and text features into a single representation before the annotation process. In the transmedia fusion, one of the two modalities (visual or text) is used to gather the relevant images before switching to the other modality to aggregate tag features of these relevant images and perform the image annotation. However, in the late fusion, there is a separation in each modality processing, then the results of two previous processes are combined at the decision level of the annotation process.

III. Association Rules Mining

One of the well-known and widely used data-mining technique is the association rules mining. In a knowledge discovery operation, data mining is defined as a procedure that attempts to discover a new and meaningful pattern in data. Analyzing the association rules between groups of items in a set is useful for uncovering the interesting relationships that may be hidden in a huge database. The example of the content of a market baskets is considered a classical example for the extraction of association rules. Items are defined as things that anyone may buy from the market, and transactions are the diverse items contained in market baskets. Below is an example of a simple association rule extracted from market baskets:

$$\text{Bread} \rightarrow \text{Eggs}$$

This rule shows that there is a strong relationship between the selling of bread and eggs, which results when customers who buy bread also buy eggs. The goal of discovering such rules is to describe the customers purchase behavior, which can help companies to find opportunities for sales and better guide category management in order to increase profits. Association-rule mining is usable in several applications in diverse fields, such as web mining, Bioinformatics, and medical diagnosis. The method proposed in this paper aims to provide a semantic annotation by using the ARM algorithm to perform an association between text clusters and visual clusters that are semantically related. Generating a good set of ARs depends mainly on the setup of support and confidence, their calculations are given in the equations (1) and (2) respectively as defined in [23]:

$$Supp(X) = \frac{count(X)}{N} \tag{1}$$

Where N refers to all transactions in the transaction database T.

$$Conf(X \rightarrow Y) = \frac{count(X \cup Y)}{count(X)} \tag{2}$$

In the rule $X \rightarrow Y$, the calculation that defines how many items in Y appear also in other transactions that include X is known as confidence of the rule. While determining how often a rule is applied in a given database is known as the support of the rule.

IV. Proposed Method

A novel multimodal image annotation method using association rules mining and clustering algorithms is proposed, it takes advantage of using both text and visual modalities to improve AIA. The proposed method uses clustering algorithms to regroup text and visual features into clusters, then association rules mining are applied to generate the rules that associate text clusters to visual clusters. This can be considered as a late fusion between text and visual clusters. The novelty of the proposed method is in the use of clustering and associations rules mining. The Fig. 2 is an illustration of the method framework. The method comprises two main phases: the training phase and the annotation phase. In the first phase, the training dataset includes images with their associated text files (each image has a text file containing the list of user tags) [14]. All the steps of the training phase are described in the following sub-sections.

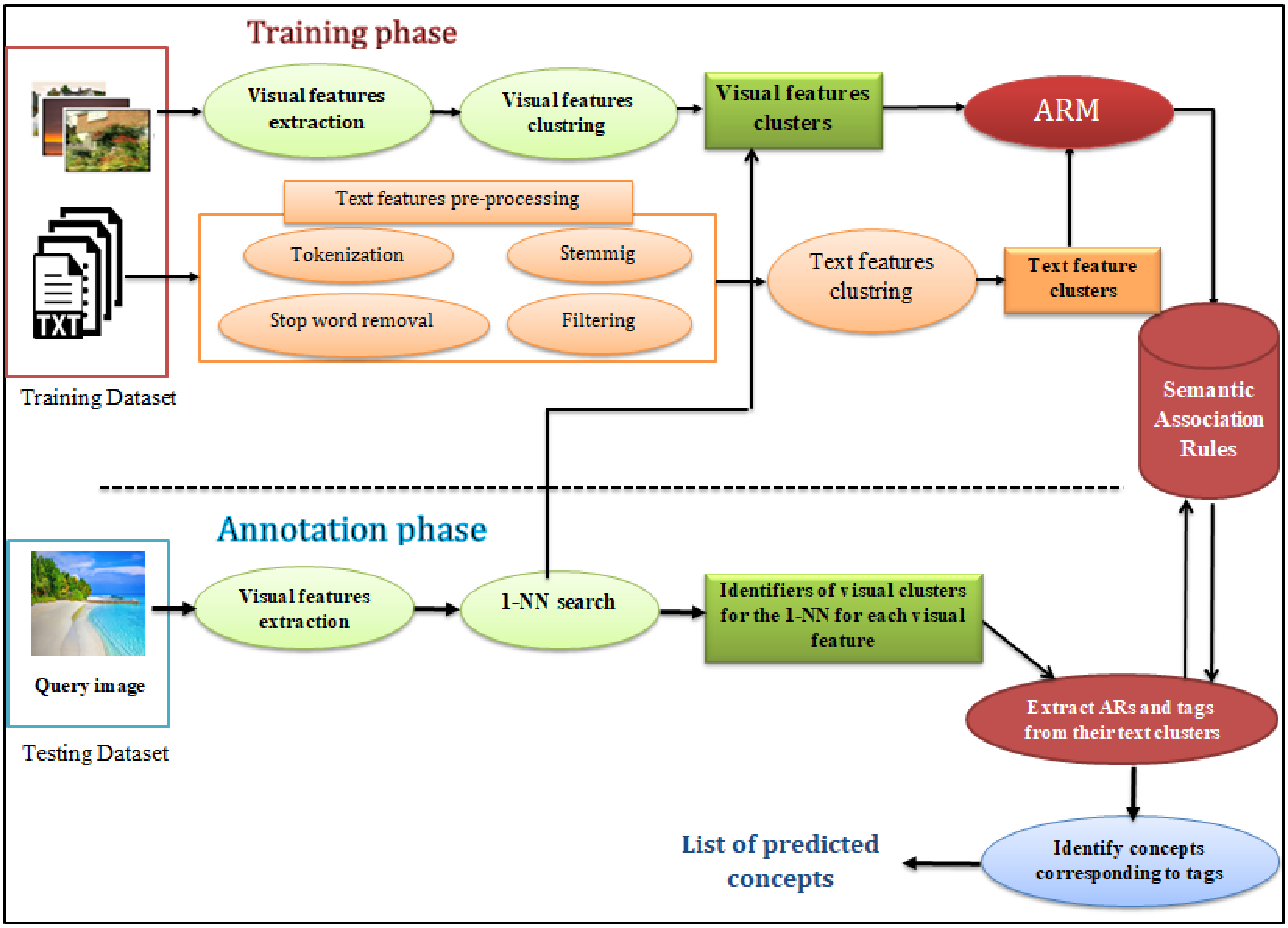


Fig. 2. Flowchart of the proposed annotation method

A. Training Phase

In the training phase, a series of processes is performed. It starts by extraction of visual and text features, then the clustering of both modalities. And ends with the generation of association rules. The different processes are detailed in the following.

1) Features Extraction: In this phase, two MPEG-7 global visual features are extracted; CSD (color structure descriptor) and EHD (edge histogram descriptor). In the other hand, the text files associated to images are considered as text features, they are first pre-processed by performing tokenization, stemming, stop word removal and filtering. The latter pre-processing step is performed by comparing the set of all words with American and British English dictionaries to ignore any non-English words. For example, the number of words (tags) in the dataset ImageCLEF 2011 was 88,083 and after pre-processing, 52,527 words remained. The number of remaining words have been reduced to less than 52,527 words, but we have noticed that short words like countries and cities names were removed (such as USA), hence, the final number of words was maintained at 52,527 words. It is worth to mention that only a few of the remaining words have been manually associated to the concepts predefined in the two datasets (ImageCLEF 2011 and ImageCLEF 2012) used in the experimental evaluation of the proposed annotation method. As an example, from ImageCLEF 2011, the following tags: cello, bass, viola, guitar, piano, drum, and organ are associated to the concept called Musical Instrument. For the dataset ImageCLEF 2011, the total number of words associated to concepts is 292, but these words are repeated several times in the final list of filtered words.

2) Clustering: The features extracted in the previous step are regrouped into clusters using clustering algorithms. The visual features are clustered using the hierarchical clustering method, also called indexing technique, NOHIS [24]. This indexing technique groups the visual features into clusters and the clusters are organized in a hierarchal structure called NOHIS-tree. For each of the visual features (CSD and EHD), a NOHIS-tree is constructed. The text features are clustered as well using K-means [25]. After the clustering of both visual and text features, the ARM algorithm is performed on visual and text clusters to find relations between them. The output is the list of association rules that will be used later in the annotation phase.

3) Generation of Association Rules: In order to generate the association rules, it is important to define first set of items (itemsets) *I* and the transaction database *T*. In the proposed method, the text and visual clusters represent the itemsets, where the text clusters are denoted by *C*ti and the visual features clusters are denoted by *C*cj for color clusters and *C*ek for edge clusters. After determining the features model space for all of the feature modalities, the transaction database *T* can be constructed and the ARM algorithm can be run over it.

The relationship between one text cluster *C*t and at least one of the visual clusters (color or/and edge, *C*c and *C*e respectively) is considered as transaction. In other words, the association is between a text cluster and one or two visual clusters is a transaction. A text cluster is associated to one or two visual clusters if the clusters of different modalities have images in common (since each tag in the text clusters and each visual feature in the visual clusters belong to an image). Hence, a transaction is constructed if the number of common images that are in both text and visual clusters is greater than zero, as illustrated in the following example:

$$If \mid Cti \cap Ccj \mid > 0, then \{Cti, Ccj\} is \quad added \quad to \quad T$$

Examples of transactions are given in the following:

$\{Ct_0, Cc_{62}\}, \{Ct_{33}, Ce_{57}\}, \{Ct_{46}, Cc_{25}\}, \{Ct_{51}, Cc_{42}, Ce_{75}\}, \{Ct_{68}, Cc_{67}, Ce_{76}\}$.

The formal equations of support and confidence have been modified as in [26] to be adapted to the proposed method for the following reasons; first, if the calculated support/confidence values of the association rules for the whole transaction database *T* result in a low support value, this could affect later the generated association rules. Second, the goal is to explore the semantic relations among text and visual clusters, hence the support and confidence calculation have to be restricted to the text clusters results. Therefore, the support and confidence of the rule defined as *C*ti $\rightarrow$ *C*vj (where *C*v refers to the visual cluster) are as follows:

$$Supp(Ct_i \rightarrow Ctv_i) = \frac{count(Ct_i, Ctv_i)}{count(Ct_i)} \tag{3}$$

$$Conf(Ct_i \rightarrow Ctv_i) = \frac{count(Ct_i, Ctv_i)}{max_k(count(Ct_i, Ctv_i))} \tag{4}$$

When there are multiple items on the right side, the similarity of the rule is as follows:

$$Supp(Ct_i \rightarrow Cv_j \mid j = 1, ..., m) = \frac{count(Ct_i, Cv_j \mid j = 1, ..., m)}{(count(Ct_i))} \tag{5}$$

$$Conf(Ct_i \rightarrow Cv_j \mid j = 1, ..., m) = \frac{count(Ct_i, Cv_j \mid j = 1, ..., m)}{max_k(count(Ct_i, Cv_k))} \tag{6}$$

The modified definitions of support and confidence are used to identify the frequent itemsets by applying *A*priori algorithm [27] as defined in [26]. The algorithm needs as parameters the minimum value of support $minsup$ and the transaction database *T*. Then the frequent itemsets are used along with the minimum value of confidence $minconf$ to generate the association rules.

Each association rule is stored with its support and confidence values, as shown in Fig. 3. Later, these association rules will be retrieved during the annotation phase to predict the list of tags (or concepts) to annotate a new image.

{Ct2} => {Cc82} (support: 6.9 confidence: 80)
{Ct22} => {Cc66} (support: 33.33 confidence: 100)
{Ct31} => {Ce73} (support: 2.48 confidence: 91.3)
{Ct74} => {Cc85} (support: 66.67 confidence: 100)

Fig. 3. Example of generated associations rules

It is worth to mention that the association rules has been used in a tag-based image annotation method [13], in this method the association rules is employed in a completely different way where each web page is considered as a transaction and the set of words in the web page is regarded as the itemset.

B. Annotation Phase

In the annotation phase, and for each image of the test dataset, the following processes are performed:

- Extraction of the color and edge visual features (CSD and EHD respectively).
- Search of the first nearest neighbor 1-NN for the visual features CSD and EHD extracted in the previous step. The search is performed on the two hierarchical structures NOHIS-tree constructed in the training phase for the two visual feature spaces. The distance used in the nearest neighbor search is the Euclidian distance. The nearest neighbor for each visual feature is returned along with its cluster identifier. In other words, the identifier of the visual cluster that contains the nearest neighbor of the visual feature CSD (extracted from the test image) is returned. And the same is done for the visual feature EHD. An example is provided in Fig. 4, where Cc_{66} and Ce_{54} are the clusters that contain the nearest neighbor of CSD and EHD features respectively.
- Each of the visual cluster identifiers is used to extract all the rules that contain this visual cluster. The text clusters of all the extracted rules are considered by taking the tags they contain, as shown in Fig. 4.
- The concepts that correspond to the previous tags are considered as annotation of the image. We remind that the image datasets used in the experimental results are using a list of concepts to annotate the image, and the tags have been linked to the concepts in the training phase.

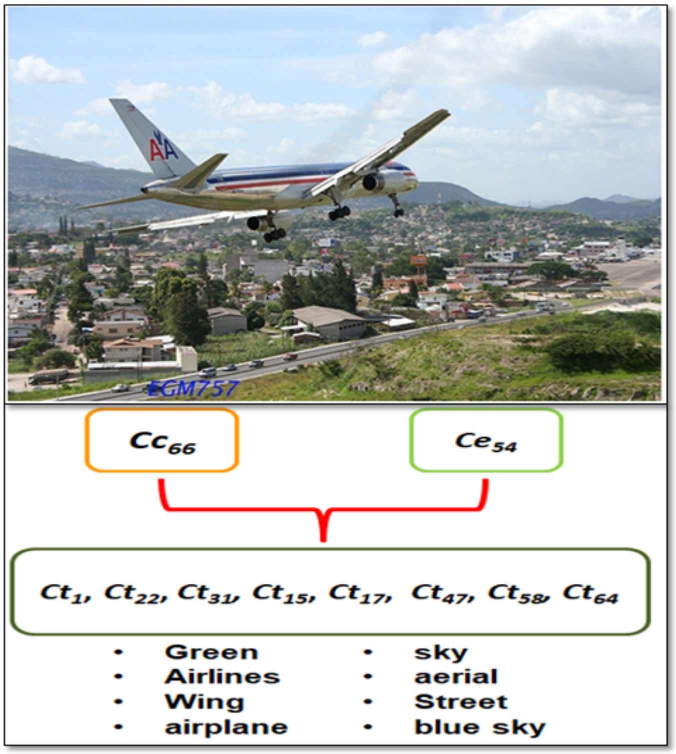


Fig. 4. List of tags extracted from the text clusters

V. Experimental Evaluation

Two datasets are used in the performance evaluation of the proposed automatic annotation method; ImageCLEF 2011 and ImageCLEF 2012. These datasets are considered because of two reasons; the availability of ground truth and their wide use. In the following, a detailed description is provided regarding the datasets, evaluation metrics and results. The values of *minsupp* and *minconf* are set respectively at 2% and 70% based on a set of experiments.

A. Datasets

The dataset of the annotation task in ImageCLEF 2011[1] is used; is consists of a training dataset of 8000 images with their associated user tags (in text files), and a test dataset of 10,000 images. The objective is to annotate the test dataset with 99 concepts. The second dataset is ImageCLEF 2012[2] used in the sub-task 1; concept annotation, for the year 2012. The size of the training dataset is 15,000 images provided with used tags and the test dataset is a collection of 10,000 images to annotate with 94 concepts.

B. Evaluation metrics

The mean interpolated average precision MiAP is used for the evaluation per concept, and F-Measure is used for the evaluation per example (per image). To calculate the MiAP, first 11 recall values are considered from 0.0 to 1.0 with steps of 0.1. Then for each recall value *R*, its interpolated precision *P*interp is the highest precision of any recall value $R' \geq R$:

$$P_{interp}(R) = max_{R' \geq R} P(R) \quad (7)$$

Then the average interpolated precision of the 11 recall values is calculated as follows:

$$AP_{interp} = \frac{1}{11} \sum_{R=0}^{1} P_{interp}(R) \quad (8)$$

The MiAP is the average of the average interpolated precisions of all concepts, it is calculated as follows:

$$MiAP = \frac{1}{C} \sum_{i=0}^{c} AP_{interp} \quad (9)$$

C is the number of concepts used for the annotation.

The calculation of F-Measure is given in the following equation:

$$F - Measure = \frac{2(P * R)}{(P + R)} \quad (10)$$

Where *P* is the precision and *R* is the recall for each annotated image.

C. Results

The results obtained with the proposed method are compared to results of multimodal methods proposed by participants in the annotation tasks of ImageCLEF 2011 and ImageCLEF 2012 and methods with best results we found in literature [6], [12], [21]. The MiAP [3] and F-Measure [4] are used in the comparison with participants. However, only the MiAP is available for the state-of-the-art methods.

The proposed method achieved results that outperforms those of all methods considered in table I in terms of MiAP and F-Measure. The proposed method obtains a gain of 23.3% of MiAP compared to the best MiAP achieved by the participant "TUBFI", and a gain of 17.7% of F-Measure compared to the best F-Measure achieved by the participant "ISIS". In addition, the MiAP obtained with the proposed method outperforms the best MiAP of [12] by 22.8% as show in in table II.

The table III illustrates the comparison of the proposed method results to results (of the multimodal methods) of participants in ImageCLEF 2012 photo annotation task and the multimodal annotation method proposed in [6]. The proposed method outperforms the results of 7 participants out of 11. The best MiAP outperforms ours by 11%. This is due to the few number of tags that have associated to the predefined list of concepts.

An example of images annotated with the proposed method is illustrated in Fig. 5. The concepts predicted by the proposed method are provided along with ground truth concepts by indicating the correct and wrong predicted labels.

[1]https://www.imageclef.org/2011/Photo
[2]https://www.imageclef.org/2012/photo-flickr
[3]https://www.imageclef.org/2011/PhotoAnnotationMAPResults
[4]https://www.imageclef.org/2011/PhotoAnnotationExampleBasedResults

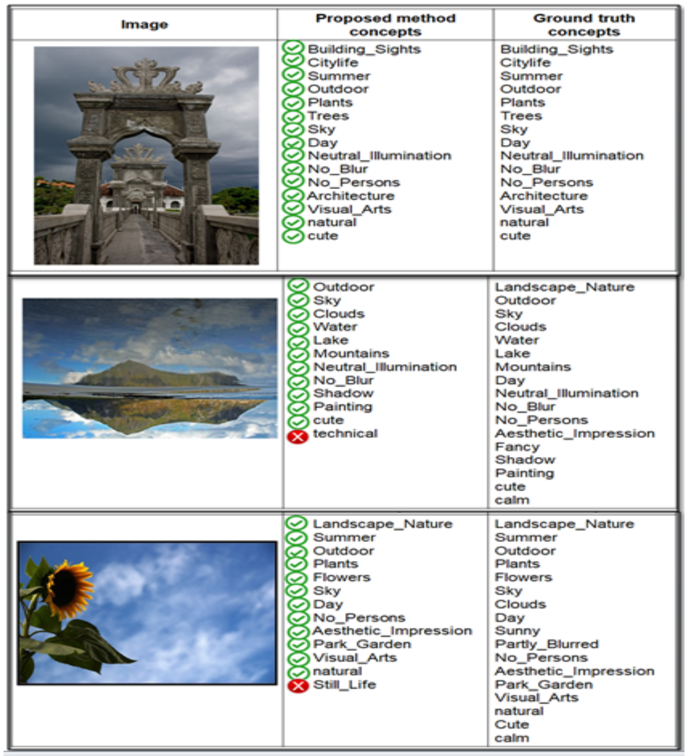


Fig. 5. Example of images annotated with the proposed method. The predicted concepts are compared to the ground truth concepts

TABLE I. Results of the proposed method and participants in ImageCLEF 2011 in terms of MiAP and F-Measure

Method	MiAP	F-Measure
MUFIN	0.299001	0.461820
MRIM	0.377179	0.552276
BPACAD	0.436294	0.593088
IDMT	0.370975	0.551224
ISIS	0.432758	0.622038
LIRIS	0.436968	0.566935
MLKD	0.401642	0.558795
TUBFI	0.443449	0.565980
Our method	**0.676530**	**0.799936**

TABLE II. Results of the proposed method and state-of-the-art methods for ImageCLEF 2011 in terms of MiAP

Method	MiAP
Multimodal method proposed in [21]	43.6 %
Multimodal method proposed in [12]	45.3 %
Multimodal method proposed in [6]	44.8 %
Our method	**67.6 %**

TABLE III. Comparison of the proposed method to results of participants in ImageCLEF 2012 photo annotation task in terms of MiAP

	Method	MiAP
Participants in ImageCLEF 2012 photo annotation task	BUAA AUDR	13.07 %
	CEA LIST	40.86 %
	CERTH	32.10 %
	DMS-SZTAKI	42.58 %
	ISI	41.36 %
	KIDS NUTN	17.17 %
	LIRIS	**43.67 %**
	MLKD	31.85 %
	UAIC	18.63 %
	UNED	7.56 %
	URJCyUNED	6.22 %
State-of the-art multimodal method	Multimodal method proposed in [6]	43.1 %
Our method		32.15 %

VI. Conclusion

A new multimodal annotation method is proposed. The method relies on the use of association rules mining and clustering; where the text and visual clusters obtained as output from the clustering algorithms are associated using the association rules mining. In order to evaluate the performance of the proposed method, two available and widely used datasets were considered to carry out the experiments. The results achieved with the proposed method outperforms all or most of the considered multimodal methods. The following improvements are considered as future work; the linking of tags to concepts using WordNet to find the semantic similarity between tags and concepts, the use of local visual features instead of global features. Improvement of text pre-processing and text features extraction in the training phase is necessary as well.

Acknowledgment

We thank King Abdulaziz City for Science and Technology for funding this research (Grant No. 1-17-02-009-0003).

References

[1] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, December 2000.

[2] Z. Mehmood, T. Mahmood, and M. A. Javid, “Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine,” *Applied Intelligence*, vol. 48, no. 1, pp. 166–181, 2018.

[3] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán, “Multiscale conditional random fields for image labeling,” in *the Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, vol. 2. IEEE, June 2004, pp. 695–702.

[4] G. Carneiro and N. Vasconcelos, “Formulating semantic image annotation as a supervised learning problem,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 163–168.

[5] S. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation,” in *the Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*. IEEE, July 2004, pp. 1002–1009.

[6] A. Znaidia, “Handling imperfections for multimodal image annotation,” Ph.D. dissertation, Ecole Centrale Paris, 2014.

[7] W. B. Li, J. Min, and G. J. Jones, “A text-based approach to the imageclef 2010 photo annotation task,” September 2010.

[8] B. Sigurbjörnsson and R. Van Zwol, “Flickr tag recommendation based on collective knowledge,” in *the Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 327–336.

[9] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, “Image annotation refinement using random walk with restarts,” in *the Proceedings of the 14th ACM international conference on Multimedia*. ACM, 2006, pp. 647–650.

[10] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in *the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 119–126.

[11] S.-B. Chan, H. Yamana, D.-D. Le, and S. Satoh, “Image annotation fusing content-based and tag-based technique using support vector machine and vector space model,” in *the Proceedings of the 10th International Conference on Signal-Image Technology and Internet-Based Systems*. IEEE, April 2014, pp. 272–276.

[12] Y. Zhang, S. Bres, and L. Chen, "Semantic bag-of-words models for visual concept detection and annotation," in *the Proceedings of the 8th International Conference on Signal Image Technology and Internet Based Systems*. IEEE, January 2012, pp. 289–295.

[13] C.-m. Huang, Y.-S. Lee, C.-Y. Lin, and C.-Y. Chen, "Using LSA and Association Rules to Enhance Web Image Annotation," in *the Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 04 2011, p. 1.

[14] S. Nowak, K. Nagel, and J. Liebetrau, "The imageclef2011 photo annotation and concept-based retrieval tasks," in *CLEF (Notebook Papers/Labs/Workshop)*, 2011, pp. 1–25.

[15] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *the Proceedings of the European conference on computer vision*, vol. 2353. Springer, April 2002, pp. 97–112.

[16] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 127–134.

[17] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *the Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management*. Citeseer, 1999, pp. 1–9.

[18] M. Bakalem, N. Benblidia, and S. Ait-Aoudia, "A novel image auto-annotation based on blobs annotation," in *Image Processing and Communications Challenges 3*. Springer, 2011, vol. 102, pp. 113–122.

[19] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, August 2015.

[20] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *the Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[21] N. Liu, E. Dellandréa, L. Chen, C. Zhu, Y. Zhang, C.-E. Bichot, S. Bres, and B. Tellez, "Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 493–512, May 2013.

[22] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *European conference on computer vision*. Springer, 2008, pp. 316–329.

[23] R. He, N. Xiong, L. T. Yang, and J. H. Park, "Using multi-modal semantic association rules to fuse keywords and visual features automatically for web image retrieval," *Information Fusion*, vol. 12, no. 3, pp. 223–230, 2011.

[24] M. Taileb, S. Lamrous, and S. Touati, "Non-overlapping hierarchical index structure for similarity search," *International Journal of Computer Science*, vol. 3, no. 1, 2007.

[25] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[26] R. A. Alghamdi, M. Taileb, and M. Ameen, "A new multimodal fusion method based on association rules mining for image retrieval," in *the Proceeding of 17th IEEE Mediterranean Electrotechnical Conference (MELECON). Beirut, Lebanon*. IEEE, April 2014, pp. 493–499.

[27] R. Agrawal, R. Srikant *et al.*, "Fast Algorithms for Mining Association Rules in Large Databases," in *the Proceeding of the 20th International Conference on Very Large Data Bases*, 1994, pp. 487–499.

Comparison between Commensurate and Non-commensurate Fractional Systems

Khaled HCHEICHI and Faouzi BOUANI
University Tunis EL Manar
National Engineering School of Tunis
Analysis, Conception and Control of Systems Laboratory, Tunis, Tunisia

***Abstract*—This article deals with fractional systems that represent better physical process and guarantee a very small number of parameters that can reduces the computation time. It focuses in particular on the state-space representation which highlights the state variables and allows to study the internal behavior of the system taking into account the initial state. Moreover, this representation adapts better to the multiple input multiple-output case. It also discusses the discretization of fractional system to finally adapt the Model Predictive Control to apply it and shows its efficiency and performance in these systems. The main objective of this article is to compare the commensurate and non-commensurate fractional models performance, calculation time and ease of use.**

Keywords*—*discretization; state-space; fractional; calculation time

I. Introduction

Over the last few years, many theoretical and practical contributions showed the importance of fractional systems and their interest in different applications of modeling, identification and control of physical systems. Several control strategies have been developed and adapted for fractional systems such as the PID controller, sliding mode control and predictive control. In different disciplines such as electricity, chemistry, biology, economics, automation and signal processing. Some researchers have compared the fractional models with the classical integer models. The results have shown that fractional model represents systems with much lower number of parameters than those of integer model.

The most studied physical phenomena and most used during modeling or identification are mentioned in [1].

One of these phenomena is the attenuation of the movement of water on dikes, especially those with cavities or depressions trapping air pockets that can be compressed by water.

There is also the viscoelasticity of the materials having mechanical properties strongly dependent on the frequency over many decades of frequency, where the number of parameters is very large. As a result, calculations on the model take time and produce high order differential equations. Which is solved by using a fractional model with a number of paramtrestres very reduced in [2].

Ref [3] treated another phenomenon that requires the use of the fractional model which is the Randles model that is frequently used in the literature for modeling lead-acid batteries. This model results from a simplified solution of the electrochemical diffusion equation in the batteries (Fick's law).

To finish, the diffusion of heat in a semi-infinite environment subjected to a heat flux $q(t)$ on its surface boundary S. In [4, 5], authors demonstrate that the mathematical equations describing the unidirectional heat transfer in the environment reveal a real order of derivation of the temperature at the point of abscissa $x = 0$.

Almost all articles dealing with fractional systems focus on commensurate order systems and the developed controls are based on the transfer function model. But in practice the identification programs give better results in free identification. Therefore the orders of systems can be arbitrary, this type is named non-commensurate order. This can impose problem because in the case of non-commensurate the use of the classic tools becomes imopossible, moreover the programming becomes more complicated.

This paper compares the commensurate and non-commensurate fractional system discussing the advantages and the disadvantages of each model. The models used in this paper are in the form of a state-sapce representations that make it easier to study the internal behavior of systems and adapt better to MIMO systems.

The first section of this article quotes the different representations of fractional systems as well as the transition between them. A second section explores the method used to pass from a continuous model to a discrete model. Section 3 will present the predictive control that will be applied on the fractional model. In Section 5 we have implemented different methods to compare commensurate and non-commensurate fractional systems.

II. General Fractional System Representation

A generalized fractional system can be represented by the following equation [6, 7]:

$$y(t) + \sum_{i=1}^{n} a_i D^{\alpha_i} y(t) = \sum_{i=1}^{m} b_i D^{\beta_i} u(t) \tag{1}$$

As the system is relaxed $y(t) = u(t) = 0$ for $t \leqslant 0$, the Laplace transform of $D^{\alpha_i} y(t)$ and $D^{\beta_i} u(t)$ are respectively $s^{\alpha_i} Y(s)$ and $s^{\beta_i} U(s)$, where $Y(s)$ and $U(s)$ are the Laplace transforms of $y(t)$ and $u(t)$. Applying the Laplace transformation to (1) we obtain:

$$Y(s) + \sum_{i=1}^{n} a_i s^{\alpha_i} Y(s) = \sum_{i=1}^{m} b_i s^{\beta_i} U(s) \tag{2}$$

The transfer function can be deduced from the previous equation:

$$G(s) = \frac{Y(s)}{U(s)} = \frac{\sum_{i=1}^{m} b_i s^{\beta_i}}{1 + \sum_{i=1}^{n} a_i s^{\alpha_i}} \tag{3}$$

The generalized state space model corresponding to the multivariable transfer function is:

$$\begin{cases} D^{(\alpha)}(x) &= Ax + Bu \\ y &= Cx + Du \end{cases} \tag{4}$$

Where:

$$D^{(\alpha)}(x) = \left[\frac{d^{\alpha_1}}{dt^{\alpha_1}}x_1, \frac{d^{\alpha_2}}{dt^{\alpha_2}}x_2 \cdots \frac{d^{\alpha_n}}{dt^{\alpha_n}}x_n\right]^T \tag{5}$$

The passage from the state-space representation to the transfer-function representation can be deduced as in the integer case by taking the Laplace transform and considering the zero initial conditions [8], it is given by:

$$G(s) = C\left[\left(s^{(\alpha)} I_n - A\right)^{-1}\right] B + D \tag{6}$$

On the other hand, to go from the transfer function to the state representation there is a difference. This difference is due to the presence of fractional derivation orders that must be taken into account during transformation. As mentioned, the transfer function of fractional systems is represented as follows:

$$G(s) = \frac{b_m s^{\beta_m} + b_{m-1} s^{\beta_{m-1}} + \cdots + b_1 s^{\beta_1} + b_0}{s^{\alpha_n} + a_1 s^{\alpha_{n-1}} + \cdots + a_{n-1} s^{\alpha_1} + a_n} \tag{7}$$

With

$$\alpha_n > \alpha_{n-1} > \cdots > \alpha_2 > \alpha_1$$

and

$$\beta_m > \beta_{m-1} > \cdots > \beta_2 > \beta_1$$

Let $\tilde{\alpha}$ be the vector obtained from the concatenation of the fractional numbers α_i and β_i such that:

$$\tilde{\alpha} = \left[\tilde{\alpha}_{n+m}\ \tilde{\alpha}_{n+m-1}\ \tilde{\alpha}_{n+m-2}\ \cdots\ \tilde{\alpha}_2\ \tilde{\alpha}_1\right] \tag{8}$$

With

$$\tilde{\alpha}_{n+m} > \tilde{\alpha}_{n+m-1} > \tilde{\alpha}_{n+m-2} > \cdots > \tilde{\alpha}_2 > \tilde{\alpha}_1$$

Let consider the continuous state-space model given by:

$$\begin{cases} D^{\gamma}(x) &= A_c x + B_c u \\ y &= C_c x \end{cases} \tag{9}$$

with

$$A_c = \begin{pmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 0 & 1 \\ \tilde{a}_{n+m} & \tilde{a}_{n+m-1} & \cdots & \tilde{a}_2 & \tilde{a}_1 \end{pmatrix}, \qquad B_c = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

$$C_c = \begin{pmatrix} \tilde{c}_1 & \tilde{c}_2 & \cdots & \tilde{c}_{n+m-1} & \tilde{c}_{n+m} \end{pmatrix}$$

and

$$D^{\gamma}x = \left[D^{\tilde{\alpha}_1}x_1 \quad \cdots \quad D^{(\tilde{\alpha}_{n+m} - \tilde{\alpha}_{n+m-1})}x_{n+m}\right] \tag{10}$$

Where the corresponding transfer function model $H(s)$ is given by:

$$H(s) = \frac{\tilde{c}_1 + \tilde{c}_2 s^{\tilde{\alpha}_1} + \tilde{c}_3 s^{\tilde{\alpha}_1 + \tilde{\alpha}_2} + \cdots}{s^{\tilde{\alpha}_1 + \tilde{\alpha}_2 + \cdots + \tilde{\alpha}_{n+m}} + \cdots} \frac{\cdots + \tilde{c}_{n+m} s^{\tilde{\alpha}_1 + \tilde{\alpha}_2 + \cdots + \tilde{\alpha}_{n+m-1}}}{\cdots + \tilde{a}_{n+m-2} s^{\tilde{\alpha}_1 + \tilde{\alpha}_2} + \tilde{a}_{n+m-1} s^{\tilde{\alpha}_1} \tilde{a}_{n+m}} \tag{11}$$

Since $H(s)$ numerator and denominator contain $(n+m)$ terms, it is then sufficient to sort them in order to isolate m terms for witch the fractional orders correspond to those of $G(s)$ numerator and n terms for witch the fractional orders correspond to those of $G(s)$ denominator. The procedure of selection of the terms $\tilde{c}_i$ and $\tilde{a}_i$ is summarized in (12).

$$\begin{cases} \tilde{a}_{n+m} = a_n \qquad \tilde{c}_1 = b_0 \\ if\ \tilde{\alpha}_i = \beta_j \quad then\ \tilde{c}_{i+1} = b_j\ and\ \tilde{a}_{n+m-i} = 0 \\ if\ \tilde{\alpha}_i = \alpha_j \quad then\ \tilde{c}_{i+1} = 0\ and\ \tilde{a}_{n+m-i} = a_{n-j} \end{cases} \tag{12}$$

$$i = 1, \cdots, n+m-1 \quad \text{and} \quad j = 1, \cdots, n+m-1$$

In this case of generalized fractional systems, the number of state variables is equal to the sum of dimensions of the numerator and denominator polynomials of the transfer function $(n+m)$.

III. Discritization of Fractional State-Space Model

Contrary to the commensurate case, discretization in the case of non-commensurate fractional systems must take into account the plurality of derivations of state variables.

To move from a continuous model to a discrete model it is necessary to use this approximation [9, 10, 11]:

$$D^{\gamma}x(t) = \frac{1}{T_s^{\gamma}} \sum_{j=0}^{p} (-1)^j \binom{\gamma}{j} x((k-j)T_s) \tag{13}$$

Let's assume that the vector of continuous model derivation $\gamma = \begin{bmatrix}\gamma_1 & \gamma_2 & \cdots & \gamma_r\end{bmatrix}^T$, T_s is the sampling time and $p \in \mathbb{N}$ is the number of past samples with which the derivation was computed.

If $(i = 1,\ \cdots,\ r)$, the term $\binom{\gamma}{j}$ can be written as follows:

$$\binom{\gamma}{j}^T = \left[\binom{\gamma_1}{j} \quad \binom{\gamma_2}{j} \quad \cdots \quad \binom{\gamma_r}{j}\right] \tag{14}$$

$$\binom{\gamma_i}{j} = \begin{cases} 1 & for \quad j = 0 \\ \dfrac{\gamma_i(\gamma_i - 1)...(\gamma_i - j + 1)}{j!} & for \quad j > 0 \end{cases} \tag{15}$$

By multiplying (13) by T_s^γ and developing the terms of $j=0$ and $j=1$ the following result is found:

$$T_s^\gamma D^\gamma x(t) = x(kT_s) - \gamma x((k-1)T_s) + + \sum_{j=2}^{p} (-1)^j \binom{\gamma}{j} x((k-j)T_s) \tag{16}$$

Now let consider the following continuous fractional state-space model [12]:

$$\begin{cases} D^\gamma x(t) &= A_c x(t) + B_c u(t) \\ y(t) &= C_c x(t) \end{cases} \tag{17}$$

With $A_c \in \mathbb{R}^{r\times r}$, $B_c \in \mathbb{R}^{r\times 1}$ and $C_c \in \mathbb{R}^{1\times r}$ are the state matrices of the continuous fractional model and r is the number of varibales in state-space model.

$$T_s^\gamma A_c x(kT_s) - x(kT_s) = -\gamma x((k-1)T_s) + + \sum_{j=2}^{p} (-1)^j \binom{\gamma}{j} x((k-j)T_s) - B_c T_s u(kT_s) \tag{18}$$

Note that $I_r \in \mathbb{R}^{r\times r}$ the identity matrix and T_s^γ the diagonal matrix filled by ($T_s^{\gamma_1} \cdots T_s^{\gamma_r}$).

To facilitate writing, note

$$\mathbb{Z} = (T_s^\gamma A_c - I_r)^{-1} \tag{19}$$

$$x(kT_s) = -\mathbb{Z}\gamma x((k-1)T_s) + \mathbb{Z}\sum_{j=2}^{p} (-1)^j \binom{\gamma}{j} x((k-j)T_s) - \mathbb{Z} B_c T_s^\gamma u(kT_s) \tag{20}$$

and with ($i = 1, \cdots, r$)

$$c_j = diag\{(-1)^j \binom{\gamma_i}{j} \tag{21}$$

The above equation can be written as:

$$x(k) = \mathbb{Z}\, c_1 x(k-1) + \mathbb{Z}\sum_{j=2}^{p} c_j x(k-j) - \mathbb{Z}\, B_c T_s^\gamma u(k) \tag{22}$$

To simplify the equation:

$$A_j = \mathbb{Z}\, c_j \tag{23}$$

By expanding all terms and simplifying, (22) becomes in the form:

$$x(k) = A_1 x(k-1) + A_2 x(k-2) + \cdots + A_k x(0) - \mathbb{Z} B_c T_s^\gamma u(k) \tag{24}$$

The system can therefore be described by a discrete state-space representation [13]:

$$\begin{cases} X_d(k+1) = A_d X_d(k) + B_d u(k) \\ y(k) = C_d X_d(k) \end{cases} \tag{25}$$

With

$$A_d = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} \\ I & 0 & \cdots & 0 \\ 0 & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{pmatrix}, \quad B_d = \begin{pmatrix} -\mathbb{Z}\, B_c T_s^\gamma \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$C_d = (C \quad 0 \quad \cdots \quad 0)$$

$$X_d(k+1) = \begin{pmatrix} x(k+1) \\ x(k) \\ \vdots \\ x(k-p+1) \end{pmatrix} \quad \text{and}$$

$$X_d(k) = \begin{pmatrix} x(k) \\ x(k-1) \\ \vdots \\ x(k-p) \end{pmatrix}$$

with u, y, X_d are respectively the input, output and variables state of the process. p is the number of past iterations which the system takes into account for calculating a variable, $A_d \in \mathbb{R}^{rp\times rp}$, $B_d \in \mathbb{R}^{rp\times 1}$, $C_d \in \mathbb{R}^{1\times rp}$ and $X_d \in \mathbb{R}^{rp\times 1}$.

By increasing the number of iterations taken into account p the computation time increases, so it is reasonable to choose a number large enough to represent the system correctly but not too large to reduce the calculation time.

IV. Fractional Model Predictive Control

The principle of the predictive control is to create an anticipatory effect for the system with respecting the trajectory to follow known in advance, based on the prediction of the future behavior of the system and minimizing the gap of these predictions to the trajectory and by minimizing a certain cost function J, while respecting operating constraints [14, 15, 16].

This section will develop a predictive control from the discrete fractional state-space model described in previous section. For that we will make a variable change : $\Delta X_d(k) = X_d(k) - X_d(k-1)$ the input variable difference: $\Delta u(k) = u(k) - u(k-1)$, and using it in (25) this transformation is found:

$$\Delta X_d(k+1) = A_d \Delta X_d(k) + B_d \Delta u(k) \tag{26}$$

The new state variable vector is:

$$X(k) = [\Delta X_d(k)^T \, y(k)]^T$$

with $y(k)$ is the output and:

$$y(k+1) - y(k) = C_d A_d \Delta X_d(k) + C_d B_d \Delta u(k) \tag{27}$$

The system can be written in the form:

$$\begin{cases} X(k+1) = AX(k) + B\Delta u(k) \\ y(k) = CX(k) \end{cases} \tag{28}$$

$$A = \begin{pmatrix} A_d & 0_d^T \\ C_d A_d & 1 \end{pmatrix}; \; B = \begin{pmatrix} B_d \\ C_d B_d \end{pmatrix};$$

$$C = (0_d \quad 1);\ 0_d \in \mathbb{R}^{1\times rp}$$

Future state variables can be predicted and written in the form:

$$\begin{cases} X(k+1) &= AX(k) + B\Delta u(k) \\ X(k+2) &= AX(k+1) + B\Delta u(k+1) \\ &= A^2X(k) + AB\Delta u(k) + B\Delta u(k+1) \\ &\vdots \\ X(k+H_p) &= A^{H_p}X(k) + A^{H_p-1}B\Delta u(k) + \\ & A^{H_p-2}B\Delta u(k+1) + \cdots \\ & +A^{H_p-H_c}B\Delta u(k+H_c-1) \end{cases} \tag{29}$$

Based on (29) future system outputs can be predicted:

$$\begin{cases} y(k+1) &= CAX(k) + CB\Delta u(k) \\ y(k+2) &= CAX(k+1) + CB\Delta u(k+1) \\ &= CA^2X(k) + CAB\Delta u(k) + \\ & CB\Delta u(k+1) \\ &\vdots \\ y(k+H_p) &= CA^{H_p}X(k) + CA^{H_p-1}B\Delta u(k) + \\ & CA^{H_p-2}B\Delta u(k+1) + \cdots \\ & +CA^{H_p-H_c}B\Delta u(k+H_c-1) \end{cases} \tag{30}$$

H_p and H_c are respectively the prediction horizon and the control horizon with $H_p \geq H_c$. Assume the vector Y which contains H_p system's predicted future outputs and Δu contains H_c future controls:

$$Y^T = [y(k+1)\ y(k+2) \cdots y(k+H_p)]$$

$$\Delta u^T = [\Delta u(k)\ \Delta u(k+1)\ \cdots \Delta u(k+H_c-1)]$$

The vector Y can also be written as :

$$Y = FX(k) + \Phi\Delta u \tag{31}$$

$$F = \begin{pmatrix} CA \\ CA^2 \\ CA^3 \\ \vdots \\ CA^{H_p} \end{pmatrix} \tag{32}$$

$$\Phi^T = \begin{pmatrix} CB & CAB & CA^2B & \cdots & CA^{H_p-1}B \\ 0 & CB & CAB & \cdots & CA^{H_p-2}B \\ 0 & 0 & CB & \cdots & CA^{H_p-3}B \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & CA^{H_p-H_c}B \end{pmatrix} \tag{33}$$

The aim of predictive control is to find the control vector Δu which forces the system's output y to follow the setpoint y_s. In order to achieve this we must optimize a criterion J which represents the control objective:

$$J = \sum_{i=1}^{H_p}(y_s(k+i) - y(k+i))^2 + \lambda \sum_{i=0}^{H_c-1} \Delta u^2(k+i) \tag{34}$$

The criterion J can be written in matrix form:

$$J = (Y_s - Y)^T(Y_s - Y) + \Delta u^T \lambda \Delta u \tag{35}$$

With $Y_s^T = [y_s(k+1)\ y_s(k+2) \cdots y_s(k+H_p)]$ is the vector filled by the future values of the set-points and λ is weight coefficient on the control.

By minimizing J we obtain optimal control sequence [17]:

$$\Delta u = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T(Y_s - FX(k)) \tag{36}$$

V. Simulation Results

In this section the matlab FOMCON toolbox is used to find a continuous fractional transfer function from the input-output data of a thermal system. The input signal used for identification is shown in the Fig.1. In this case we chose a free identification to find the better result.

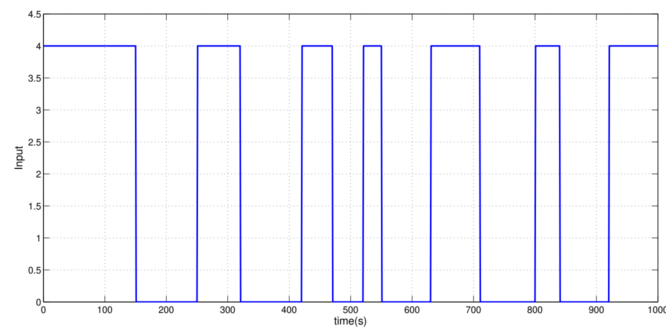


Fig. 1. System input.

The result obtained is compared with the output signal used in the identification, the Fig.2 illustrates the two signals as well as the error.

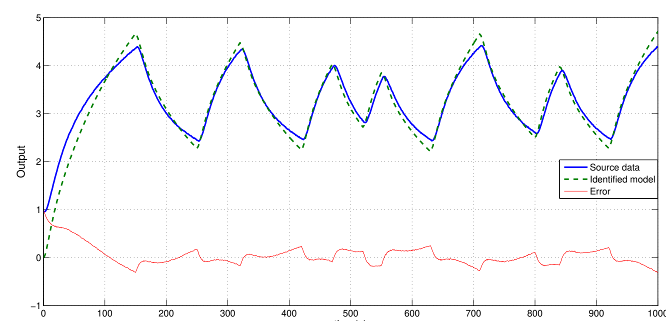


Fig. 2. Identification result.

This method gaves as result a continuous fractional transfer function (7)

$$G(s) = \frac{b_0}{s^{2.46} + \tilde{a}_1 s^{1.97} + \tilde{a}_2 s^{1.39} + \tilde{a}_3 s^{0.97} + \tilde{a}_4 s^{0.88} + \tilde{a}_5} \tag{37}$$

So the parameters are deduced: $b_0 = -0.0041$, $\tilde{a}_1 = -1.5977$, $\tilde{a}_2 = 0.551$, $\tilde{a}_3 = -0.3517$, $\tilde{a}_4 = -0.0011$and $\tilde{a}_5 = -0.002$.

By applying the above method, the founded result is the fractional state-space representation for previous transfer function and which is described by the matrices:

$$A_c = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0.002 & 0.0011 & 0.3517 & -0.551 & 1.5977 \end{pmatrix},$$

$$B_c = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \qquad C_c = \begin{pmatrix} -0.0041 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$\tilde{\alpha}_1 = 0.88$, $\tilde{\alpha}_2 = 0.97$, $\tilde{\alpha}_3 = 1.39$, $\tilde{\alpha}_4 = 1.97$ and $\tilde{\alpha}_5 = 2.46$.

and using (8) and (10) we can fill the vector γ by :

$$\gamma = \begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 - \tilde{\alpha}_1 \\ \tilde{\alpha}_3 - \tilde{\alpha}_2 \\ \tilde{\alpha}_4 - \tilde{\alpha}_3 \\ \tilde{\alpha}_5 - \tilde{\alpha}_4 \end{pmatrix}, \quad \text{then} \quad \gamma = \begin{pmatrix} 0.88 \\ 0.09 \\ 0.42 \\ 0.58 \\ 0.49 \end{pmatrix}$$

From transfer function (37) the commensurate model that can be found by matlab FOMCON toolbox has a very high order, in this case $r_c = 246$ and $\alpha = 0.01$ With $A_{c_c} \in \mathbb{R}^{r_c \times r_c}$, $B_{c_c} \in \mathbb{R}^{r_c \times 1}$ and $C_{c_c} \in \mathbb{R}^{1 \times r_c}$ are the state matrices of the continuous fractional model, α is the commensurate order and r_c is the number of varibales in state-space commensurate model.

For discretization the chosen sampling period is $T_s = 5s$ and the history will be limited to 20 past values so p = 20, these parameters will be the same for the rest of this paper.

In what follows (Fractional Model Predictive Control) FMPC will be applied for both models, the necessary parameters for the FMPC will be fixed for this section $H_p = 10$, $H_c = 1$ and $\lambda = 2$.

The system output (temperature) and the set-point are shown in the Fig.3 while the Fig.4 shows the control signal generated by the FMPC. A disturbance is added to the system in the interval of time $k \in [120, 150]$ to test the ability of FMPC to anticipate it, the amplitude of this disturbance is $+10\%$.

Fig.3 compares the output of the non-comensurate model with the commensurate model, these two models are deduced from the same transfer function presented by (37).

The Fig.3 and Fig.4 show that the FMPC is able to mitigate the effect of the disturbance on the output. On the other hand the Fig.5 shows peaks with a very large amplitude which appears in the control increment signal.

The figures show that the predictive control is able to force the system to follow the set-point in both cases, but it is clear that for the non-commensurate model the control expends less energy to achieve it.

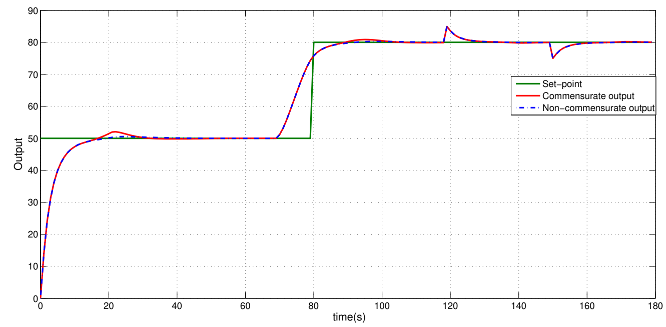


Fig. 3. Output with disturbance.

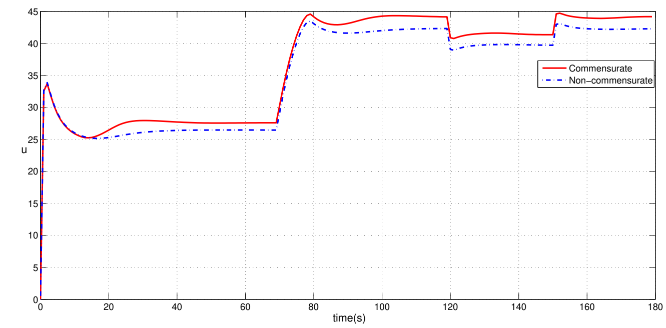


Fig. 4. Control with disturbance.

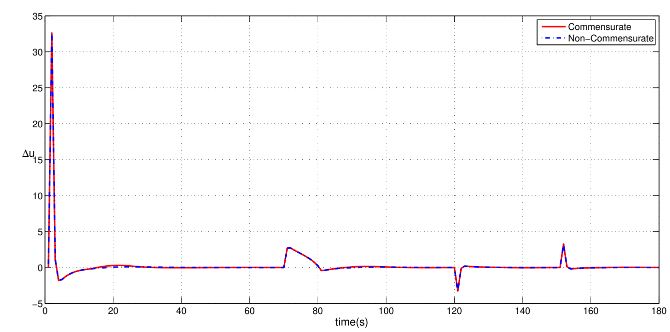


Fig. 5. Control increment with disturbance.

In the Fig.5, the control increases by a large amount instantly. To cancel this undesirable spike we will add a constraint to the calculation of control.

The principle of constraint control [18, 19] is to add to the main problem one or more conditions, it is therefore necessary to minimize criterion J while respecting these conditions. The most used constraint form is the one on the rate of change of the control variables $\Delta u(k)$. Suppose that for a single-input system the upper limit is Δu_{max} and the lower limit is Δu_{min}. The constraints are specified in the form:

$$\Delta u_{min} \le \Delta u(k) \le \Delta u_{max} \tag{38}$$

Assuming that the control variable $\Delta u(k)$ can only increase or decrease in a unit of magnitude less than 2.5, the operational constraint is:

$$-2.5 \le \Delta u(k) \le 2.5$$

Now let's test the performance of the system under constraint and with a disturbance of amplitude $+10\%$ in the interval of time $k \in [120,\ 150]$, the constraint is the same used before.

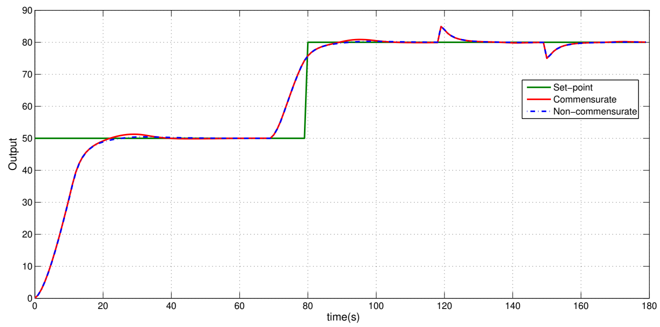


Fig. 6. Output with constraint and disturbance.

Even under constraint the FMPC can ensure that the output (temperature) follows the set-point. Constrains on the control increment guarantees that there is no peaks in the control signal. In return, the output porsuite becomes slower as shown in the Fig.6.

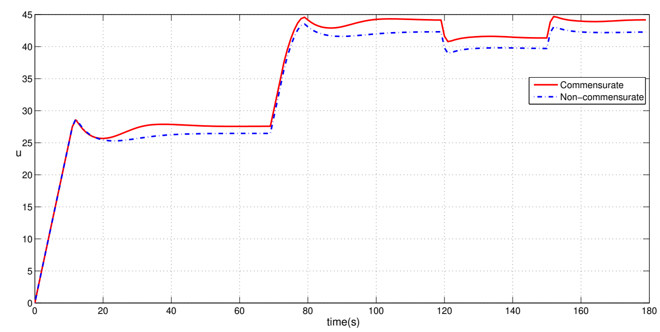


Fig. 7. Control with constraint and disturbance.

The choice of the the interval $[\Delta u_{min},\ \Delta u_{max}]$ is very important because if the interval is too wide the condition will not be taken into account when minimizing criterion J, and if the interval is too small the control will no longer be able to bring the output to follow the set-point, even if it happen the system will be too slow.

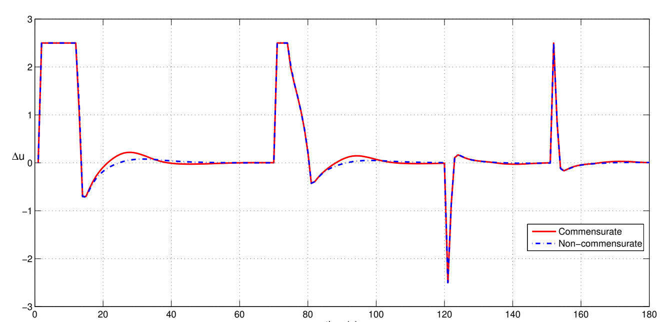


Fig. 8. Control increment with constraint and disturbance.

Fig.7 and Fig.8 show that the constraint is taken into account in optimisation and show that the non-commensurate model still expends less energy than the commensurate model.

One of the disadvantages of fractional systems is the increase of the dimensions of the matrices at each iteration. For example if $p = 20$, the calculation of the state variables will depend on 20 previous values. The dimensions of the matrices will increase to multiply by p. This increase has a negative effect on the calculation time which increases considerably.

Now let's focus on the calculation time for both models. This part compares the time spent on each iteration to calculate the control increment Δu for different value of p. In this simulation the used processor is a $i3$ with $1.9\ GHz$ frequency, the results are presented in Table I:

TABLE I. CALCULATE TIME (MS)

p	20	50	100	180
Non-commensurate	22.7	23.6	27.9	30.5
Commensurate	53.2	56.1	61.5	74.1

Even if fractional models better represent physical systems, their use can be complicated. The commensurate model has a very high order $r_c = 246$ whereas in the case non-commensurate model $r = 5$, this difference is manifested in the computation time.

These results show that the use of the commensurate model can provide the same performance as non-commensurate model with a longer computing time. The non-commensurate model consecutively reduces the calculation time for the control because the matrices have a smaller dimension. On the other hand, the determination of a non-commensurate model is more complicated than the commensurate model.

VI. CONCLUSION

The use of fractional models becomes more and more frequent given the efficiency they provide in the discription of certain physical systems. Nevertheless it remains a little difficult to handle. The majority of research [20, 21, 22] deals with commensurate fractional models that have proved efficiency at describing several physical phenomena, but the major inconvenience is that computation time increases in discrete models because of the use of history in the calculation. This article has shown that the limitation of the used history and the use of non-commensurate models in modeling or identification can remedy this problem. It has also adapted predictive control to apply it to a non-commensurate fractional system. The importance of this work is that it deals with the state-space representation of the non-commensurate fractional systems from identification to control. It also compares the use of commensurate and non-commensurate fractional systems and explores the advantages and disadvantages of each model. At first it introduced the transition between the transfer-function and the state-space representation in the non-commensurate fractional case. Then for the same kind of system, it explained the discretization and we closed by a comparison with the commensurate model in therm of performance and computation time.

REFERENCES

[1] A. Oustaloup, J. Sabatier, P. Lanusse, et al. "An overview of the CRONE approach in system analysis, modeling and identification, observation and control" *IFAC Proceedings Volumes: 17th IFAC World Congress*, 2008, 41 (2): 14254 - 14265

[2] H. Sun, Y. Zhang, D. Baleanu, W. Chen and Y. Chen ”A new collection of real world applications of fractional calculus in science and engineering” *Communications in Nonlinear Science and Numerical Simulation*, 2018, 64: 213 – 231

[3] T.J. Machado, M. Silva, R. Barbosa, et al. ”Some Applications of Fractional Calculus in Engineering” *Mathematical Problems in Engineering* 2010: 34 pages

[4] S. Blasiak ”Time-fractional heat transfer equations in modeling of the non-contacting face seals” *International Journal of Heat and Mass Transfer*, 2016, 100: 79 – 88

[5] L. Ouhsaine, Y. Boukal, M. El-Ganaoui et al. ”A General fractional-order heat transfer model for Photovoltaic/Thermal hybrid systems and its observer design” *Energy Procedia*, 2017, 139: 49 – 54

[6] C.F. Lorenzo and T.T. Hartley ”Initialization in fractional order systems” In: *European Control Conference (ECC)*, Porto, Portugal, September 2001, pp. 1471–1476

[7] A. Rhouma and F. Bouani ”Robust model predictive control of uncertain fractional systems: a thermal application” *IET Control Theory Applications*, 2014, 8 (17): 1986–1994

[8] C.A. Monje, Y. Chen, B.M. Vinagre, D. Xue and V.Feliu-Batlle, (2010) ”Fractional-order Systems and Controls: Fundamentals and Applications.” London: Springer-Verlag, 1st edn

[9] Z. Belkhatir and T.M. Laleg-Kirati ”Parameters and fractional differentiation orders estimation for linear continuous-time non-commensurate fractional order systems” *Systems and Control Letters* , 2018, 115: 26 – 33

[10] A. Dzieliǹski and D. Sierociuk ”Adaptive Feedback Control of Fractional Order Discrete State-Space Systems.” 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*, Vienna, Austria, December 2011, pp. 1–6

[11] P. Ostalczyki and D. Mozyrska ”The second form of the variable-, fractional-order discrete-time integrator.” In: *21st International Conference on Methods and Models in Automation and Robotics (MMAR)*, Miedzyzdroje, Poland, August 2016, pp. 859–864

[12] A. Dzieliǹski and D. Sierociuk ”Ultracapacitor modelling and control using discrete fractional order state-space models and Fractional Kalman Filters.” In: *European Control Conference (ECC)*, Kos, Greece, July 2007, pp. 2916–2922

[13] M. Abedini, M.A. Nojoumian, H.Salarieh and A. Meghdari ”Model reference adaptive control in fractional order systems using discrete-time approximation methods” *Communications in Nonlinear Science and Numerical Simulation*, 2015, 25 (1-3): 27 – 40

[14] H. Aboukheir ”Predictive Control of Fractional Order Systems” *IFAC Proceedings Volumes*, 2012, 45 (13): 622–626

[15] M.M. Joshi, V.A. Vyawahare and M.D. Patil ”Model predictive control for fractional-order system a modeling and approximation based analysis.” In: *International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH)*, Vienna, Austria, August 2014, pp. 361–372

[16] J.B. Rawlings ”Tutorial overview of model predictive control” *IEEE Control Systems*, 2000, 20 (3): 38–52

[17] A. Rhouma, F. Bouani, B. Bouzouita and M. Ksouri ”Model Predictive Control of Fractional Order Systems” *Journal of Computational and Nonlinear Dynamics*, 2014, 9: 310111–310117

[18] B. Sikora and J. Klamka ”Constrained controllability of fractional linear systems with delays in control” *Systems and Control Letters*, 2017, 106: 9 – 15

[19] Q. Zou, Q. Jin and R. Zhang ”Design of fractional order predictive functional control for fractional industrial processes” *Chemometrics and Intelligent Laboratory Systems*, 2016, 152: 34 – 41

[20] Z. Belkhatir and T.M. Laleg-Kirati ”High-order sliding mode observer for fractional commensurate linear systems with unknown input” *Automatica*, 2017, 82: 209 – 217

[21] R. Martìnez-Guerra, C.D. Cruz-Ancona and C.A. Pérez-Pinacho ”Estimators for a class of commensurate fractional order systems with caputo derivative.” In: *14th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, Mexico City, Mexico, October 2017

[22] S. Saxena, V. Yogesh and P.P. ”Arya Reduced-order modeling of commensurate fractional-order system.” In: *14th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Porto, Portugal, Nov 2016, pp. 1–6

A Case Study for the IONEX CODE-Database Processing Tool Software: Ionospheric Anomalies before the M$_w$ 8.2 Earthquake in Mexico on September 7, 2017

Guillermo Wenceslao Zarate Segura, Carlos Sotomayor-Beltran
Imaging Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades
Lima, Peru

***Abstract*—A software tool was developed in the Imaging Processing Research laboratory (INTI-Lab) that automatically downloads several IONEX files around a specific user input date and also performs statistical calculations to look for ionospheric anomalies through the generation of differential vertical total electron content (ΔVTEC) maps. The IONEX CODE-Database Processing Tool (ICPT) software allows to save a considerable amount of time spended in gathering the necessary IONosphere map EXchange (IONEX) files for the production of differential VTEC maps. Using the ICPT software we were able to detect ionospheric anomalies before the devastating earthquake that happened in Mexico on September 7, 2017. A positive and negative ionospheric anomalies were detected nine days and one day before the seismic event. Due to stable geomagnetic conditions we suggest that the anomalies are assocciated to the earthquake event. Furthermore, It is very likely that the collision between the North American and Coco's plate is producing the ionization necessary of the air to generate the disturbances observed.**

Keywords*—Earthquake; IONEX CODE-Database Processing Tool; ionospheric anomalies; geomagnetic storm; global ionospheric maps(GIM); IONosphere map EXchange (IONEX)

I. Introduction

Several natural disasters arise around the globe from time to time and even in this era of high-performing computers devices, we cannot still predict the location of an earthquake. However, global ionospheric maps (GIMs) may provide some relevant information [1], [2], [3] that may indicate some possible relation between ionospheric anomalies and the occurrence of an earthquake . Due to the fact that there are several earthquakes around the globe and most of them occurring near populated areas, we believe it is necesary to keep researching these topics so in the near future we may predict seismic events and save uncountable lives.

Ionospheric anomalies are not 100% caused by earthquakes, they can also by generated by, for example, geomagnetic storms and also even by humans [4]. In this study we focused in the earthquake that happened in Mexico on September 7, 2017 at 04:49 UTC with a moment magnitude of 8.2. Using the IONEX CODE-Database Processing Tool (ICPT) software developed by the two authors of this study, we generate differential vertical total electron content (VTEC) maps from GIMs.

In this paper, differential VTEC (ΔVTEC) maps around the date of the Mexico seismic event are presented. ICPT software is a useful tool to quick detect any anomalies in some days close to the event day.Futhermore, It quickly performs multiple tasks to finally plot ΔVTEC maps. Finally, the detected ionospheric disturbances are shown and briefly discuss their possible origin.

II. ICPT Software

The software tool automatically downloads IONEX files for an interval of days selected according to a certain date given by the user [5].

The IONEX files are provided in a web ftp server (ftp.aiub.unibe.ch/CODE/) which commonly has hundreds of files in different folders arranged per year (e.g., for the year 2017 the path folder would be: 2017/). In this study, It was builded a software to enhance the productivity time (i.e., saving it) rather than spending it in downloading manually several files. Additionally, the ICPT software carries out a statistical process [5] to generate ΔVTEC maps .

Even though this process could be done by any user manually, there are some issues before downloading a IONEX file. The desired date should be translated into the IONEX labelling file name which is the current day number of the year.

A detailed graphical description of what the ICTP software does step-by-step is presented in Fig. 1. In reference [5] describes in details each step of the software tool.The ICPT software is avaible at https://sourceforge.net/projects/ionex-processing-icpt/files/ to download for free.

A. Statistical method to produce differential VTEC maps

The statistical method to find and analize ionospheric anomalies requires to process a group of IONEX files with continuous dates. Assuming the VTEC values follow a normal distribution within a range of days, a sliding window is then applied to obtain the VTEC mean value (μ) and the interquartile range (IQR). ICTP performs the task downloading 17 days before and 11 days after the given date chosen by the user. In this range, ICTP calculates the VTEC means

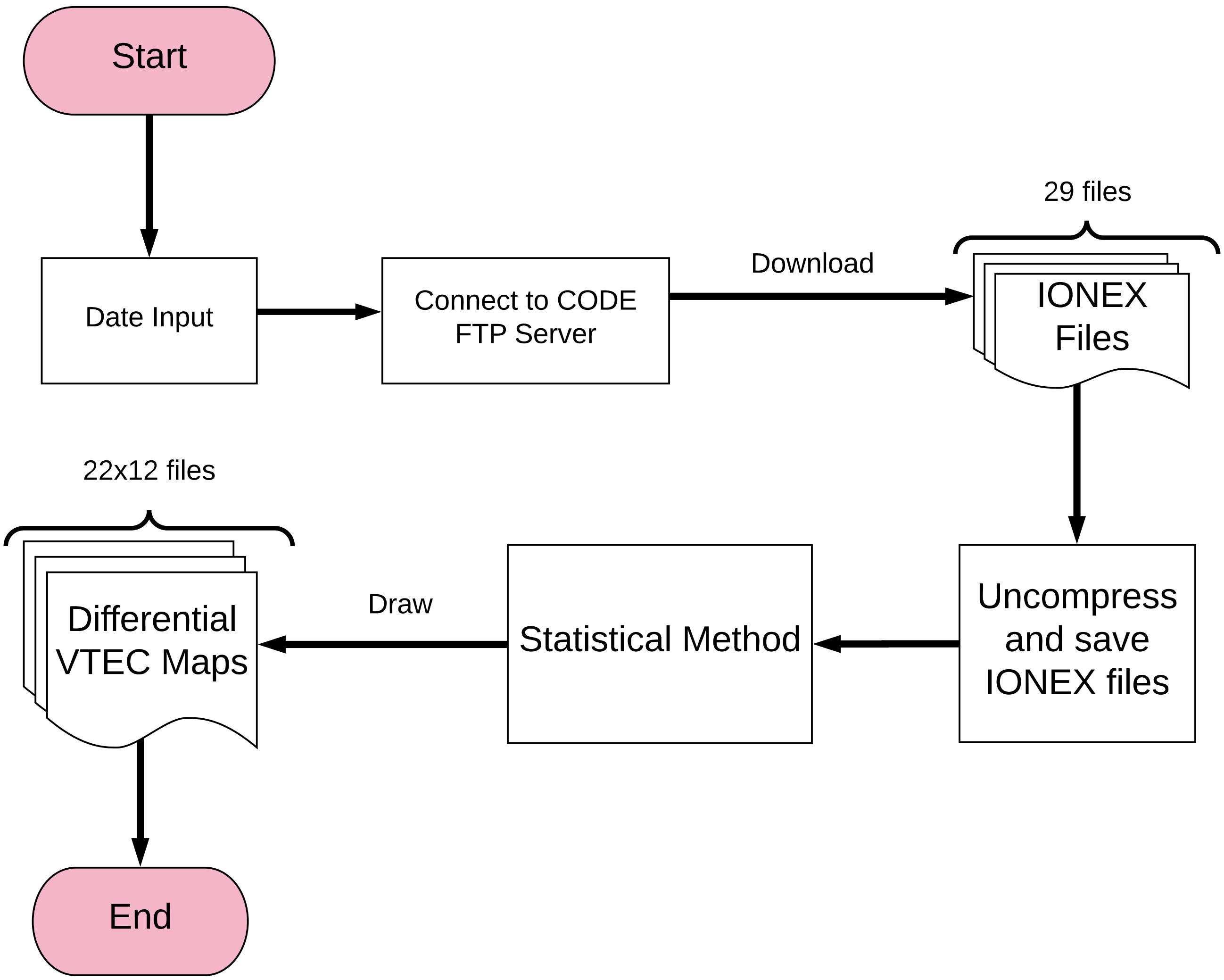


Fig. 1. Flowchart of the ICTP software.

and the associated IQRs. The statistical method to generate ΔVTEC maps consists in calculating the upper and lower bounds (UB and LB, respectively) of the normal background VTEC variation. Under the assumption that the VTEC follows a Gaussian distribution within a certain period of days for a particular hour of the day, the UB and LB can be defined as:

$$UB = \mu + 2\sigma \tag{1}$$

$$LB = \mu - 2\sigma \tag{2}$$

were μ is the mean and σ is the standard deviation. If the VTEC for a certain day at a particular hour of the day falls above the UB, we say that there is a positive anomaly. Otherwise if the VTEC falls below the LB, a negative anomaly is identified. In case that the VTEC is between the UB and LB there is no detection of a ionospheric disturbance. Given a tolerance of 2 used in equation 1 and 2 we say that there is a detection of a positive or negative ionospheric disturbance with a 95% confidence level.

III. Geomagnetic Conditions

Because ionospheric anomalies can also be produced by geomagnetic activitiy we have looked into two important geomagnetic indices (Dst and Kp). The Kp index measures the disturbance of the horizontal component of the geomagnetic field. Data for this index was obtained from the German Research Center for Geosciences (GFZ). The Dst index monitors the level of intensity of geomagnetic storms. It was retrieved Dst index data from the World Data Center for Geomagnetism in Kyoto. In Fig. 2, data for the Dst and Kp indices are shown between August 18, 2017 and September 13, 2017.

IV. Pre-Earthquake Ionospheric Anomalies

Figure 3 shows ΔVTEC maps for August 30, 2017 (9 days before the seismic event). It can be seen that at 16:00 UT on August 8, 2017 there are no ionospheric anomalies visible inside the preparation's zone to the earthquake. This preparation region is defined by radius obtained from the Dobrovolsky equation [6]:

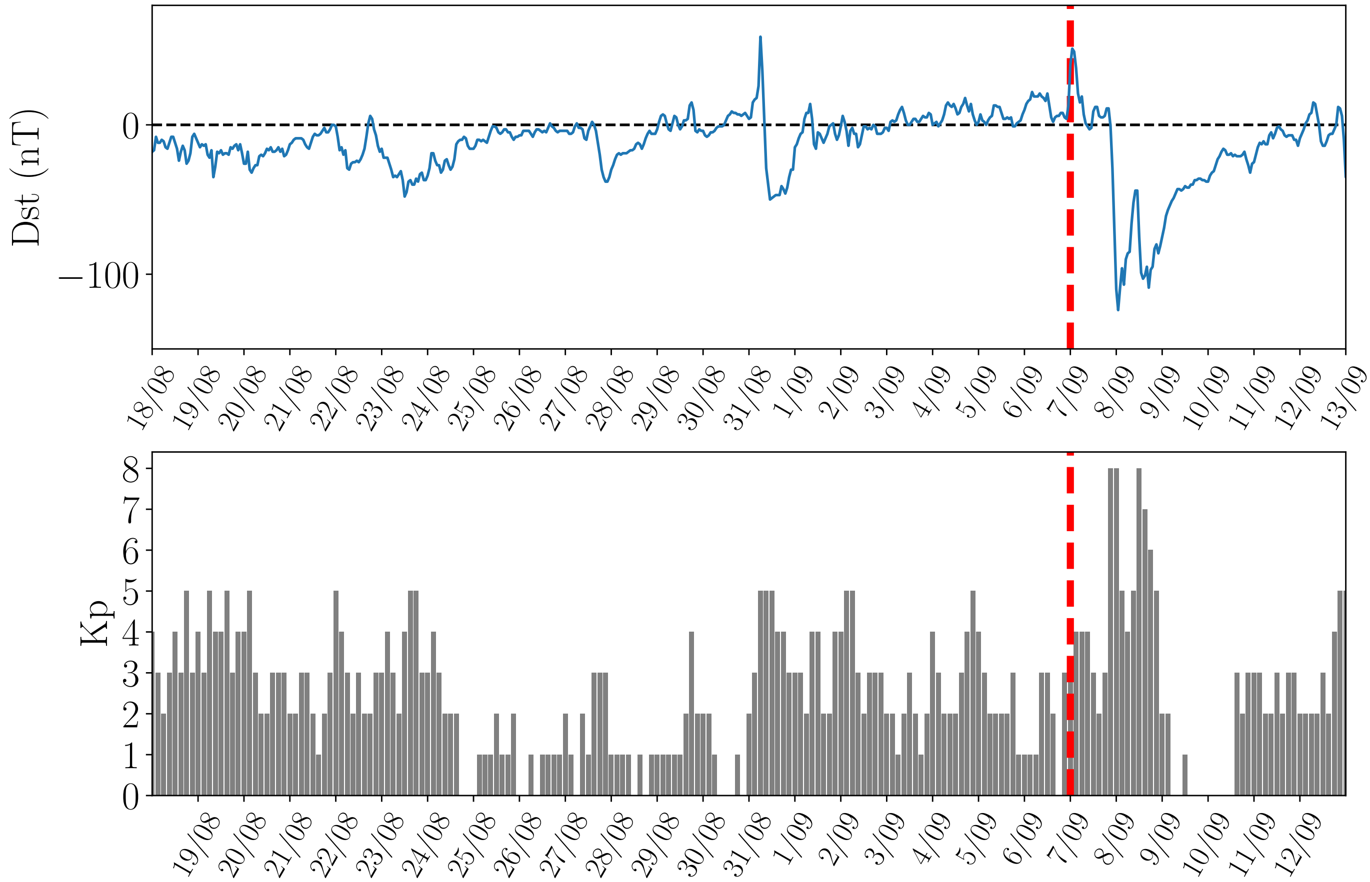


Fig. 2. Dst and Kp geomagnetic indices during August and September 2017.

$$R = 10^{0.43M}[\text{km}], \quad (3)$$

where M is the moment magnitude of an earthquake. For the M_w 8.2 earthquake, the preparation radius is ~3357 km. At 18:00 UT and 20:00 UT on September 7, 2017, a negative ionospheric anomaly can be seen to appear and displace in the same direction as the Sun. Moreover, this anomaly falls well within the area defined by the radius R. At 22:00 UT this anomaly is nearly gone. Looking into the geomagnetic indices from Fig. 2 we notice that September 7, 2017 was a day of relatively quiet geomagnetic conditions; hence, we can suggest that the origin of the anomaly is product of the earthquake's preparation.

In Fig. 4, ΔVTEC maps for September 6, 2017 (one day before the earthquake) are presented. It can be observe that a positive ionospheric anomaly appears to the north of the epicenter at 14:00 UT moving westwards until it disappears at 18:00 UT. This positive anomaly also falls inside the earthquake's preparation region at different hours. In a similar approach as the observed negative anomaly on August 30, 2017, the geomagnetic conditions during this day were stable (Kp $<$ 3 and Dst $<$ 20nT). Thus, It can be also hinted that this anomaly is also associated to the earthquake. The previous findings suggest that ionosphere anomalies were produced by the earthquake's preparation. However, according to some studies [7], [8], ionization of the air due to radiation from the Earth's crust (due to the collission in this case of the North American and Cocos' plates) can produce ionospheric anomalies which are visible in VTEC and ΔVETC maps. Depending on the level of ionization positive or negatives anomalies can be observed. Consequently, we consider that this mechanism may be at work producing the anomalies observed 9 and 1 days before the September 7, 2017 earthquake. The ICTP software gives quick results about ionosphere disturbances, however, It cannot be concluded that those anomalies came from earthquakes's preparations. Nonetheless, the software can detect anomalies around the globe base on the IONEX files downloaded. The user must discard any possible source of anomalies such as geomagnetic storms to correspond the anomaly to an earthquake event.

V. Conclusions

This paper has used a software that allows to produce differential VTEC maps using public data (IONEX files) from CODE. The ICTP software also allows to considerably reduce the time spended in gathering the data for the respective analysis. We used our software to search for anomalies before the catastrophic seismic event that happen on September 7, 2017 in Mexico. We detected 9 days before the occurrence of a negative ionospheric anomaly. On the other hand, 1 day before the earthquake we identified a positive one to the north

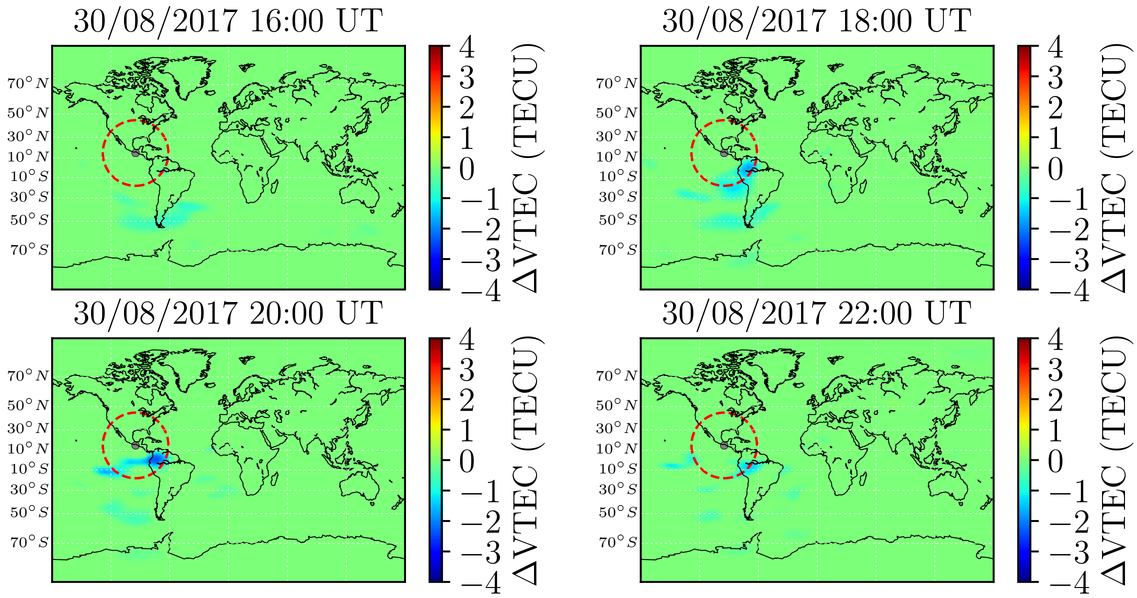


Fig. 3. Differential VTEC maps for August 30, 2017. The grey circle indicates the epicenter and the red dashed circle delineates the earthquake's preparation region.

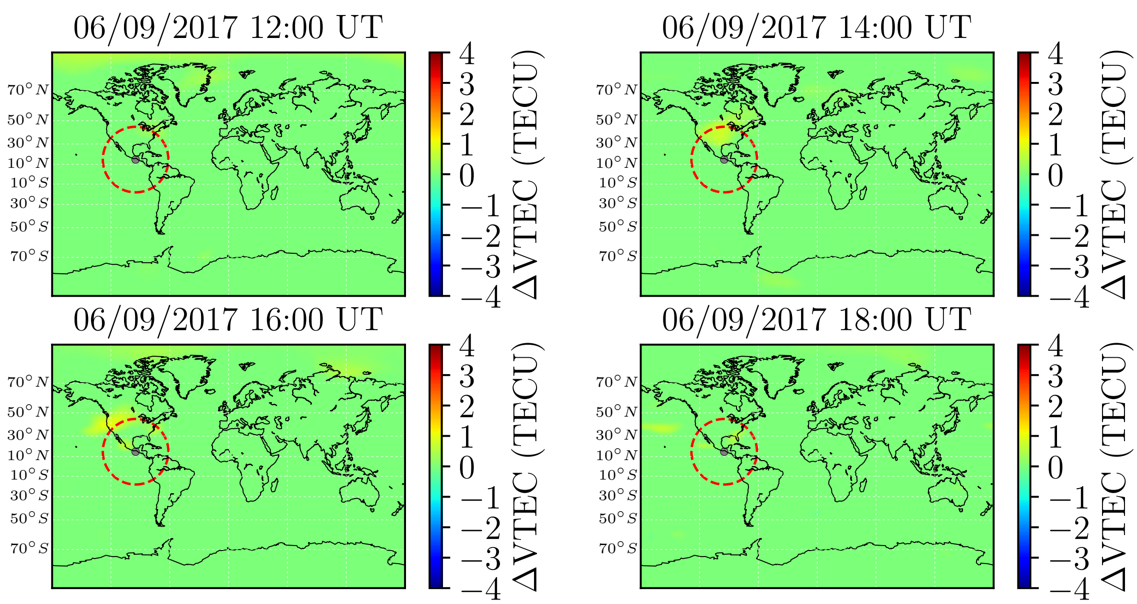


Fig. 4. Differential VTEC maps for September 6, 2017. The grey circle indicates the epicenter and the red dashed circle delineates the earthquake's preparation region.

of the epicenter. Due to the quiet geomagentic conditions during these two days, we attribute the disturbances observed to the earthquake's preparation. Moreover It is very likely that radiation from the Earth's crust is causing these ionospheric disturbances. Finally, we encourage students and researches to use this software for their own research since our main objective is to give a computational tool to test any other similar research and compare your results about ionosphere anomalies with other scientific works.

REFERENCES

[1] W. A. Mubarak, M. Abdullah, and N. Misran, "Ionospheric TEC response over Simeulue earthquake of 28 March 2005," in *2009 International Conference on Space Science and Communication*, 10 2009, pp. 175–178.

[2] J. Guo, W. Li, H. Yu, Z. Liu, C. Zhao, and Q. Kong, "Impending ionospheric anomaly preceding the Iquique M_w 8.2 earthquake in Chile on 2014 April 1," *Geophysical Journal International*, vol. 203, no. 3, pp. 1461–1470, 2015.

[3] J. Li, X. You, R. Zhang, G. Meng, H. Shi, and Y. Han, "Ionospheric total electron content disturbance associated with may 12, 2008, wenchuan earthquake," *Geodesy and Geodynamics*, vol. 70, 4 2015.

[4] Afraimovich, Edward L., Astafyeva, Elvira I., Demyanov, Vladislav V., Edemskiy, Ilya K., Gavrilyuk, Nadezhda S., Ishin, Artem B., Kosogorov, Eugene A., Leonovich, Lyudmila A., Lesyuta, Oleg S., Palamartchouk, Kirill S., Perevalova, Natalia P., Polyakova, Anna S., Smolkov, Gennadyi Y., Voeykov, Sergey V., Yasyukevich, Yury V., and Zhivetiev, Ilya V., "A review of gps/glonass studies of the ionospheric response to natural and anthropogenic processes and phenomena," *J. Space Weather Space Clim.*, vol. 3, p. A27, 2013. [Online]. Available: https://doi.org/10.1051/swsc/2013049

[5] Z. G. W. and S. C., "Processing tool for generating global ionosphere maps (gim)," in *Proc. 2018 IEEE COngreso Argentino de Ciencias de la Informática y Desarrolls de Investigación (CACIDI)*, in press.

[6] I. P. Dobrovolsky, S. I. Zubkov, and V. I. Miachkin, "Estimation of the size of earthquake preparation zones," *pure and applied geophysics*, vol. 117, no. 5, pp. 1025–1044, Sep 1979. [Online]. Available: https://doi.org/10.1007/BF00876083

[7] Pulinets, "Physical mechanism of the vertical electric field generation over active tectonic faults," *Advances in Space Research*, vol. 44, pp. 767–773, Sep. 2009.

[8] S. A. Pulinets, "Low-Latitude Atmosphere-Ionosphere Effects Initiated by Strong Earthquakes Preparation Process," *International Journal of Geophysics*, vol. 2012, 2012.

Hybrid Non-Reference QoE Prediction Model for 3D Video Streaming Over Wireless Networks

Ibrahim Alsukayti
College of Computer
Qassim University
Saudi Arabia

Mohammed Alreshoodi
Unizah Communitiy College
Qassim University
Saudi Arabia

***Abstract*—With the rapid growth in mobile device users, and increasing demand for video applications, the traffic from 2D/3D video services is expected to account the largest proportion of internet traffics. User's perceived quality of experience (QoE) and quality of service (QoS) are the most important key factors for the success of video delivery. In this regard, predicting the QoE attracts high importance for provisioning of 3D video services in wireless domain due to limited resources and bandwidth constraints. This study presents a cross-layer no-reference quality prediction model for the wireless 3D video streaming. The model is based on fuzzy inference systems (FIS), and exploits several QoS key factors that are mapped to the QoE. The performance of the model was validated with unseen datasets and even shows a high prediction accuracy. The result shows a high correlation between the objectivley measured QoE and the predicted QoE by the FIS model.**

***Keywords*—QoE; QoS; Video; Fuzzy Logic; Prediction**

I. Introduction

The success of the modern communication networks and systems has made many applications and services widely accessible and more usable. Among of these are multimedia applications which become popular over the Internet. One critical multimedia service is video streaming that is nowadays not only delivered as 2D stream but also in 3D. It is expected that 82% of the global Internet traffic will be video, and more than 65% of the video traffic will be transferred over wireless communication links and using mobile devices [1]. It is evident that maintaining video quality, particularly in the scarce wireless and mobile environments, is of critical importance. While this can be tackled focusing on the performance of the network, it is also important to highly consider it from the end user perspective. Focusing on consumers QoE rather than the networks QoS leads to evaluating the quality of videos as perceived by end-users. Compared to 2D videos, a 3D video introduces a new dimension, namely the depth, which needs to be additionally taken into account when processing users QoE.

Maintaining high QoE for the end users requires the ability for 3D videos to be efficiently monitored, predicted, and controlled [2]. Nevertheless, the prediction of QoE is highly based on understanding the impact of QoS parameters [3], [4]. In this regards, the relationship between QoE and QoS could be established in a defined prediction model. This can be done based on no-reference QoE prediction without the need for a reference 3D video. In order to achieve such capability, it is important to develop QoE prediction models that consider more QoS factors related to users QoE. One of the effective techniques for developing objective QoE prediction models is the learning-based technique with the different types of machine learning methods [5]. Developing QoE prediction models using the machine learning methods would result in a model that can dynamically and intelligently learn and then make a decision like human reasoning.

There is a number of machine learning techniques that have been adopted to realize QoE prediction models [6]. Among the popular examples are Fuzzy Inference Systems (FIS), Artificial Neural Networks (ANN), and Decision Tree. However, it is critical to comprehensively incorporate the key QoS factors for video traffic, instead of relying on a basic model of limited factors which is the case in the majority of the proposed solutions. According to the ITU classification of objective quality assessment models [7], a hybrid QoE prediction model is the best choice to build a generic prediction model, which is proposed in this paper.

The FIS method has been the choice for many solutions in telecommunications and engineering. It provides an efficient system for addressing the innate uncertainty, which can be caused by internal or external factors. Developing a QoE prediction model based on FIS with predefined rules is an effective approach to make effective decision with imprecise information. Typically, fuzzy rules in FIS cannot be automatically formulated and need to be manually updated when the input dataset is updated. We proposed in this paper a hybrid no-reference prediction model using an automated FIS for predicting the quality of wireless 3D video streaming. The main target of the proposed solution is to support real-time QoE prediction at an intermediate measuring point over a wireless network for 3D video streaming.

The structure of the paper is as follows; the related works is discussed in Section 2. Section 3 provides a description of the experimental setup. The validation and statistical analysis of the QoE measurements resulted from the simulations are presented in Section 4. Section 5 discusses the methodology of the proposed video quality prediction system. The QoE prediction model performance validation and evaluation are provided in Section 6, respectively. Finally, the conclusion is in section 7.

II. Related Works

Compared to other multimedia services, video streaming is a demanding service that is more sensitive to any degradation over the network. The is more evident in the case of 3D video communications over wireless networks with mobile end-user devices. Without careful and efficient system design, degradation in user experience would occur. Therefore, it is critical to consider the different application and networking aspects in this regard. There is a variety of video quality measurement methods in the literature with different operational and computational requirements [3]. The measurement of video quality is typically achieved with one of the two modes: non-intrusive no-reference (NR) and intrusive full-reference (FR). The FR mode has to use both the original and distorted video streams. In contrast, the NR mode is only based on the distorted video streams, which is more reliable for real-time applications.

The effects of QoS can happen at the access network layer, referred to as NQoS, or at the application layer, referred to as AQoS. The NQoS can be described with a number of NQoS parameters including delay, packet loss, and random loss. These have been considered in many research efforts such as [8], [9], and [10] for estimating QoE in 2D video streaming applications. On the other hand, AQoS parameters such as frame rate bit rate were the focus in [11], [12], [13], [14]. However, QoE estimation for 3D video streaming received less attention in the literature. This would require the consideration of additional parameters relevant to 3D videos including depth perception, naturalness, and comfort levels. The study in [15] focused on examining random packet losses and its effect on the overall 3D perception. With a subjective test, the results showed that the increase in packet loss rate led to a negative trend in 3D perception. In [16] and [17], the researchers considered the degradation on the 3D video quality when delivered over wireless mobile networks. The aim was to understand how the overall 3D perception can be affected by random packet loss using a subjective test. It is apparent that there was no consideration in these studies for evaluating the quality of 3D videos with VBR streams in different resolutions. Moreover, there is no need to completely rely on both NQoS and AQoS parameters when adopting hybrid prediction models.

Recently, there is a growing interest in the literature in the utilization of machine learning techniques for the development of objective QoE prediction models [5], [18], [19]. As the self-learning capability can be provided by such methods, non-intrusive prediction of video quality can be implemented to dynamically adapt to any update in QoS parameters. One of these methods is the Adaptive Neural Fuzzy Inference System (ANFIS) [20] which was adopted in [21] and [22] for the estimation of the quality score, with the focus on 2D video in a single resolution, QCIF (176x144). The authors in [23] presented a real-time prediction engine for 3D video using Random Neural Network (RNN). Another research work focused on 3D video is [24], which presented a RR metric based on Peak Signal-to-Noise ratio (SNR).

It is observed from the investigated literature that most of the video quality prediction models considered either network impairments, encoders compression artifacts, or the features of video content. It is uncommon that these factors are considered all together in one solution. The proposed work in this paper introduces a a non-intrusive QoE prediction model based on the use of an automated FIS method considering a collection of key QoS parameters for wireless 3D video streaming.

III. Experimental Set-Up

A. Video Encoding Parameters

Three classes of H.264 coded 3D video streams were evaluated based on a temporal activity by using the spatio-temporal classification in the ITU-T P.910 recommendation [25]. The temporal and spatial features were extracted from the 3D video stream, and then a temporal index (TI) and spatial index (SI) were assigned by the Sobel filter. The temporal activity and spatial complexity of the video sequence are indicated by the computed index. Consequently, 6 3D video sequences (2 in each class) were chosen, as shown in Table I. The sequences Music, Poker and BMX were used for training, while Fencing, Poznan and Pantomime were used for testing and validation.

TABLE I. Video Sequences

Video Sequence	TI	SI	Class
Music	4.90	74.41	Low Motion
Fencing	7.78	77.20	
Poker	12.20	85.69	Moderate Motion
Poznan	11.53	87.78	
BMX	22.35	99.42	High Motion
Pantomime	37.17	104.43	

The H.264/AVC JM Reference Software [26] was used to encode and decode all video sequences with the H.264/AVC video coding standard [27] (for both the colour image and the depth map). Table II presents the configuration parameters of the encoding process. The frame rate (FR) was fixed at 25 fps as this is typical in wireless video streaming [28]. The videos sequences were encoded with different quantization parameter (QP) and resolutions. The network abstraction layer (NAL) units were RTP packetized and encapsulated in IP packets. The Group of Picture (GOP) size was 16, where each group included one I-frame and all remaining frames were P-frames. This reduced the computation time arising from bi-predictive B-frames, and is the structure recommended for wireless video streaming [28].

TABLE II. Parameters H.264/AVC Coding

Parameters	Values
Level IDC	30 (SD), 32 (HD)
Sequence GoP	IPPP
Reference Frames' Number	2
Search-Range	32
Slice-Mode	Packetized (bytes)
Format of Output-File	RTP packet

B. The selected QoS Parameters

The selected QoS parameters in this study were resolution (R), quantization parameter (QP) and content type (CT) from the AQoS level, while mean burst loss (MBL) and packet loss rate (PLR) from the NQoS level. Table III summarises the values of the simulated QoS parameters. Moreover, the simulation of each tested condition is repeated 10 different times to increase data confidence.

C. Simulation Scene

The simulation scene illustrated in Fig 1, which is designed and conducted for mapping the QoS to QoE. The coded

3D video streams (2D colour image and depth map) were simulated on wireless transmission environment. The packet loss traces with varying MBL and PLR metrics were generated by using the Gilbert-Elliot model [29]. The degraded 3D video streams were then assessed by an objective 3D video quality metric, which is explained in the following subsections.

TABLE III. QOS PARAMETERS

Parameters	Values
R	SD (720 × 540), HD (1280 × 720)
CT	Low, Moderate, High motion
QP	16, 24, 32, 40, 48
MBL	1, 2.5, 5, 7.5
PLR	0%, 0.1%, 1%, 2.5%, 5%, 7.5%, 10%

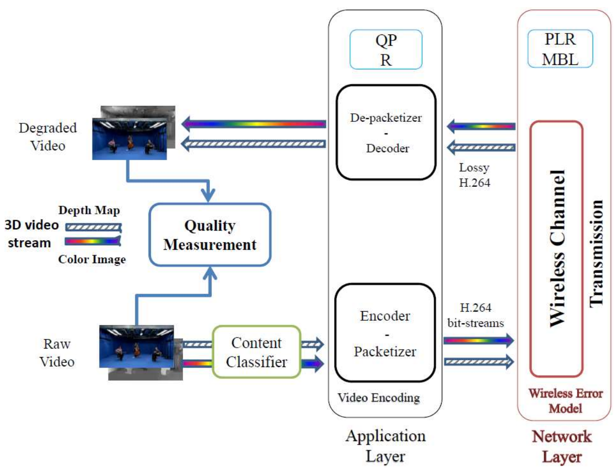


Fig. 1. The Conducted Simulation Scene

IV. QOE ASSESSMENT AND EVALUATION

The QoE measurement methods can be categorised into subjective and objective methods [30]. In this paper, an objective quality metric was used for quality assessment and then validated by a subjective assessment.

A. Objective Test

A validated full reference perceptual 3D video quality metric (Q) [31] was used for the objective measurements. This metric uses VQM (Video Quality Metric) [32] for the 2D colour images' assessment, and then uses a joint mathematical model [31] to be combined with the corresponding depth map. For the VQM scale, 1 represents severe distortion and 0 represents original quality. The 3D quality scale is mapped to the subjective metric called MOS (Mean Opinion Score) [25] by means of the equations [33]:

$$MOS = 5 - 4VQM \tag{1}$$

$$MOS = 5 - 4(1 - Q) \tag{2}$$

B. Subjective Test

In this work, the subjective assessment test was conducted to assure the credibility of the huge measured objective dataset. The standard recommendation ITU-R BT.500-13 [34] was followed for this test. Because the total number of test conditions for the measured objective dataset was huge (about 1080 conditions), a systematic approach (called Kennard and Stone algorithm) [35] was followed to select a subset of 64 video sequences for subjective testing. A single stimulus (SS) quality evaluation method was applied with a panel of 21 viewers in a lab under controlled conditions. The viewers were marked their MOS scores between 1 to 5, as illustrated in Table IV. Figure 2 illustrates the observers' MOS scores and their corresponding 95% confidence intervals.

TABLE IV. MOS SCORES

Quality	Bad	Poor	Fair	Good	Excellent
MOS	1	2	3	4	5

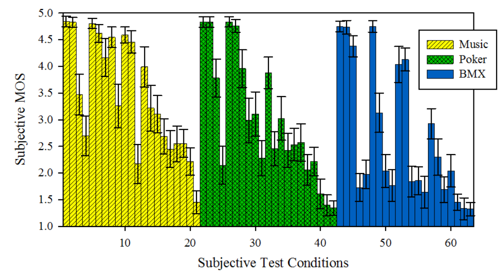


Fig. 2. Subjective test with corresponding 95% confidence intervals

C. Correlation of Subjective and Objective QoE Scores

The objective scores that obtained through simulation were validated by correlating them with the subjective scores, as shown in Fig 3. In addition, the PCC (Pearson Correlation Coefficient) was used to express the correlation. The measured PCC indicated a high correlation level 92%.

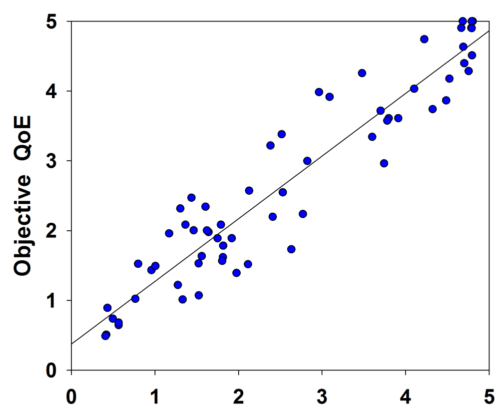


Fig. 3. Objective and Subjective Correlation

The analysis of variance (ANOVA) test [36] was also conducted in this study to explore the impact of AQoS and NQoS parameters on the 3D video quality. Tables V shows the results of the ANOVA test. A small p value $(p <= 0.01)$ indicates that the QoS parameter highly affectes the video quality. It is clear from the table that the PLR metric had the highest impact. The interactions between these QoS parameters also affect the video quality

V. QOE PREDICTION METHODOLOGY

The peoposed 3D video quality prediction was built by using the automated Type-1 FIS, which outperforms other

TABLE V. FIVE-WAY ANOVA ON QOE OF 3D VIDEO

Source	Degree of freedom	F-statistics	p-value
CT	2	141.7391	0
R	2	124.416	0.02975
QP	4	165.981	0
PLR	5	639.172	0
MBL	3	73.354	0.01003
PLR+CT	10	10.218	0.11909
PLR+R	8	5.182	0.3342
PLR+QP	18	71.955	0.1264
PLR+MBL	15	30.466	0.2921
MBL+CT	5	2.868	0.1942
MBL+R	5	7.940	0.52301
MBL+QP	12	13.533	0.29968

methods in terms of making decisions and modelling capabilities [37]. The FIS control consists of three main modules; fuzzifier, fuzzy inference engine and defuzzifier. A functional block of the proposed model is presented in Figure 4.

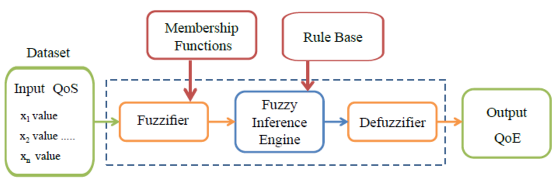


Fig. 4. Functional Diagram of The Proposed Model

Inputs: QoS parameters $x1$, $x2$, ,,,, x_n, and their values.

Membership functions: $\mu^{(P_i)}A_j(x)$, $i \in (1, \ldots, m)$, $j \in (1, \ldots, n)$,

where A_j the class number, n, of corresponding parameter values, and x represents the input parameter value (Pm_{value}).

The form of the fuzzy rules: $IF < antecedent_1 > AND < antecedent_2 > THEN < consequent >$;

Output: QoE (MOS) scores.

Generally, the procedure of designing a FIS model can be divided into two main steps; learning (initialization) step and control step. The learning step includes defining the linguistic variables, designing the MFs and extracting the fuzzy rules base. While, the control step includes fuzzification, inference engine and defuzzication. Algorithm 1 gives an overview about the process of fuzzy logic system [38]. The following subsections describe the procedure of designing the proposed model, its MFs and fuzzy rules extraction.

Algorithm 1: process of the FIS

1. Define the linguistic expressions (Initialisation)
2. Design the membership function using triangle shape (Initialisation)
3. Convert crisp input value to fuzzy value using the MFs (Fuzzification)
4. Automatically extract the fuzzy rule base (Fuzzy inference engine)
5. Evaluate the fuzzy rules in the rule base (Fuzzy inference engine)
6. Aggregate the results of each rule (Fuzzy inference engine)
7. Convert the fuzzy value to crisp output value (Defuzzification)

A. Identifying the Inputs and Output

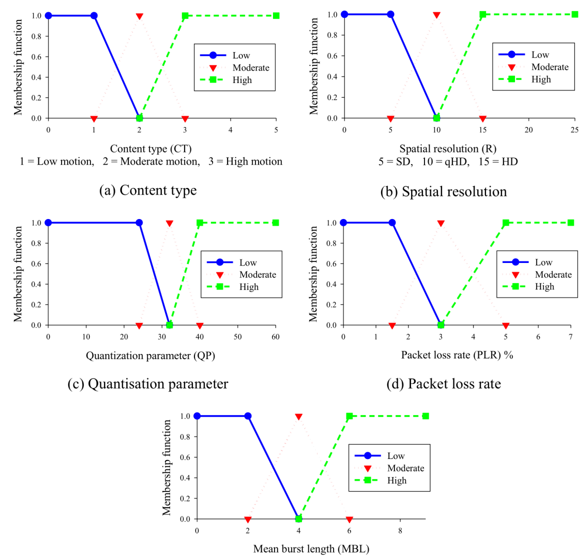
the chosen QoS parameters are outlined in table III and the output MOS scores (QoE) in table IV. The collected dataset consists of multiple data pairs of the input and output using the following form:

$$(x^{(t)}; y^{(t)})(t = 1, 2, ..., N), \tag{3}$$

where N is the data instances number, $x^{(t)} \in R^n$, and $y^{(t)}$ R^k. Once the inputs and the output were identified, both were converted into linguistic expressions (the MFs design) to represent the quantification of the output (QoE scores).

B. Membership Functions Design

The correlation between the input and output parameters was transferred into fuzzy MF. In this study, the MFs were derived using probability distribution functions (PDF) [39] for every QoS parameter. The PDF is divided by its peak to convert the probabilistic information into fuzzy sets, which are a set of rules that take linguistic expressions form. In this work, the QoS input parameters were assigned by three fuzzy sets (low, moderate, high), while the output was assigned by five fuzzy sets, which are the MOS scales. All the fuzzy sets were presented in an equivalent triangular shape due to its simplicity. Figure 5 shows the MFs of the inputs parameters, while the MF of the output is presented on Figure 6.



(a) Content type (b) Spatial resolution (c) Quantisation parameter (d) Packet loss rate (e) Mean burst length

Fig. 5. MF of the Input Parameters

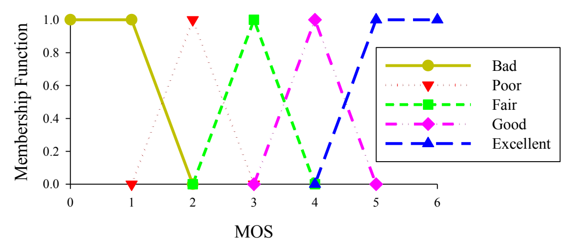


Fig. 6. MF of the Output Parameters

C. Automated Fuzzy Rules Extraction

The proposed model automatically adapts the rules necessary for the fuzzy inference system using learning from example (LFE) approach [40]. A developed version of the (Mendal-Wang) method [41], [38] was used to apply the LFE approach in designing the proposed automated FIS model. The extracted rules can take different forms. In the proposed system, the following fuzzy IF-THEN rules was used to represent the relationship between the input pattern $x = (x_1, ..., x_n)^T$ and the output $y = (y_1, ..., y_n)^T$:

$$IF\ x_1\ is\ A_1^{(i)}\ and\ ...\ x_n\ is\ A_n^{(i)}\ THEN\ y_k\ is\ B_k^{(i)} \quad (4)$$

where the rules index $i = (1, 2, ...M)$, and M represent the rules number. A number of V fuzzy sets A_s^q, for $q \in \{1, 2, \ldots, V\}$, defined for each input X_s, and W fuzzy sets B^h, for $h \in \{1, 2, \ldots, W\}$, defined for each output y_c. Moreover, the AND operator, which selects the minimum value of the fuzzy sets was used in this model. The process of extracting the fuzzy rules from the data was performed using the following two steps, as described in [41], [38]:

STEP A:
For a fixed N pair $(x^{(t)}; y^{(t)})$ of input-output parameters, the membership values $\mu_{A_s^q}(x_s^{(t)})$ are computed for each MF ($q \in \{1, \ldots, V\}$), and each input variable $s \in \{1, \ldots, n\}$ to give $q^* \in \{1, \ldots, V\}$:

$$\mu_{A_s^{q^*}}\ (x_s^{(t)})\ \geq\ \mu_{A_s^q}\ (x_s^{(t)}) \quad (5)$$

Thus, this step defines the process of choosing the fuzzy set that reaches the highest membership degree at the data point as the one in the *IF part* of the rule.

$$IF\ x_1^t\ is\ A_1^{q^*}\ and\ x_2^t\ is\ A_2^{q^*}\ and\ .....\ x_n^t\ is\ A_n^{q^*}\ THEN\ y\ is\ centred\ at\ y^{(t)} \quad (6)$$

Note that each of the fuzzy sets X_s associated to the input variables is characterised by V fuzzy sets A_s^q, where $q \in \{1, \ldots, V\}$, and V^n is the maximum number of possible generated rules. Nevertheless, depends on the dataset, only those rules were generated from the V^n possibilities whose dominant regions contain at least one data point. In this STEP, one rule was generated for each pair of the input and output data. This rule is modified to create its final form in step 2. The rule's weight was computed using the formula [41]:

$$W^{(t)}\ =\ \prod_{s=1}^{n}\ \mu_{A_s^q}(x_s^{(t)}) \quad (7)$$

STEP B:
The STEP (A) was conducted several time in order to create the N data generated rules by Equation (7) for all data points t, from 1 to N. Since the number of data points is usually large, many generated rules in the first step were shared the same IF part but dissimilar the THEN part. In this step, the N rules divided into groups with the same IF part. For example, suppose there are M such groups i ($i \in \{1, \ldots, M\}$) that includes N_i rules of the form:

$$IF\ x_1\ is\ A_1^{(q^i)}\ and\ x_2\ is\ A_2^{(q^i)}\ and\ .....\ x_n\ is\ A_n^{(q^i)}\ THEN\ y\ is\ centred\ at\ y^{(t_u^i)} \quad (8)$$

Where $u \in \{1, \ldots, N_i\}$ and (t_u^i) is the data point index in the group (i). After that the rule weighted average in this conflict group is calculated by the following equation [41]:

$$average^{(i)} = \frac{\sum_{u=1}^{N_i} y^{(t_u^i)} w^{(t_u^i)}}{\sum_{u=1}^{N_i} w^{(t_u^i)}}, \quad (9)$$

Where the weights $w^{(t_u^i)}$ are computed using equation (9) from Step (A). These N_i rules are then combined into a single rule as the following form:

$$IF\ x_1\ is\ A_1^{(i)}\ and\ ....\ x_n\ is\ A_n^{(i)}\ THEN\ y\ is\ B^{(i)} \quad (10)$$

The fuzzy set of the output B^i was selected on the basis of finding the set B^{h^*} among the W output fuzzy sets B^1, ..., B^w that satisfies:

$$\mu_{B^{h*}}\ (average^{(i)})\ \geq\ \mu_{B^h}\ (average^{(i)}) \quad (11)$$

Where $h \in \{1, 2, ..., W\}$, and $B^{(i)}$ was chosen to be B^{h*}.

The calculations invoke Equations (9) and (11), and are repeated for each output value [41], [38]. Table VI shows a sample of the index fuzzy rules. The final fuzzy rule that was used in the FIS controller had the form:

IF (CT is Moderate motion) AND (QP is Moderate) AND (Resolution is Moderate) AND (PLR is High) AND (MBL is High) THEN (QoE is Bad)

IF (CT is Low motion) AND (QP is Moderate) AND (Resolution is Low) AND (PLR is Low) AND (MBL is Moderate) THEN (QoE is Fair)

TABLE VI. EXAMPLES OF THE QOE INDEX DECISION MAKING RULES, (H= HIGH, M= MODERATE, L= LOW)

CT	R	QP	PLR	MBL	QoE
L	M	L	L	L	Excellent
M	M	M	L	L	Good
H	L	M	L	M	Fair
L	M	H	M	M	Poor
M	M	L	L	M	Good
H	M	H	H	H	Bad
L	L	M	H	M	Poor

D. Defuzzification: Predicting the Output

The overall result after the inference step was a set of fuzzy values. This result was defuzzified to give a crisp output value (QoE), based on the MF of the output variable. The defuzzification process examined all of the rule outcomes after they were logically added and then computed a value that was the final output of the fuzzy controller. In this work, the defuzzification was conducted using the centroid method that is based on the centre of gravity (COG) formula [38]:

$$y(x) = f_s(x) = \frac{\sum_{i=1}^{M} y^{-i} \prod_{l=1}^{n} \mu F_l^i(x_l)}{\sum_{i=1}^{M} \prod_{l=1}^{n} \mu F_l^i(x_l)}. \quad (12)$$

Here, y^{-i} is the centroid point of the output fuzzy set B^i, M is the rule's number in the rule base, and $\Pi_{l=1}^{n} = \mu F_l^i(x_l)$ is the membership degree product of each rule's inputs.

E. Implementation of The Automated FIS model

In this work, the designed membership functions, fuzzy rule extraction and fuzzy controller for the automated FIS model were built in a Java programming environment. The main components of FIS controller were programmed as Java classes. The main java class (Fuzzy_inference_engine) includes the equations of Mendal-Wang method.

VI. Model Validation

A. Validation by The Testing Dataset

As listed in Table I, 6 3D video sequences were chosen, two in each class. The sequences Music, Poker and BMX are used for model training and Fencing, Poznan and Pantomime are used for model testing. The testing dataset was used to validate the proposed prediction model. The measured QoE results are compared with the predicted QoE by the proposed prediction model. The used validation metrics were R^2 correlation and RMSE (root mean squared error). R^2 scored 0.951 and RMSE was 0.1058. The validation of the proposed system is illustrated in Fig 7 and 8. In Fig 7, the measured MOS (QoE) represented by the line, while each point shows the estimated MOS (QoE) of a particular test condition. The obtained result indicates that the measured QoE is greatly correlated with predicted QoE. So, the proposed FIS-based model succeeds in estimating the users perception, and shows how the relationship between the AQoS/NQoS parameters and the video QoE is consistent.

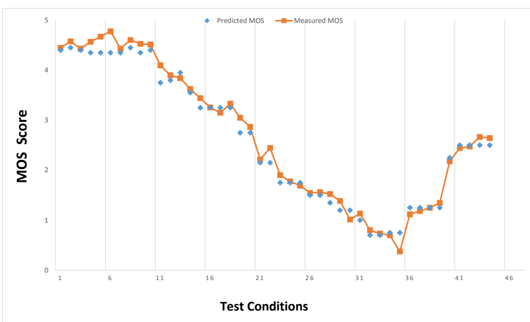


Fig. 7. Predicted MOS vs. Measured MOS

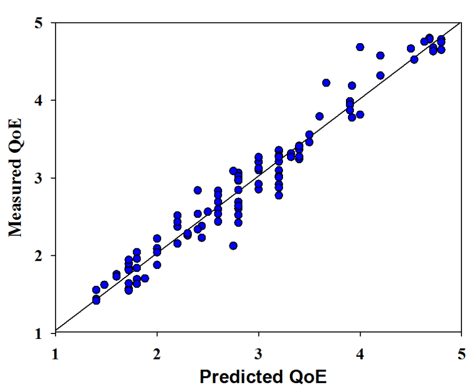


Fig. 8. Predicted MOS vs. Measured MOS (scatter graph)

B. Validation by The External Dataset

A further validation was conducted by using an external dataset that has been used in [21] and publicly available in their website. This dataset is for H.264 encoded QCIF 2D videos with network conditions of PLR, MBL for three types of video sequences (CT). In this case, we decrease the input parameters in the proposed model to three inputs; PLR, MBL and CT. Fig 9 illustrates the model validation results after using the external database. This external dataset was also used in [21] for the ANFIS-based prediction model. Thus, we compared the results of our FIS-based model with the ANFIS-based model developed in [21]. As shown in table VII, the proposed model achieved a correlation coefficient of 90.2%, while 87.1% for the ANFIS-based model in [21].

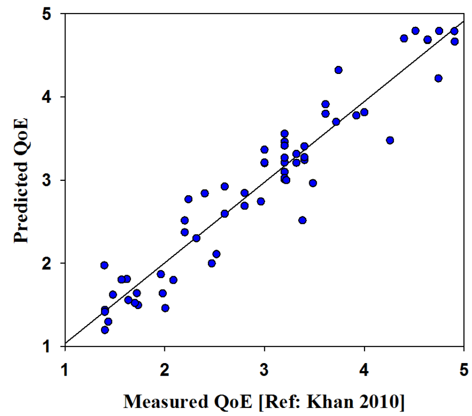


Fig. 9. Predicted MOS vs. Measured MOS from External Dataset [21]

TABLE VII. Performance Comparison

Model Name	R^2	RMSE
Automated FIS-Based	90.23%	0.206
ANFIS-Based	87.19%	0.2412

VII. Conclusions

In this proposed work, a non-reference prediction model was developed to predict the 3D video quality in wireless transmission environment. The proposed model was built by using an automated FIS method. Moreover, a selection of AQoS and NQoS parameters were identified and mapped to the MOS scores of the transmitted 3D video streams for end-to-end quality prediction. A subjective assessment was conducted to validate the objectively measured QoE dataset. Furthermore, in order to identify the most influential QoS parameters that affect the video QoE, the collected dataset was also analysed by the ANOVA test. After that the validated dataset was then used to build the proposed model. From the results, the proposed FIS-based model shows a high correlation between the objectively measured QoE and the predicted QoE. The results also confirmed that the choice of the AQoS/NQoS parameters is essential to achieve a high prediction accuracy.

This work advances the development of non-reference quality prediction models for wireless 3D video streaming. For future work, the usage of the FIS-based model can be investigated to implement a potential application, such as content provisioning for network/service providers. Furthermore, additional AQoS and NQoS parameters can be considered for end-to-end quality prediction.

References

[1] I. Cisco, “Cisco visual networking index: Forecast and methodology, 2016?2021,” *CISCO White paper*, 2017.

[2] A. Khan, L. Sun, and E. Ifeachor, “Qoe prediction model and its application in video quality adaptation over umts networks,” *Multimedia, IEEE Transactions on*, vol. 14, no. 2, pp. 431–442, 2012.

[3] T. H. Lea Skorin-Kapov, Martín Varela and K.-T. Chen, "A survey of emerging concepts and challenges for qoe management of multimedia services," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2018.

[4] M. Yang, S. Wang, R. Calheiros, and F. Yang, "Survey on qoe assessment approach for network service," *IEEE Access*, vol. 6, 2018.

[5] M. S. Mushtaq, B. Augustin, and A. Mellouk, "Empirical study based on machine learning approach to assess the qos/qoe correlation," in *Networks and Optical Communications (NOC), 2012 17th European Conference on*. IEEE, 2012, pp. 1–7.

[6] S. Aroussi and A. Mellouk, "Survey on machine learning-based qoe-qos correlation models," in *Computing, Management and Telecommunications (ComManTel), 2014 International Conference on*, April 2014.

[7] A. Takahashi, D. Hands, and V. Barriac, "Standardization activities in the itu for a qoe assessment of iptv," *Communications Magazine, IEEE*, vol. 46, no. 2, pp. 78–84, February 2008.

[8] P. Pradip, B. Federica, and C. Marco, "Impact of video content and transmission impairments on quality of experience"," *Multimedia Tools and Applications*, vol. 75, 2016.

[9] J. Pokhrel, B. Wehbi, A. Morais, A. Cavalli, and E. Allilaire, "Estimation of qoe of video traffic using a fuzzy expert system," in *Consumer Communications and Networking Conference (CCNC), 2013 IEEE*. IEEE, 2013, pp. 224–229.

[10] J. Nightingale, Q. Wang, C. Grecos, and S. Goma, "The impact of network impairment on quality of experience (qoe) in h. 265/hevc video streaming," *Consumer Electronics, IEEE Transactions on*, vol. 60, 2014.

[11] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 3, pp. 286–298, 2011.

[12] A. Eden, "No-reference estimation of the coding psnr for h. 264-coded sequences," *Consumer Electronics, IEEE Transactions on*, vol. 53, 2007.

[13] X. Lin, H. Ma, L. Luo, and Y. Chen, "No-reference video quality assessment in the compressed domain," *Consumer Electronics, IEEE Transactions on*, vol. 58, no. 2, pp. 505–512, 2012.

[14] S.-O. Lee, K.-S. Jung, and D.-G. Sim, "Real-time objective quality assessment based on coding parameters extracted from h. 264/avc bitstream," *Consumer Electronics, IEEE Transactions on*, vol. 56, no. 2, pp. 1071–1078, 2010.

[15] C. T. Hewage, M. G. Martini, M. Brandas, and D. De Silva, "A study on the perceived quality of 3d video subject to packet losses," in *Communications Workshops (ICC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 662–666.

[16] M. M. Nasralla, C. T. Hewage, and M. G. Martini, "Subjective and objective evaluation and packet loss modeling for 3d video transmission over lte networks," in *Telecommunications and Multimedia (TEMU), 2014 International Conference on*. IEEE, 2014, pp. 254–259.

[17] M. M. Nasralla, O. Ognenoski, and M. G. Martini, "Bandwidth scalability and efficient 2d and 3d video transmission over lte networks," in *Communications Workshops (ICC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 617–621.

[18] E. Aguiar, A. Riker, M. Mu, S. Zeadally, E. Cerqueira, and A. Abelem, "Real-time qoe prediction for multimedia applications in wireless mesh networks," in *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*. IEEE, 2012, pp. 592–596.

[19] P. Indira, P. Jeevan, W. Bachar, C. A. Rosa, and J. Badii, "Estimation of video QoE from MAC parameters in wireless network : a Random Neural Network approach," pp. 51 – 55, 2014.

[20] J.-S. R. Jang, "Anfis: adaptive-network-based fuzzy inference system," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 23, no. 3, pp. 665–685, 1993.

[21] A. Khan, L. Sun, and E. Ifeachor, "An anfis-based hybrid video quality prediction model for video streaming over wireless networks," in *The Second International Conference on Next Generation Mobile Applications, Services and Technologies*. IEEE, 2008.

[22] C. N. Pitas, D. E. Charilas, A. D. Panagopoulos, and P. Constantinou, "Adaptive neuro-fuzzy inference models for speech and video quality prediction in real-world mobile communication networks," *Wireless Communications, IEEE*, vol. 20, no. 3, pp. 80–88, 2013.

[23] H. Malekmohamadi, W. Fernando, E. Danish, and A. Kondoz, "Subjective quality estimation based on neural networks for stereoscopic videos," in *Consumer Electronics (ICCE), 2014 IEEE International Conference on*. IEEE, 2014, pp. 107–108.

[24] C. T. Hewage and M. G. Martini, "Reduced-reference quality assessment for 3d video compression and transmission," *Consumer Electronics, IEEE Transactions on*, vol. 57, no. 3, pp. 1185–1193, 2011.

[25] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 2008.

[26] J. V. T. J. of ISO/IEC MPEG & ITU-T VCEG. (2011) H.264/avc jm reference software. [Online]. Available: http://iphome.hhi.de/suehring/tml/

[27] ITU-T, "advanced video coding for generic audiovisual services, recommendation itu-t h.264," 2013.

[28] R. A. Farrugia, *Multimedia Networking and Coding*. IGI Global, 2012.

[29] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell system technical journal*, vol. 42, no. 5, 1963.

[30] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *Image Processing, IEEE transactions on*, vol. 19, no. 6, pp. 1427–1441, 2010.

[31] S. Yasakethu, S. Worrall, D. De Silva, W. Fernando, and A. Kondoz, "A compound depth and image quality metric for measuring the effects of packet loss on 3d video," in *Digital Signal Processing (DSP), 2011 17th International Conference on*. IEEE, 2011, pp. 1–7.

[32] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *Broadcasting, IEEE Transactions on*, vol. 50, no. 3, pp. 312–322, 2004.

[33] J. Joskowicz, R. Sotelo, and J. C. Lopez Ardao, "Towards a general parametric model for perceptual video quality estimation," *Broadcasting, IEEE Transactions on*, vol. 59, no. 4, pp. 569–579, 2013.

[34] I. Rec, "Bt.500-13, methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, 2012.

[35] R. W. Kennard and L. A. Stone, "Computer aided design of experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.

[36] G. W. Snecdecor and W. G. Cochran, *Statistical Methods*. Wiley, 1991.

[37] A. R. Gray and S. G. MacDonell, "A comparison of techniques for developing predictive models of software metrics," *Information and software technology*, vol. 39, no. 6, pp. 425–437, 1997.

[38] L.-X. Wang, "The wm method completed: a flexible fuzzy system approach to data mining," *Fuzzy Systems, IEEE Transactions on*, vol. 11, no. 6, pp. 768–782, 2003.

[39] B. S. de Lima and N. F. Ebecken, "A comparison of models for uncertainty analysis by the finite element method," *Finite Elements in Analysis and Design*, vol. 34, no. 2, pp. 211–232, 2000.

[40] K. M. Passino, S. Yurkovich, and M. Reinfrank, *Fuzzy control*. Citeseer, 1998, vol. 42.

[41] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 22, no. 6, pp. 1414–1427, 1992.

The Proposal of a Distributed Algorithm for Solving the Multiple Constraints Parking Problem

Khaoula Hassoune[1], Wafaa Dachry[2], Fouad Moutaouakkil[1], Hicham Medromi[1]
[1]Systems Architecture Team ,Laboratory of Research in Engineering, Hassan II University, ENSEM
Casablanca, Morocco
[2]Laboratory of Mechanical Engineering, Industrial Management and Innovation Hassan I University, FST
Settat, Morocco

***Abstract*—The parking problem in big cities has become one of the key causes of the city traffic congestion, driver frustration and air pollution.So to avoid these problems, parking monitoring is an important solution. Recently many new technologies have been developed that allows vehicle drivers to effectively find the free parking places in the city but these systems still limited because they don 't take into consideration road networks constraints. In this paper, We design a distributed system that will help drivers to find the optimal route between their positions and an indoor parking in the city taking into consideration a set of constraints such as (distance, traffic, amount of fuel in the car, available places in te parking, and parking cost). We propose a distributed technique based on multi objective Ant Colony Optimisation (ACO). The proposed method aim to manage multi objective parking problem in real time using the behavior of real ants and multi agent systems to decrease the traffic flow and to find the optimal route for drivers.**

***Keywords*—ant colony optimization(ACO); multi-agent system; parking monitoring; intelligent systems**

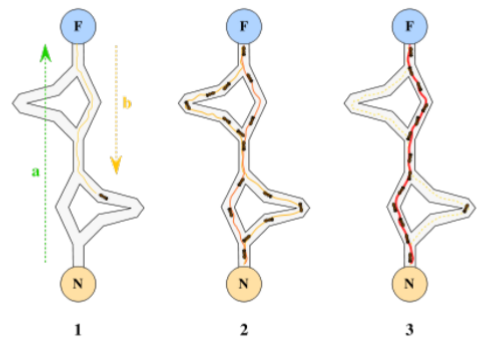


Fig. 1. The behavior of real ants

I. Introduction

Cities noticed that their drivers had real problems to find a parking space easily especially during peak hours, the difficulty roots from not knowing where the parking spaces are available at the given time. Even if this is known, many vehicles may pursue a small number of parking spaces which in turn leads to traffic congestion. So researchers are recently turned to apply intelligent transportation technologies for management of parking area by designing and implementation of smart parking systems that allows vehicle drivers to effectively find the free parking places. The main idea is to reduce traffic and time searching for a parking place.

Many systems are developed for parking but each system is based on different Technologies . To control the parking problem we need to use an intelligent system that can be adaptative with the environment and give us the optimal path to reach an indoor parking place taking into consideration multiple constraints.

Parking problem is situated in several researches based on different technologies and algorithms. Some of these systems are based on intelligent systems. In this domain, we have systems that are based on internet of things and cloud computing [1] [2] [3]. and others which are based on Artificial intelligence like fuzzy logic [4] and neural networks [5].

Ant colony optimization (ACO) [6] is a very powerful optimization heuristic for combinatorial optimization problems, Vehicle Routing algorithms [7] [8], and NP-complete problems. Ant Colony Optimization (ACO) is inspired by ants behavior to find the shortest paths between their colony and the source of food. Ants manage to discover the optimal path in an organized and distributed manner (Figure 1). In an ACO algorithm ants are programmed agents to find an optimal combination of elements in a given set of utility functions. The key ingredient in ACO is pheromones which are deposited chemicals by ants and their concentration traces a map of trajectories. Paths that contain a highest concentration of pheromones represent better trajectories. In this paper, we propose a mathematical model to decrease the time searching for an indoor parking taking into consideration a set of constraints (distance, traffic, amount of fuel in the car, available places and parking cost) . With this approach, we propose a new method to optimize parking timing based on a distributed multi-objective ACO algorithm and multi agent systems.

The outline of this work is as follows. In the next section, we define related works done to solve this type of problems. In Section 3, we describe the multi-objetive problem and how can be modeled as a graph. The Mathematical model for solving the multi objective parking problem is proposed in Section 4 that aims to find the optimal solution for drivers. Section 5 explains the distributed multi objective ACO based on multi-agent system for parking problem. Section 6 summarizes the perspectives. Finally, the conclusion is given in the last section.

II. Previous Works

In the literature, many researches about smart parking systems have been presented. These researches have used different information technologies like: wireless sensor network, internet of things, cloud Computing systems, mobile application, and artificial intelligence techniques.

The authors in [9] [10] present the design and implementation of a smart parking system based on wireless sensor networks that allow vehicle drivers to find the free parking places.

In [11] the author presents a wireless system for locating parking spots remotely via smartphone. This system automates the process of locating an available parking spot and paying for it.

Authors in [12] have proposed a scalable and low cost car parking framework (CFP) based on the integration of networked sensors and RFID technologies. These framework include driver guidance, automatic payment, parking lot retrieval, security and vandalism detection.

In others studies the authors have choose to design an automatic smart parking using internet of things which enables the user to find the nearest parking area and the available slot in that area [1] [2] [3].

In [13], an intelligent parking guidance algorithm using dynamic city parking conditions has been proposed. Simulation results were provided to validate the effectiveness of the proposed algorithm.

Samaras [4] proposed a fuzzy logic system based on the energy efficiency of the city 's parking lots. The wireless sensor network sends the actual data to generate fuzzy rules in the study.

Wang [14] has developed a based reservation smart parking that allows drivers to find and reserve the vacant parking places.Cai [15] proposed a navigation system based on the parking sensor network using a Dijkstra optimization algorithm to obtain the optimal parking route. . Zheng [5] has developed a prediction system for parking problem using some machine learning methods such as artificial neural network, support vector machines, and regression) using real data acquired from two cities like Melbourne and San Francisco.Mejri [16] proposed a multi-purpose intelligent parking approach based on a mata heuristic to optimize parking location using real data.

We notice that the guidance systems proposed by researches dont provide an optimal multi objective solution to find an indoor parking with available places.We design a distributed system that will help drivers to find the optimal route between their positions and an indoor parking in the city taking into consideration a set of constraints such as (distance, traffic, amount of fuel in the car, available places and parking cost). The proposed method is based on multi objective ACO and intelligent agents .

III. Problem Definition

In this case, the parking problem can be modelled as a graph with distributed drivers, service stations and parkings

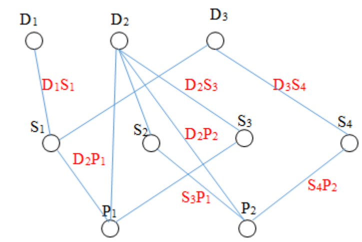


Fig. 2. The problem graph

where $D, = \{D_1, D_2, D_3, ...D_m\}$ is The set of drivers searching for parking,$P = \{P_1, P_2, P_3, P_4, ...P_n\}$ is the set of indoor parkings and $S_f = \{S_1, S_2, S_3.....S_p\}$ the set of service stations.

Let G be the graph(Figure 2) with vertices in $D \cup P \cup S$ (called nodes) and the edges (arcs) in E defined as follows: $E = \{D_iS_j \ / \ 1 \le i \le m \text{ and } 1 \le j \le p\} \cup \{S_iP_j \ / \ 1 \le i \le p \text{ and } 1 \le j \le n\} \cup \{D_iP_j \ / \ 1 \le i \le m \text{ and } 1 \le j \le n\}$. We denote D_iS_j the distance between the vertice D_i and S_j (between the driver D_i and the service station S_j) , S_iP_j the distance between the vertice S_i and P_j (between the service station S_i and the parking P_j) and D_iP_j the distance between the vertice D_i and P_j (between the driver D_i and the parking P_j).

A given route may contain a service station. the objective of the problem is to find the optimal path for each driver searching for a parking place by considering the following constraints:

- A Driver searches for an indoor parking with available places.
- Every route starts from the driver and finishes in a parking.
- The amount of fuel in the vehicle must allow the driver to reach at least one service station and a car park.
- Other parameters such us (traffic, distance, Available places, cost of parking, and amount of fuel) need to be taken into consideration in our optimization problem.

IV. Ant Colony Optimisation based Mathematical Model

In an ACO algorithm ants are programmed agents to find an optimal combination of elements in a given set of utility functions. The key ingredient in ACO is pheromones which are deposited chemicals by ants and their concentration traces a map of trajectories. Paths that contain a highest concentration of pheromones represent better trajectories. ACO is among heuristic methods that use distributed optimization concepts and pheromone concentration to solve combinatorial optimization problems [17]. The Ant Colony System Algorithm is divided into three main steps:

- All ants will construct their tours based on a probabilistic transition which is related to the amount of pheromone in their paths. After the construction of the trajectories, the ants will update the pheromone amount on each arc visited.

- The system will find the optimal path provided by the ant colony and chooses the optimal solution for each driver.
- In the end, the best ants update the overall pheromone on the paths that are part of the best solution.

The whole procedure repeats until the end of a number of iterations. The ACO (Ant Colony Optimization) algorithm for parking problem starts with a group of ants. These ants work together simultaneously to build routes. Each ant will build a complete circuit based on a given parking map. The construction of the route can be described as follows: First, the artificial ants (an intelligent developed agent) start from the driver 's location 's and select the next node to visit. the ant has two possibilities: The first is to select a gas station location to visit from the list of available gas station locations to reach a parking lot with available spaces and the second is to select a parking location to visit from the list of available parking spaces. When the ant reaches the parking 's position, then the system will restart the procedure of searching for another optimal path for an indoor parking with available places. In this version of ACO for parking problem, each ant needs to create a driver route that visits an indoor parking. Each ant select the next neighbor to visit using a probabilistic transition [18], if the ant is in a driver's position we have two possibilities (1) (2) :

$$p^k(D_i, S_j) = \frac{\tau^\alpha(D_i, S_j) \prod_{l \in \widetilde{P1}} \eta^{\alpha_l}(D_i, S_j)}{\sum_{\beta \in [1,p]} \tau^\alpha(D_i, S_\beta) \prod_{l \in \widetilde{P1}} \eta^{\alpha_l}(D_i, S_\beta)} \quad (1)$$

$$p^k(D_i, P_j) = \frac{\tau^\alpha(D_i, P_j) \prod_{l \in \widetilde{P}} \eta^{\alpha_l}(D_i, P_j)}{\sum_{\beta \in [1,n]} \tau^\alpha(D_i, P_\beta) \prod_{l \in \widetilde{P}} \eta^{\alpha_l}(D_i, P_\beta)} \quad (2)$$

where:

$\widetilde{P1}$={ distance,traffic,amount of fuel in the car } is the set of parameters of edges D_iS_j.

$\widetilde{P}$={ distance,traffic,available places ,parking cost ,amount of fuel in the car}is the set of parameters of edges D_iP_j.

$\tau^\alpha(D_i, S_j)$:value of pheromone trail on edge from D_i to S_j

$\tau^\alpha(D_i, P_j)$:value of pheromone trail on edge from D_i to P_j

α :coefficient that controls importance level of level of pheromone

$\eta^{\alpha_l}(D_i, S_j)$:value of cost functions for parameter l for edge from D_i to S_j

$\eta^{\alpha_l}(D_i, P_j)$:value of cost functions for parameter l for edge from D_i to P_j

α_l :coefficient that controls importance level of parameter l

The cost functions for all parameters for edge D_iS_j are :

$$\eta^d(D_i, S_j) = \frac{1}{D_iS_j} \quad (3)$$

The equation (3) presents the cost function for distance D_iS_j between the driver D_i and the service station S_j.

$$\eta^d(D_i, P_j) = \frac{1}{D_iP_j} \quad (4)$$

The equation (4) presents the cost function for distance D_iS_j between the driver D_i and the parking P_j.

$$\eta^t(D_i, S_j) = \frac{1}{\text{tr}(D_i, S_j)} \quad (5)$$

The equation (5) presents the cost function for traffic where tr(D_i, S_j) is the measured traffic flow from the driver D_i to the service station S_j.

$$\eta^t(D_i, P_j) = \frac{1}{\text{tr}(D_i, P_j)} \quad (6)$$

The equation (6) presents the cost function for traffic where tr(D_i, P_j) is the measured traffic flow from the driver D_i to the parking P_j.

$$\gamma(D_i) = \frac{fuel}{distance} \quad (7)$$

The equation (7) presents the amount of fuel consumed in litres by the driver D_i per km.

$$\eta^f(D_i, S_j) = sup(F(D_i) - \gamma(D_i) * (D_iS_j), 0) = \eta^f(D_i, S_{j0}) \quad (8)$$

$\eta^f(D_i, S_{j0})$ is the minimal fuel cost to reach the service station S_{j0} and Where F(D_i) is the amount of fuel in the vehicle of the driver D_i.

$$\eta^f(D_i, P_j) = sup(F(D_i) - \gamma(D_i) * (D_iP_j), 0) = \eta^f(D_i, P_{j0}) \quad (9)$$

$\eta^f(D_i, P_{j0})$is the minimal fuel cost to reach the parkingP_{j0} and Where F(D_i) is the amount of fuel in the vehicle of the driver D_i.

$$\eta^c(D_i, P_j) = \frac{1}{\text{cost}(P_j)} \quad (10)$$

The equation (10) presents the cost function for parking cost cost(P_j) in the parking P_j .

$$\eta^v(D_i, P_j) = \text{Card}(P_jv) \quad (11)$$

where P_j=P_{jv}∪ $P_{j\widetilde{v}}$ and P_{jv}∩ $P_{j\widetilde{v}}$=∅

P_{jv} is the set of available places in the parking P_j and $P_{j\widetilde{v}}$ is the set of non available places in the parking P_j. $Card(P_{jv}$)is the number of elements of the set $(P_{jv}$.

When the ant reach a Service station 's position we use the probabilistic transition to select the next parking to visit :

$$p^k(S_i, P_j) = \frac{\tau^\alpha(S_i, P_j) \prod_{l \in \widetilde{P}} \eta^{\alpha_l}(D_i, P_j)}{\sum_{\beta \in [1,n]} \tau^\alpha(S_i, P_\beta) \prod_{l \in \widetilde{P}} \eta^{\alpha_l}(S_i, P_\beta)} \quad (12)$$

where:

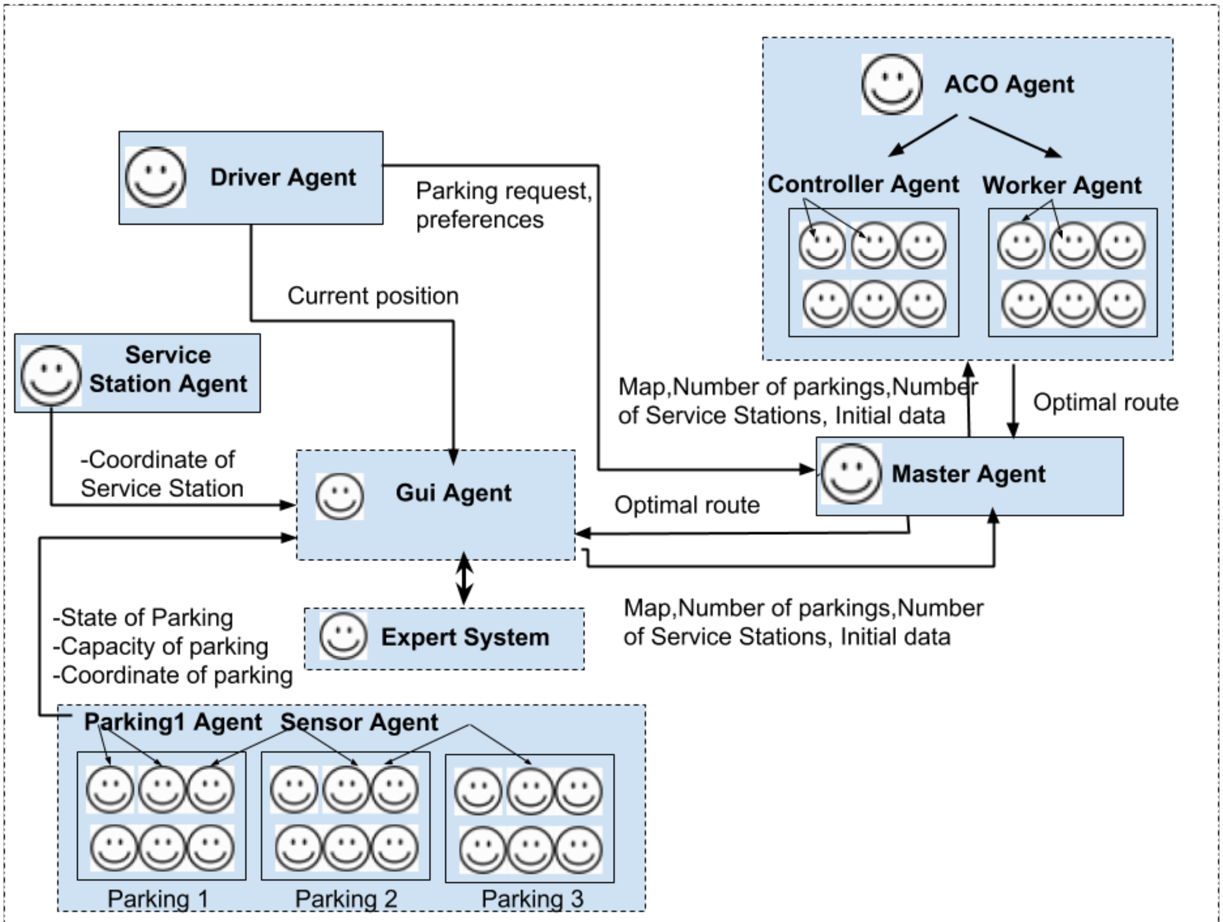


Fig. 3. Distributed Architecture for Parking

$\widetilde{P}$={ distance,traffic,available places,parking cost , amount of fuel in the car} is the set of parameters of edges D_iP_j.

$\tau^{\alpha}(S_i,P_j)$:value of pheromone trail on edge from S_i to P_j

α :coefficient that controls importance level of pheromone

$\eta^{\alpha_l}(S_i,P_j)$:value of cost functions for parameter l for edge from S_i to P_j

α_l :coefficient that controls importance level of parameter l

The cost functions for all parameters for edge S_iP_j are :

$$\eta^d(S_i,P_j) = \frac{1}{S_iP_j} \tag{13}$$

The equation (13) presents the cost function for distance S_iP_j between the service station S_i and the parking P_j.

$$\eta^t(S_i,P_j) = \frac{1}{\text{tr}(D_i,S_j)} \tag{14}$$

The equation (14) presents the cost function for traffic where tr(S_i,P_j) is the measured traffic flow from the service station S_i to the parking P_j.

$$\eta^f(S_{j0},P_k) = sup((\eta^f(D_i,S_{j0}) - \gamma(D_i)*(S_{j0}P_k)),0) \tag{15}$$

S_{j0} is the station that gives us the highest probability to go from D_i to a service station . $\eta^f(S_j0,P_k)$is the minimal fuel cost to reach the parking P_{i0}.

$$\eta^v(S_i,P_j) = \text{Card}(P_jv) \tag{16}$$

where P_j=$P_{jv}\cup$ $P_{\widetilde{jv}}$ and $P_{jv}\cap$ $P_{\widetilde{jv}}$=Ø

P_{jv} is the set of available places in the parking P_j and $P_{\widetilde{jv}}$ is the set of non available places in the parking P_j. $Card(P_jv)$ is the number of elements of the set P_{jv}.

This phase can be divided into two steps: a local and a global update [18]. The local update is done by each ant during the construction of the different routes. The global update of pheromone is performed when all ants reach parkings. The system will then be able to update all routes pheromone. The following rule (16) describes the process of the local update:

$$\Delta\tau_{ij}^k = (1-\rho)\tau_{ij}^{old} + \tau_{ij} \tag{17}$$

$$\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^k = \begin{cases} \frac{Q}{L_k} & (if\, ant\, k\, use\, the\, edge\, ij) \\ 0 & else \end{cases} \tag{18}$$

$(1-\rho)$is the pheromone reduction parameter $(0 < \rho < 1)$ .The parameter m is the number of ants , L_k is the length of the tour performed by ant k and Q is a positive arbitrary constant set to be equal to 100.

V. Distributed Multi-Objective ACO Algorithm Applied to Parking Problem

In this article, a distributed system based on multi-agent is proposed to solve the parking problem using Multi-Objective ACO (Ant colony Optimization) algorithm.the main idea is to find an optimized path for the driver that will allow him to find a place in an indoor parking while respecting certain constraints like (Traffic ,distance ,amount of fuel in the car ,parking price , Parking Capacity).

A. Distributed architecture based on multi agent

We propose a distributed system to control the parking problem in the city, our solution is based on a set of intelligent agents which communicate and collaborate together in real time to provide the best solution. In Figure 3, we present our architecture to model the new distributed ACO algorithm for parking problem. In this solution, our system contains a set of intelligent agents, each one with a specific behavior. Different types of agents with different objectives have been used in this distributed architecture. The first type is the Ant Colony Optimization Agent (ACOA) which is divided into two agents:

-the first one is the Worker agent (WA) and its function is nding the optimal route to an indoor parking in the given map by respecting the constraints . The behavior of WA is dened by the set of visited nodes (current tour), the best tour encountered and the best tour for this agent to reach the parking.

-The second is the Controller agent (CA),its function is the search for optimizing the best solution. Each Controller Agent in the environment has some information like the coordinate of the driver, the coordinate of the parking, the coordinate of the service station, the distance, traffic between itself and each controller agent in the environment, amount of fuel in the car of the driver, the number of available places, parking cost. The pheromone quantity is stored in each Controller Agent, so each one knows the level of pheromone on each link to another agent.

-An other type is Master agent and its function is to collect informations from the Gui agent and send it to the master agent.This agent controls the communication between agents and send the current best tour to the Gui agent.

B. Multi-agent Interaction Model

The interaction model (figure 4) proposed consists of a series of messages such as:

The user initiates the map of the parking locations, driver location and service stations locations, The GUI Agent interact with the knowledge base to store all this information. Then send it to the Master Agent. The Master Agent receives the Map data and initial parameters .sends a request message to the ACO Agent which will apply the (ACO) Algorithm to search for the optimal route which fits preferences of the user.

The Worker Agent sends the origin node and current best path to controller agent to record the present path and calculates cost to each node. When it finished the CA sends the list of neighbors, cost to each controller agent and best current path to worker agent. After that the worker agent chooses the next node and sends it to controller agent.

C. The Behavior of ACO Agent

In this solution, our system contains a set of intelligent agents; each one has a specific behavior. Three types of agents with different objectives were used in this distributed architecture. The first type is ant colony optimization agent (ACOA) and its function is constructing the new solution for finding the optimal routes to find a parking place in the given map by respecting the preferences of the driver. The behavior of ACO Agent is defined by The set of visited nodes (current tour), the best tour encountered and the best tour for this agent to reach the driver.

The second set of agents is the Manager agent (MA), its function is the search for optimizing the best solution. Each Manager Agent in the environment has some information like the coordinate of his parking, the number of places available in the parking, the number of reserved spaces, parking fees, the coordinate of other parking, also, the distance between itself and each Manager Agent in the city, and the distance between itself and the service stations in the city. The amount of pheromone is stored in each Manager Agent, so each Manager Agent knows the level of pheromone on each link to another agent.

The third type is Master agent and its function is to control the communication between agents and send the current best tour to the Interface agent

D. Ant colony algorithm based on multi agent system for solving parking problem

In this section, we present in detail the behavior of each intelligent agent in our distributed architecture. The Controller Agent has a collection of attributes: the quantity of pheromone, the level of evaporation, best solution cost, lists of nearest neighbor agents, the coordinate of the parking, the distance and traffic between itself and each controller agent in the environment, the number of available places, and parking cost. At each iteration, the Controller Agent records the present path. Table I show a pseudo-code of the Controller Agent behavior

TABLE I. The Controller Agent Behavior

INPUT:Graph, Initial parameters, Initial Preferences
OUTPUT : List of nearest nodes
GetAnt { RecordPathAnt(ant) CalculateCostFunctions(ant) UpdatePheromone(ant) ListBestneighbors(ant) Return BestNeighbors }

The behavior of CA is based on a set of functions

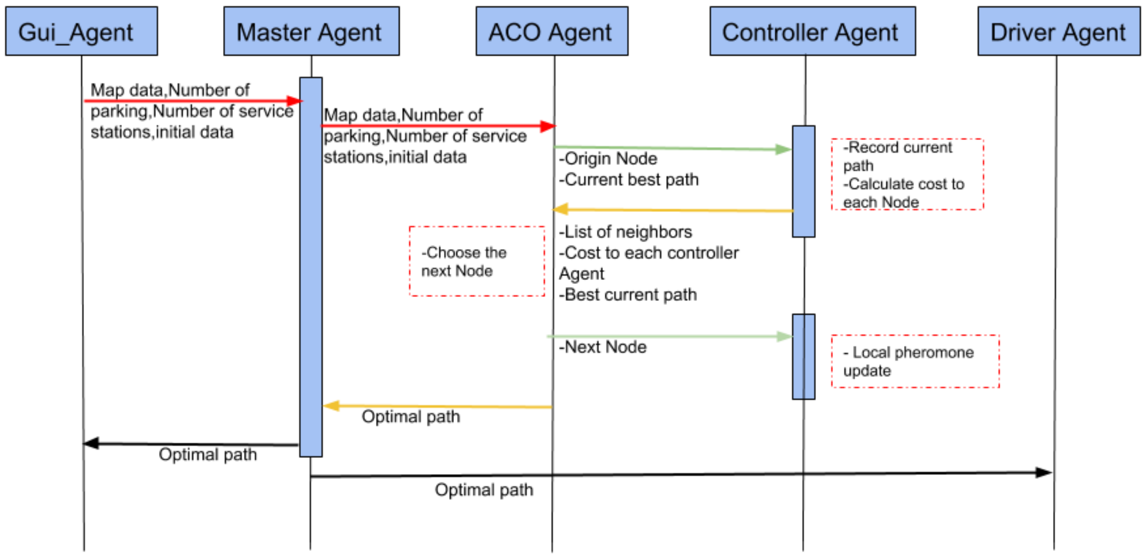


Fig. 4. Multi-agent Interaction Model

- The Controller agent saves the path of the present ACO Agent.
- Controller Agent calculates the cost functions for all problem parameters defined by the driver .updates the amount of pheromone for the received CA according to the information about its tour cost .
- The Manager agent constructs the list of best neighbors that will be sent to the Worker Agent for selecting the next node.

The objective of Worker Agent is computing the optimal path that connects a driver to an indoor parking in the city taking into consideration some constraints like (Traffic, distance, amount of fuel, Number of available places, parking cost). Table II presents the behavior of the Worker Agent.

- The Worker Agent(WA) Calculates the probabilistic transition to Select the next neighbor to visit and start a local pheromone update on this route. If the probabilistic transition of the path used by this agent is higher than the path stored at the current CA, the CA will update its path .
- At the end, the WA compares his solution with the best solution in the environment .If the cost of the path used by the WA is less than the best path in the environment, the WA will update the global amount of pheromone on this route and this path will be the best solution in the environment.

VI. Future Work

In the future work we will implement our solution using Java-Agent Development Environment (JADE) as a distributed and parallel platform. The proposed solution is deployed on several containers. At the beginning, we will need to fix the number of iterations. Each iteration ends when a set of Ant agents reaches an indoor parking and the parameters are initialized at the start of the iteration. In each iteration we collect the data (the best cost, the amount of pheromone, the best solution ...). At the end of each round, we can use these data to find the optimal solution for the driver taking into consideration network constraints like (Traffic, distance, amount of fuel,cost of parking, number of available places in the parking).

TABLE II. The Behavior of the Worker Agent

INPUT:Graph,Initial parameters
OUTPUT : Optimal path
• OptimalPath() • UpdateGlobalPheromone() • DisplayBestPath() OptimalPath() { CalculateProbabilisticTransition() SelectBestTransition() UpdateLocalPheromone() UpdateListBestPath() Return Bestpath }

VII. Conclusion

This paper presents a distributed solution based on multi objective ACO and multi agent systems to solve multi constraints parking problem .The proposed solution uses the behavior of real ants based on artificial agents .

The new solution was applied to solve the parking problem by creating an optimal that feets the preferences of the driver (Traffic, distance, amount of fuel, Number of available places in the parking, parking cost). The Parking problem is presented as a graph with multiple parking, each one is managed by a Manager agent and the process of communication between agents is described. In our proposed architecture, agents work together in parallel to find the optimal path for the driver. When one of the agents fails in his behavior, the system keeps running because we have a set of agents available to achieve this task.as future works we will consider other forms of parallel Ant Colony optimization algorithm and we will use this algorithm to solve other problems.

ACKNOWLEDGMENT

I would like to thank my advisors Ms.W.Dachry, Mr. F. Moutaouakkil and Mr. H. Medromi for the efforts they made to review this paper and for the many useful suggestions they gave me during the work on this paper.

I would also express my gratitude to every person that has contributed to the achievement of this works.

REFERENCES

[1] Y. R. Rao, "Automatic smart parking system using internet of things (iot)," *Int J Eng Technol Sci Res*, vol. 4, no. 5, 2017.

[2] Z. Suryady, G. R. Sinniah, S. Haseeb, M. T. Siddique, and M. F. M. Ezani, "Rapid development of smart parking system with cloud-based platforms," in *Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on*. IEEE, 2014, pp. 1–6.

[3] B. M. K. Gandhi and M. K. Rao, "A prototype for iot based car parking management system for smart cities," *Indian Journal of Science and Technology*, vol. 9, no. 17, 2016.

[4] I. Samaras, A. Arvanitopoulos, N. Evangeliou, J. Gialelis, and S. Koubias, "A fuzzy rule-based and energy-efficient method for estimating the free size of parking places in smart cities by using wireless sensor networks," in *Emerging Technology and Factory Automation (ETFA), 2014 IEEE*. IEEE, 2014, pp. 1–8.

[5] Y. Zheng, S. Rajasegarar, and C. Leckie, "Parking availability prediction for sensor-enabled car parks in smart cities," in *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on*. IEEE, 2015, pp. 1–6.

[6] Z. Ali and W. Shahzad, "Comparative analysis and survey of ant colony optimization based rule miners," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 1, 2017. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2017.080108

[7] H. Salehinejad, F. Pouladi, and S. Talebi, "A new route selection system: multiparameter ant algorithm based vehicle navigation approach," in *Computational Intelligence for Modelling Control & Automation, 2008 International Conference on*. IEEE, 2008, pp. 1089–1094.

[8] H. Salehinejad and S. Talebi, "A new ant algorithm based vehicle navigation system: a wireless networking approach," in *Telecommunications, 2008. IST 2008. International Symposium on*. IEEE, 2008, pp. 36–41.

[9] J. Yang, J. Portilla, and T. Riesgo, "Smart parking service based on wireless sensor networks," in *IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society*. IEEE, 2012, pp. 6029–6034.

[10] H. C. Yee and Y. Rahayu, "Monitoring parking space availability via zigbee technology," *International Journal of Future Computer and Communication*, vol. 3, no. 6, p. 377, 2014.

[11] O. Orrie, B. Silva, and G. Hancke, "A wireless smart parking system," in *Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE*. IEEE, 2015, pp. 004 110–004 114.

[12] E. Karbab, D. Djenouri, S. Boulkaboul, and A. Bagula, "Car park management with networked wireless sensors and active rfid," in *Electro/Information Technology (EIT), 2015 IEEE International Conference on*. IEEE, 2015, pp. 373–378.

[13] J.-H. Shin and H.-B. Jun, "A study on smart parking guidance algorithm," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 299–317, 2014.

[14] H. Wang and W. He, "A reservation-based smart parking system," in *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. IEEE, 2011, pp. 690–695.

[15] W. Cai, D. Zhang, and Y. Pan, "Implementation of smart parking guidance system based on parking lots sensors networks," in *Communication Technology (ICCT), 2015 IEEE 16th International Conference on*. IEEE, 2015, pp. 419–424.

[16] N. Mejri, M. Ayari, R. Langar, and L. Saidane, "Reservation-based multi-objective smart parking approach for smart cities," in *Smart Cities Conference (ISC2), 2016 IEEE International*. IEEE, 2016, pp. 1–6.

[17] M. Dorigo and C. Blum, "Ant colony optimization theory: A survey," *Theoretical computer science*, vol. 344, no. 2-3, pp. 243–278, 2005.

[18] M. Dorigo and L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem," *IEEE Transactions on evolutionary computation*, vol. 1, no. 1, pp. 53–66, 1997.

A Review on Event-Based Epidemic Surveillance Systems that Support the Arabic Language

Meshrif Alruily
Department of Computer Science
Jouf University, Sakaka, Saudi Arabia

Abstract—**With the revolution of the internet, many event-based systems have been developed for monitoring epidemic threats. These systems rely on unstructured data gathered from various online sources. Moreover, some systems are able to handle more than one language to cover all news reports related to disease outbreaks worldwide. The aim of this paper is to examine existing systems in terms of supporting the Arabic language. The 28 identified systems were evaluated based on different criteria. The results of this evaluation show that only 5 systems support the Arabic language using translation tools; hence, disease outbreaks in news reports written in Arabic are not directly processed. In other words, no existing event-based system in the literature has yet been developed specifically for Arabic health news reports to monitor epidemic diseases.**

Keywords*—*Public health; infectious disease; event extraction; disease surveillance system; arabic language

I. Introduction

During the past few years, the spread of many different pandemic diseases has increased worldwide; for example, the disease caused by the Ebola virus was first reported by the World Health Organization (WHO) in Guinea in 2014, which then spread rapidly to many West African countries causing hundreds of deaths (Guinea 346, Liberia 181, Nigeria 1, Sierra Leone 37) [1], [2]. It was also transmitted to other countries outside of the African continent, including Italy, the United Kingdom, Spain and United States of America. The latest information about Ebola can be reached via the following link. http://www.who.int/csr/don/archive/disease/ebola/en/.

In addition to the outbreak of Ebola in Africa, respiratory syndrome coronavirus (SARSCoV) was identified in Asia (2002/2003), an outbreak of the pandemic disease H1N1 influenza virus occurred worldwide (2009), and the Middle East Respiratory Syndrome (MERS) was found in Saudi Arabia (2012 - to date) [3], [4]. Therefore, the threat of infectious disease outbreaks to public health has prompted countries and organizations to develop several early warning surveillance systems [5]. However, the traditional surveillance systems or indicator-based surveillance systems need public health networks (Sentinel Networks) in order to collect predefined structured data about diseases on a routine basis from indicator sources, such as over-the-counter drugs and emergency department visits [6], [7]. Therefore, in the case of using passive surveillance systems, regular submission of monthly, weekly or daily reports of disease data by all health facilities is required. Although implementing this type of system has some advantages, such as its ability to cover all parts of a country, and its statistical power, it takes a couple of weeks for disease patterns to be detected and for the results regarding possible outbreaks to be disseminated; furthermore, not all countries have the required infrastructure to implement this system [3], [7], [8], [9]. On the other hand, as a result of the technological revolution of the internet, another type of surveillance system has emerged. This type is called the event-based surveillance system [10]. Generally, event-based surveillance systems can be described as real-time monitoring of diseases 24/7 through gathering information from informal sources, such as online news. According to Keller et al. [11], WHO's investigations into the majority of disease outbreaks are obtained through diverse online informal sources.

The remainder of the paper is organized as follows: in Section II, a background to the topic and a review of related work are presented. Section III describes methods used in performing this work. Section IV explores disease outbreak surveillance systems in Arabic. Section V discusses the state of the art of outbreak surveillance systems. Section VI presents the results and discussion. Finally, the conclusion of this work is presented in Section VII.

II. Related Work

According to Agheneza et al. [12], surveillance systems are classified into two types based on the type of data being processed. The first type is the indicator-based surveillance system (syndromic surveillance). The second type is the event-based surveillance system that collects and processes unstructured data from formal and informal sources, such as newspapers, reports, and medical websites. The current paper will review and focus solely on the event-based surveillance system, which utilizes text mining techniques to process media sources (news reports related to disease outbreaks) for text understanding, i.e. detecting and extracting infectious disease outbreak-related information, such as disease type, location name, date, and number of victims, if any. Also, more focus will be placed on systems that are able to process Arabic unstructured texts in the health domain.

A. Indicator-based Surveillance Systems

To date, many existing public indicator-based surveillance applications have been introduced to detect and track increases in disease incidence rates based on structured predefined information collected from different official sources, such as emergency room visits, telephone calls, and over-the-counter drug sales (syndromic/clinical surveillance data) [6]. Many detection methods are used to perform this task. Tsui et al. [13] categorized these methods into three types: temporal (SPC, regression, time series, and forecast-based methods), spatial (scan statistics), and spatiotemporal surveillance techniques.

Detailed information on these types of health surveillance systems can be seen in [13], [14]; the most recent review of these systems can be found in [15], [16]. In addition, in [17] the national communicable diseases surveillance systems proposed between 2000 and 2016 in developed countries were reviewed.

B. Event-based Surveillance Systems

On the other hand, many event-based surveillance systems have been developed for manipulating gathered unstructured data relating to infectious disease outbreaks from informal or non-traditional sources, such as online newspapers, news reports and social media [6]. Keller et al. [11] examined and compared the performance of three existing systems: EpiSPIDER, HealthMap and Global Public Health Intelligence Network (GPHIN), which have been developed to process event-based outbreak information. However, the most comprehensive review was conducted by Velasco et al. [3], who reviewed studies of infectious disease surveillance publications between 1990 and 2011 in detail, yielding 13 event-based systems. These systems were developed between 1994 and 2006, as listed in Table I, and used 15 review criteria: system name, system category, country, year started, coordinating organization, purpose, jurisdiction, supporting Arabic, disease type, public access, data processing, dissemination of data, most avid users, system evaluation and Homepage.

Choi et al. [5] performed a systematic review of web-based infectious disease surveillance systems, published in 2016. They identified 11 web-based surveillance systems, including GOARN, GPHIN, MedISys , BioCaster, HealthMap, ProMED, and EpiSPIDER, which had already been reviewed by Velasco et al. [3]. However, they also reviewed new systems: EpiSimS (now known as Object-oriented Platform for People in Infectious Epidemic OPPIE) [18], Google Flu Trends [19], GET WELL [20] and Influenzanet [21], which were not mentioned in [3], as can be seen in Table I.

To the best of the authors' knowledge, the recent review article was carried out by Yan et al. [22] and was published in October 2017. Developed systems in articles published between 2006 and 2016 were evaluated in terms of their methods, timeliness and accuracy outcomes.

III. Methods

Web-based infectious disease surveillance systems were systematically reviewed by focusing on multilingual systems, and systems dedicated to the Arabic language. Many electronic databases, such as Google Scholar, PubMed, Web of Science, IEEE Xplore Digital Library, and CiteSeerx were visited for reviewing the English literature published between 1994 and 2018. Moreover, many different terms or keywords were used for achieving the search; these include ”surveillance systems”, ”infectious disease systems”, ”event/internet-based surveillance systems”, ”Arabic surveillance systems”, ”syndromic surveillance”, ”biosurveillance”, and ”Arabic text mining systems”.

IV. Disease Outbreak Surveillance Systems in Arabic

Some of the developed systems, such as Argus, BioCaster, GOARN, GPHIN, HealthMap, MedISys and PULS, MiTAP and ProMED are able to process texts written in languages other than English. As previously mentioned, the aim of this study is to investigate event-based surveillance systems that can process Arabic texts in the domain of disease outbreaks. Al-Mahmoud and Al-Razgan [23] conducted a systematic review of published works on Arabic text mining between 2002 and 2014. The review showed that the topics of the articles were limited to a few different domains, such as opinion mining, crime domain, social networks, Arabic Wikipedia, and Islamic studies. Therefore, it is expected that disease outbreaks in news reports written in Arabic have not been directly processed; i.e. the 5 developed systems that support the Arabic language in Table I use translation tools to translate Arabic online news reports of disease outbreaks into English in order to process them for identifying the desired information. Further investigation on these systems can be seen below.

1) ProMED-mail
Monitoring Emerging Diseases (ProMED-mail) is a multilingual early warning system of emerging disease outbreaks developed in 1994 by Hugh-Jones [24], [25]. It monitors human, plant and animal diseases worldwide. Moreover, some surveillance systems depend on health warning reports produced by ProMED-mail, such as Argus, BioCaster and HealthMap. This system is available to the public and no subscription fees are required. The ProMED-mail's source of data depends on reports obtained from its subscribers. Currently, there are more than 70,000 subscribers from over 185 countries. With regard to the reports produced by the system, these are reviewed by a number of experts before dissemination [25], [51]. The system relies on its subscribers for performing the translation task. Although ProMED-mail utilizes translated Arabic news reports, disseminating information in Arabic is not provided.

2) The Global Public Health Intelligence Network (GPHIN)
GPHIN is a multilingual internet-based system developed by Health Canada in collaboration with the World Health Organization (WHO). The system is able to collect public health reports from global media sources, such as newswires and websites on a real-time basis in order to monitor infectious disease outbreaks. As previously mentioned, GPHIN is a multilingual system supporting eight languages (English, Chinese, Spanish, Portuguese, Russian, Arabic, French and Farsi) to monitor disease outbreaks by using machine translation to translate non-English reports into English, and vice versa. According to Wang and Barry [52], the GPHIN team supports the Indonesian language. Moreover, not only can GPHIN track events related to disease outbreaks or infectious diseases but it can also track other events, such as animal diseases, chemical incidents, plant diseases, and contaminated food and water. Most of the WHO information is provided by GPHIN. Furthermore, the Centers for Disease Control and Prevention (CDC), and the Food and Agriculture Organization of the United Nations (FAO) use GPHIN on a daily basis [12]. However, GPHIN is not free and official organizations must pay to subscribe. It also presents

TABLE I. DIFFERENT DEVELOPED TYPES OF EVENT-BASED SYSTEMS

No	System	Country	Year	Supporting Arabic	Homepage	Victim type	Working
1	ProMED-mail [24], [25]	USA	1994	Yes	www.promedmail.org	Human, animal, plant	Yes 28th Sep 18
2	GPHIN [26]	Canada	1997	Yes	https://gphin.canada.ca/cepr/listarticles.jsp?language=en_CA	Human, animal and plant	Yes 28th Sep 18
3	EWRS [27]	Europe	1998	No	https://ewrs.ecdc.europa.eu	Human	Yes 28th Sep 18
4	EpiSimS [18]	USA	2000	No	http://www.lanl.gov/projects/mathematical-computational-epidemiology/agent-based-modeling.php	Human	Yes 28th Sep 18
5	GOARN [28]	USA	2000	No	https://extranet.who.int/goarn/	Human	Yes 28th Sep 18
6	MiTAP [29]	USA	2001	No	http://mitap.sdsu.edu	Human	No 28th Sep 18
7	Proteus-BIO [30], [31]	USA	2002	No	Not available	Human, animal and plant	———————-
8	Argus [32]	USA	2004	Yes	www.biodefense.georgetown.edu	Human, animal, plant	No 28th Sep 18
9	MedISys and PULS [33]	Europe	2004	Yes	http://medisys.newsbrief.eu	Human, animal and plant	Yes 28th Sep 18
10	BioCaster [34]	Japan	2006	No	http://born.nii.ac.jp/ OR http://www.biocaster.org/	Human, animal, plant	No 28th Sep 18
11	EpiSPIDER [35]	USA	2006	No	www.epispider.org	Human, animal, plant	No 28th Sep 18
12	GODSN [36]	USA	2006	No	Not available	Human	———————-
13	HealthMap [37]	USA	2006	Yes	www.healthmap.org	Human, animal, plant	Yes 28th Sep 18
14	InSTEDD [38]	USA	2006	No	http://instedd.org	Human	Yes 28th Sep 18
15	Google Flu Trends [19]	USA	2008	No	www.google.org/flutrends "No longer publishing current estimates of Flu and Dengue fever based on search patterns." [39]	Human	Yes 6th Oct 17
16	Influenzanet [21]	Europe	2008	No	www.influenzanet.eu	Human	Yes 28th Sep 18
17	Animal disease-related event recognition system [40]	USA	2010	No	Not available	Animal	———————-
18	Automatic online news monitoring and classification [41]	USA	2010	No	Not available	Human and animal	———————-
19	GET WELL [20]	Sweden	2010	No	www.smittskyddsinstitutet.se	Human	Yes 28th Oct 18
20	CIDARS [42]	China	2011	No	Not available	Human, animal, food and waterborne	———————-
21	EpiCore [43]	USA	2013	No	https://epicore.org	Human and animal	28th Sep 18
22	Alshowaib's system [44]	KSA	2014	No	Not available	Human	———————-
23	Healthtweets [45]	USA	2014	No	www.healthtweets.org	Human	Yes 28th Sep 18
24	DESRM [46]	Vietnam	2015	No	Not available	Human	———————-
25	Flutrack [47]	Greece	2015	No	www.flutrack.org	Human	Yes 28th Sep 18
26	Online Diagnostic System [48]	Nigeria	2017	No	Not available	Human	———————-
27	ARGO [49]	Mexico, Brazil, Thailand, Singapore and Taiwan	2017	No	Not available	Human	———————-
28	PADI-web[50]	France	2018	No	http://epia.clermont.inra.fr/	Animal	Yes 28th Sep 18

information in English and French languages [26], [53].

3) A Global Detection and Tracking System for Biological Events (Argus)
Argus is an event-based system developed at the Georgetown University Medical Center; its aim is to provide early warning alerts through detecting and tracking global biological events that might affect human, plant, and animal health. It is able to cover multilingual data (40 languages) through a multilingual analytic team. For biological event detection, the system relies on a taxonomy of nearly 200 indicators. For assessing biological event evolution, a heuristic staging model called the Wilson-Collmann Scale is proposed. Moreover, Argus can monitor the spread of 130 infectious disease outbreaks covering 175 countries [32]. Argus depends on official and unofficial disease reports generated from WHO and ProMed, respectively, as indicators of possible biological events [54].

4) MedISys and PULS
The Medical Information System (MedISys) is part of the Europe Media Monitor (EMM) software family developed at the Joint Research Centre of the European Commission (JRC). It is an automatic multilingual early warning system used to identify potential public health threats, such as communicable diseases, bioterrorism, and radiological, nuclear and chemical incidents. In the medical domain, the system relies on predefined keywords for gathering textual reports from several web-based sources in various languages. It monitors around 1400 medical websites, 3750 generic news portals, 20 commercial newswires, and more than 10000 Really Simple Syndication (RSS) feeds in 60 languages [33]. MedISys issues automatic notification by sending SMS and emails to the end users. Moreover, the latest trends about diseases can be seen through MedISys's web interface. With regard to extracting disease-related information, the Pattern-based Understanding and Learning System (PULS) is utilized [55]. PULS extracts the disease name, number of victims, and their conditions, location and date. According to Agheneza [12], PULS is only able to process reports in the English language. MedISys provides three types of access levels [55]:
- Free access for public
- The access level is restricted for public health professionals outside the European Commission (EC)
- Full access inside the EC

MedISys and PULS website interface supports Arabic via the link http://medisys.newsbrief.eu/medisys/homeedition/ar /home.html. However, not all information provided is in Arabic and sometimes Google translate is used for translation some none-Arabic news reports.

5) HealthMap
HealthMap is a multilingual automated real-time web-based surveillance system. This system is free and is publicly available [37]. It can also be browsed in 7 languages: English, French, Portuguese, Russian, Chinese, Arabic and Spanish. Similarly, HealthMap uses a translation engine to handle non-English articles. HealthMap comprises five components as follows: data gathering from diverse online sources: newswires, Really Simple Syndication (RSS) feeds, ProMED Mail, and WHO Classification, Database, Web Backend and Web Frontend [56]. The system's tools are Linux, Apache, MySQL and PHP. It also utilizes free services provided by other developers, such as Google Translate API, Google Maps, GoogleMap API for PHP and xajax PHP AJAX library.

In this review, few systems were found in the literature that are able to directly process Arabic health data, i.e. without using translation engines. One such system, the named entity recognition system, NAMERAMA, has been developed to identify disease related-information such as diagnosis methods, symptoms, disease names and treatment methods from textual reports in the Arabic medical domain [57]. This system uses Bayesian Belief Networks (BBN) to extract aforementioned entities and is comprised of two stages: the first is the processing stage, which includes preprocessing, data analysis and feature extraction; the second stage is based on BBN for performing the classification task. The AMIRA tool is used for applying Part of Speech (POS) to the data, and an annotated corpus is used to evaluate the proposed system. However, only 27 articles were used for evaluating the performance of the NAMERAMA system, which is considered a very small dataset. In addition, this system only focused on identifying cancer disease-related information, i.e. other types of diseases were not covered. Table II lists the evaluation results.

In [58], two methods for identifying and extracting medical terms from the Arabic medical corpus were proposed. Their

TABLE II. THE SYSTEM EVALUATION RESULTS

Measure	Disease entity	Treatment method entity	Diagnosis method	Symptom categories
Precision	96.60%	69.33%	84.91%	71.34%
Recall	90.79%	70.99%	53.36%	49.34%
F-measure	93.60%	70.15%	65.53%	58.33%

work forms part of the Multimedica project, funded by the Spanish Ministry of Science and Innovation. The aim of the project is the development of multilingual resources and tools that include the Spanish, Arabic, and Japanese languages to process published reports by news agencies in the health domain. The first proposed approach is based on a gazetteer that contains 3473 Arabic medical terms. The terms used are translated from English medical terms resources (SNOMED and UMLS) using Google translator. In contrast, the second approach uses 410 Arabic terms that are the equivalents of Latin prefixes and suffixes commonly used in the medical and health domain. The evaluation results show that the first approach achieved 100% accuracy and outperformed the second approach; however, it only achieved 54% recall, which is relatively low.

With regard to infectious disease outbreaks, Alruily et al. [59] presented a preliminary work on developing a web-based surveillance system to track infectious diseases by extracting disease-related information from Arabic news textual reports. However, no results were reported because it was a foundational study.

V. LATEST DISEASE OUTBREAK SURVEILLANCE SYSTEMS

As can be seen in Table I, the most recent system to be developed is that of Alshowaib [44], who used a rule-based approach to extract disease outbreak-related information, i.e. named entities, such as disease name, date and location, location of the reporting authority, and outbreak incident. The rules were created based on analysis of textual disease outbreak reports. This system is solely dedicated to the English language. The performance of this system can be seen in the following Table III.

TABLE III. THE SYSTEM EVALUATION RESULTS

Measure	Entities	Relations	Events
Precision	1.00	0.70	0.90
Recall	0.75	0.70	0.80
F-measure	0.88	0.70	0.85

Nguyen and Nguyen [46], [60] developed the Disease Extraction System for Real-time Monitoring (DESRM). This system is used for Vietnamese online news. The approach used for performing this task depends on semantic rules and machine learning to extract infectious disease events. DESRM consists of two components: disease event identification from textual data, and disease event information extraction. For identifying phrases and detecting disease events, semantic rules and machine learning (maximum entropy model) are used.
For extracting related information of disease events: time, disease name, and place in the second component, Name Entity Recognition (NER) rules and dictionary are utilized. Table IV and Table V present the performance of the system.

TABLE IV. DESRM CLASSIFICATION EVALUATION

Approach	Precision	Recall	F-measure
Semantic rules & machine learning	75.07	79.76	77.33
Baseline (using only machine learning)	72.35	77.84	74.97

TABLE V. DESRM EVENT EXTRACTION EVALUATION

Approach	Precision	Recall	F-measure
Rules & NER	89.47	94.44	91.89
Baseline (using Rules)	83.55	92.02	87.58

The China Infectious Diseases Automated-Alert and Response System (CIDARS) was developed in 2008 by the Chinese Center for Disease Control and Prevention [42]. CIDARS uses three early warning methods: spatial-temporal model, temporal model and fixed threshold detection method [61]. Although CIDARS is able to detect signals of infectious disease, many false positive signals are produced [62]. Investigations performed in 2017 on surveillance and early warning systems of infectious disease developed between 2012 and 2016 in China can be seen in [63].

The EpiCore global surveillance project was established by the International Society for Infectious Diseases, the Skoll Global Threats Fund, HealthMap, the Program for Monitoring Emerging Diseases (ProMED-mail) and the Public Health Interventions Network (TEPHINET) in 2013 [43], [64]. The EpiCore system is an online platform used to verify informal health alert reports related to potential disease outbreaks by health experts. Moreover, a web-based diagnostic with epidemic alerts was proposed by Okokpujie et al. [48]. It is also able to prescribe medications based on symptoms and is used for issuing alerts about the outbreak of epidemic diseases. This system relies on data provided by the users. Hyper Text Mark-up Language, Cascading Style Sheets, Javascript, Ajax, PHP, MySQL were utilized for developing this system. For analyzing the collected data, a medical diagnostic engine called Infermedica was used.

Moreover, a web-based diagnostic with epidemic alert was proposed by Okokpujie et al. [48]. Also, it is able to prescribe medication based on the symptoms and issuing alert about outbreak of epidemic diseases. This system relies on data provided by the users. Hyper Text Mark-up Language, Cascading Style Sheets, Javascript, Ajax , PHP , MySQL were utilized for developing this system. For analyzing the collected data a medical diagnostic engine called Infermedica was used.

Arsevska et al. [50] developed the Platform for Automated Extraction of Disease Information from the web (PADI-web) to monitor infectious animal diseases. It uses data collected from Google News to extract epidemic disease related-information, such as number of victims, dates and locations. The information extraction process relies on rule-based techniques of data mining. For evaluating the performance of PADI-web, 352 news reports were used, achieving F-scores 95% of the diseases, 85% of the number of cases, 83% of dates and 80% of locations, respectively.

Osaghae et al. [65] proposed a web-based grassroots epidemic alert system. However, this type of system is a passive surveillance system, as defined by WHO, because it relies on official data collected from official places, such as health centers, hospitals and registered laboratories. Therefore, it is not covered in this review.

Several existing systems have been developed for specific disease outbreaks, such as influenza epidemics, but these are limited to a specific data source, e.g. Online Social Networks (OSN), such as Twitter. For example, Elhadad et al. [66] investigated social media in order to extract information about food-borne disease outbreaks by monitoring restaurants in New York City. A prototype was developed based on supervised machine learning to detect the review comments or discussions about a food poisoning incident, or the people affected by the incident. The system yielded high results; however, no specific numbers relating to the results were reported. Furthermore, the system works only on Yelp data collected from the Yelp website. Additionally, Talvis et al. [47], 2014, proposed the flutrack system (http://flutrack.org) for monitoring the spread of influenza epidemics. Every 20 minutes, the system gathers tweets written in English using the Twitter API to detect potential flu outbreaks. In other words, the aim of this system is to track and visualize influenza epidemics in real time. The list of searching tags, namely, influenza, flu, chills, headache, sore throat, runny nose, sneezing, fever, and dry cough were used to extract and track flu-related tweets. The system was evaluated and achieved an accuracy of 92%. Related works of systems proposed for epidemics of seasonal influenza can be seen in the Google Flu Trends system [19], MappyHealth application [67], Tracking Flu Infections on Twitter [68], detecting influenza epidemics by analyzing Twitter messages [69], HealthTweets.org: a Platform for Public Health Surveillance Using Twitter [45], and the ARGO system for monitoring dengue fever epidemics [49]. For further reading on these types of systems, a systematic review of the literature conducted on proposed systems published between 2004 and 2015 that detect and track a pandemic using online social networks was published in 2016 by [70]. In addition, the most recent review was presented by Pollett et al. [71] in 2017 for evaluating the internet-based biosurveillance performance of diseases caused by bacteria, parasites and viruses. Furthermore, other important disease outbreak surveillance systems are listed below.

1) Proteus-BIO
   Grishman et al. [30], [31] developed the Proteus-BIO system for creating and automatically updating a database with information on infectious disease outbreaks. Proteus-BIO system consists of five phases:
   - Data gathering of online daily news and from medical sources, such as WHO, ProMed and a medical forum
   - Text zoner, based on rule-based techniques
   - Information extraction engine (multi-phase finite-state transducer) to extract this information: disease name, date, location and number, type, and status of victims
   - Interface to a relational database
   - Web-based database browser connected to the original texts

   The technique used for the information extraction task is based on a multiphase finite-state transducer that includes tokenization, lexical lookup, finite-state pattern matching, and noun phrase and clause. The performance of this system achieved P= 0.79, recall 0.41, and F-measure 0.54.
2) BioCaster
   BioCaster is an automated web service that monitors global online media for detecting infectious disease outbreaks [34], [72]. The system is able to process 1700 Really Simple Syndication (RSS) feeds from different sources: the World Health Organization (WHO) outbreak reports, ProMED-mail, Google News, and the European Media Monitor. The system is limited to seven languages: English, Japanese, Chinese, Spanish, Thai, Vietnamese and French; it comprises four components: event recognition, topic classification, disease/location detection and named entity recognition. In addition, the visualization is provided in this system to the user by plotting extracted information on a Google map. BioCaster has been evaluated on a gold standard corpus of annotated news articles; for all named entity classes the system achieved an F-measure of 76.97. With regard to topic classification performance, the system was able to achieve 0.89 precision, 0.97 recall, and F-measure 0.93. However, BioCaster depends on ontology, which is a limitation in the system, as it is unable to identify new diseases or locations.
3) Automatic online news monitoring and classification for syndromic surveillance
   Zhang et al. [41] developed an automatic online news monitoring and classification system for syndromic surveillance on infectious disease. The system consists of three components:
   - Data collection
     Programs of web crawler are used to collate online news of infectious disease.
   - Data representation and feature selection
     Three methods of document representation: Bag of Words, Noun Phrases, and Named Entities, are used for transforming the full text to document vectors. Regarding the feature selection implementation, Correlation-based Feature Selection (CFS) was chosen.
   - Classification and evaluation
     For performing classification tasks, learn Bayes net (LBN), K-nearest neighbour (KNN), Support Vector Machine (SVM) and Naïve Bayesian (NB) are used. SVM achieved the best performance in the evaluation results.

VI. Results and Discussion

In general, the aim of this review was to examine existing event-based surveillance systems focusing on finding systems developed for the Arabic language. It found that 5 systems supported Arabic in different ways but were originally created for the English language. These systems, such as MedISys and PULS, GPHIN and HealthMap use translator engines to translate from non-English to English. However, a small number of systems have been developed for several specific languages other than English, such as Swedish and Vietnamese. With regard to the Arabic language, however, this review shows that no existing surveillance system for monitoring infectious disease outbreaks has yet been created to directly process Arabic texts. Processing texts to identify certain entities depending on translation is not sufficient; indeed, translation may in fact lead to identifying incorrect information. As Agheneza claimed [12], most event-based surveillance systems are based in the USA and Europe, and a few systems are based in Asia. However, only one system was found in the area of the Middle East and North Africa (MENA); this was developed by Alshowaib [44] but to date is not available online and developed to handle English health news reports.

Arabic is a Semitic language with a very complex morphology, as it is highly inflectional, and therefore, dealing with texts written in Arabic is highly complicated. Arabic is comprised of 28 letters (3 vowels and 25 consonants) that are used to form words. For correct pronunciation, the diacritical marks are used and are placed around the letters. Arabic can appear in different forms: Modern Standard Arabic (MSA), dialectal Arabic, and classical Arabic [73]. Ibrahim et al. [74] claimed that the main dialects in the Arabic word are: Gulf, Iraqi, Moroccan, Levantine, Yemeni, and Egyptian.

Furthermore, the Arabic medical domain faces a number of challenges. For instance, diglossia is common in the health community in Arabic countries [58]. According to Samy et al. [58], the English language is the primary teaching language in Egypt, Iraq, Jordan, and the Arabic Gulf countries for teaching all health courses at universities whereas French is used at universities in the North African Arab countries. Moreover, the communication languages used by health workers, either verbal or written, are English and French, even for patients prescriptions or their health reports.

VII. Conclusion

The aim of this review is to examine existing event-based surveillance systems focusing on finding systems developed for the Arabic language. This review shows that no existing early warning surveillance system for monitoring outbreaks of infectious diseases has yet been created to directly process Arabic texts, i.e, without using translator tools. This result might be due to some difficulties in the Arabic language itself and in the medical domain in particular. However, other event-based surveillance systems developed for detecting and tracking disease outbreaks are presented, and the components and performance of these systems are discussed. Most systems developed in the USA and Europe to process data are written in their own native language, and are then enhanced to serve other languages by utilizing a translator engine, which helps monitor the spread of pandemic diseases worldwide.

References

[1] University of Washington. (2015) Modeling the spread of ebola. https://sites.math.washington.edu/~morrow/mcm/mcm15/38725paper.pdf.

[2] Pan American Health Organization / World Health Organization (PAHO/WHO). (2014) Ebola virus disease (evd), implications of introduction in the americas. http://www.paho.org/hq/index.php.

[3] E. Velasco, T. Agheneza, K. Denecke, G. R. Kirchner, and T. Eckmanns, "Social media and internet-based data in global systems for public health surveillance: A systematic review," *Milbank Quart*, vol. 92, pp. 7–33, 2014.

[4] Centers of Disease Control and Prevention (CDC). (2018) Middle east respiratory syndrome (mers). https://www.cdc.gov/coronavirus/mers/.

[5] J. Choi, Y. Cho, E. Shim, and H. Woo, "Web-based infectious disease surveillance systems and public health perspectives: a systematic review," *BMC Public Health*, vol. 16, no. 1, p. 1238, 2016.

[6] E. Christaki, "New technologies in predicting, preventing and controlling emerging infectious diseases," *Virulence*, vol. 6, no. 6, pp. 558–565, 2015, pMID: 26068569.

[7] N. Collier and S. Doan, "Geni-db: a database of global events for epidemic intelligence," *Bioinformatics*, vol. 28, no. 8, p. 1186, 2012.

[8] World Health Organization. (2018, Aug.) National passive surveillance. http://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/passive/en/.

[9] B. Ramalingam. (2016) Real-time monitoring in disease outbreaks: Strengths, weaknesses and future potential. http://opendocs.ids.ac.uk/opendocs/handle/123456789/9940.

[10] World Health Organization. (2008) A guide to establishing event-based surveillance. http://www.wpro.who.int/emergingdiseases/documents/eventbasedsurv/en/.

[11] M. Blench, H. Tolentino, C. C. Freifeld, K. D. Mandl, A. Mawudeku, G. Eysenbach, and J. S. Brownstein, "Use of unstructured event-based reports for global infectious disease surveillance," *Emerg Infect Dis*, vol. 15, no. 5, May 2009.

[12] T. Agheneza, "A systematic literature review on event-based public health surveillance systems," Master's thesis, Hamburg University of Applied Sciences, 2011.

[13] K. L. Tsui, D. Goldsman, W. Jiang, and S. Y. Wong, "Recent research in public health surveillance and health management," in *2010 Prognostics and System Health Management Conference*, 2010, pp. 1–7.

[14] K.-L. Tsui, W. Chiu, P. Gierlich, D. Goldsman, X. Liu, and T. Maschek, "A review of healthcare, public health, and syndromic surveillance," *Quality Engineering*, vol. 20, no. 4, pp. 435–450, 2008.

[15] F. Azzedin, J. Yazdani, S. Adam, and M. Ghaleb, "A generic model for disease outbreak notification systems," *International Journal of Computer Science & Information Technology*, vol. 6, no. 4, p. 137, 2014.

[16] C. Abat, H. Chaudet, J.-M. Rolain, P. Colson, and D. Raoult, "Traditional and syndromic surveillance of infectious diseases and pathogens," *International Journal of Infectious Diseases*, vol. 48, pp. 22 – 28, 2016.

[17] H. Bagherian, M. Farahbakhsh, R. Rabiei, H. Moghaddasi, and F. Asadi, "National communicable disease surveillance system: A review on information and organizational structures in developed countries," *Acta Informatica Medica*, vol. 25, no. 4, p. 271, 2017.

[18] S. D. Valle. (2018, Aug.) Agent-based modeling. http://www.lanl.gov/projects/mathematical-computational-epidemiology/agent-based-modeling.php.

[19] G. Jeremy, H. M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, p. 1012, 2009.

[20] A. Hulth and G. Rydevik, "Get well: an automated surveillance system for gaining new epidemiological knowledge," *BMC Public Health*, vol. 11, no. 1, p. 252, 2011.

[21] C. E. Koppeschaar, V. Colizza, C. Guerrisi, C. Turbelin, J. Duggan, W. J. Edmunds, C. Kjelsø, R. Mexia, Y. Moreno, S. Meloni *et al.*, "Influenzanet: Citizens among 10 countries collaborating to monitor influenza in europe," *JMIR public health and surveillance*, vol. 3, no. 3, 2017.

[22] S. Yan, A. Chughtai, and C. Macintyre, "Utility and potential of rapid epidemic intelligence from internet-based sources," *International Journal of Infectious Diseases*, vol. 63, pp. 77–87, 2017.

[23] H. Al-Mahmoud and M. Al-Razgan, "Arabic text mining a systematic review of the published literature 2002-2014," in *Cloud Computing (ICCC), 2015 International Conference on*, April 2015, pp. 1–7.

[24] M. Hugh-Jones, "Global awareness of disease outbreaks: The experience of promed-mail," *Public Health Reports*, vol. 116, no. 2, pp. 27–31, 2001.

[25] J. Woodall, "Stalking the next epidemic: Promed tracks emerging diseases," *Public Health Rep*, vol. 112, no. 1, pp. 78–82., 1997.

[26] A. Mawudeku and M. Blench, "Global public health intelligence network (gphin)," in *7th Conference of the Association for Machine Translation in the Americas*, 2006.

[27] P. Guglielmetti, D. Coulombier, G. Thinus, V. F. Loock, and S. Schreck, "The early warning and reponse system for communicable diseases in the eu: an overview from 1999 to 2005," *Euro Surveill*, vol. 11, no. 12, pp. 7–8, 2006.

[28] D. L. Heymann and G. R. Rodier, "Hot spots in a wired world: Who surveillance of emerging and re-emerging infectious diseases," *Lancet Infect Dis*, vol. 1, no. 5, pp. 345–353, 2001.

[29] L. Damianos, J. Ponte, S. Wohlever, F. Reeder, D. Day, G. Wilson, and L. Hirschman, "Mitap text and audio processing for bio-security: A case study," in *Proceedings of the 14th Conference on Innovative Applications of Artificial Intelligence - Volume 1*, ser. IAAI'02. AAAI Press, 2002, pp. 807–814.

[30] R. Grishman, S. Huttunen, and R. Yangarber, "Real-time event extraction for infectious disease outbreaks," in *Proceedings of the Second International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 366–369.

[31] ——, "Information extraction for enhanced access to disease outbreak reports," *J. of Biomedical Informatics*, vol. 35, no. 4, pp. 236–246, 2002.

[32] J. M. Wilson, "Argus: a global detection and tracking system for biological events," *Advances in Disease Surveillance*, vol. 4, no. 21, 2007.

[33] European Commission's Joint Research Centre. (2018, Sep.) Medical information system (medisys). http://emm.newsbrief.eu/overview.html.

[34] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi, "Biocaster: detecting public health rumors with a web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, 2008.

[35] H. Tolentino, R. Kamadjeu, P. Fontelo, F. Liu, M. Pollack, and L. Madoff, "Scanning the emerging infectious diseases horizon - visualizing promed emails using epispider," in *Advances in Disease Surveillance*, 2007.

[36] S. A. Khan, C. O. Patel, and R. Kukafka, "Godsn: Global news driven disease outbreak and surveillance," in *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2006, p. 983.

[37] J. S. Brownstein and C. C. Freifeld, "Healthmap: the development of automated real-time internet surveillance for epidemic intelligence," *Eurosurveillance*, vol. 12, no. 48, 2007.

[38] T. A. Kass-Hout and di Tada Nicolas, "International system for total early disease detection (instedd) platform," *Advances in Disease Surveillance*, vol. 5, no. 2, p. 108, 2008.

[39] Google. (2018, Sep.) Google flu trends. https://www.google.org/flutrends/about/.

[40] V. Svitlana and H. H. William, "Computational knowledge and information management in veterinary epidemiology," in *IEEE International Conference on Intelligence and Security Informatics, ISI 2010, Vancouver, BC, Canada*, 2010, pp. 120–125.

[41] Y. Zhang, Y. Dang, H. Chen, M. Thurmond, and C. Larson, "Automatic online news monitoring and classification for syndromic surveillance," *Decision Support Systems*, vol. 47, no. 4, pp. 508–517, 2009.

[42] W. Yang, Z. Li, Y. Lan, J. Wang, J. Ma, L. Jin, Q. Sun, W. Lv, S. Lai, Y. Liao *et al.*, "A nationwide web-based automated system for early outbreak detection and rapid response in china," *Western Pacific Surveillance and Response*, vol. 2, no. 1, 2011.

[43] T. S. Lorthe, M. P. Pollack, B. Lassmann, J. S. Brownstein, E. Cohn, N. Divi, D. J. Herrera-Guibert, J. Olsen, M. S. Smolinski, and L. C. Madoff, "Evaluation of the epicore outbreak verification system," *Bulletin of the World Health Organization*, vol. 96, no. 5, p. 327, 2018.

[44] W. N. Alshowaib, "Rule-based information extraction from disease outbreak reports," *International Journal of Computational Linguistics (IJCL)*, vol. 5, pp. 37–58, 2014.

[45] M. Dredze, R. Cheng, M. J. Paul, and D. Broniatowski, "Healthtweets.org: A platform for public health surveillance using twitter," in *AAAI Workshop on the World Wide Web and Public Health Intelligence*, 2014, pp. 593–596.

[46] M.-T. Nguyen and T.-T. Nguyen, "Desrm: A disease extraction system for real-time monitoring," *Int. J. Comput. Vision Robot.*, vol. 5, no. 3, pp. 282–301, 2015.

[47] K. Talvis, K. Chorianopoulos, and K. L. Kermanidis, "Real-time monitoring of flu epidemics through linguistic and statistical analysis of twitter messages," in *2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization*, 2014, pp. 83–87.

[48] K. Okokpujie, A. Orimogunje, E. Noma-Osaghae, and O. Alashiri, "An intelligent online diagnostic system with epidemic alert," *International Journal of Innovative Science and Research Technology*, vol. 2, no. 9, pp. 327–331, October 2017.

[49] S. Yang, S. C. Kou, F. Lu, J. S. Brownstein, N. Brooke, and M. Santillana, "Advances in using internet searches to track dengue," *PLoS computational biology*, vol. 13, no. 7, p. e1005607, 2017.

[50] R. Goel, S. Fadloun, S. Valentin, A. Sallaberry, M. Roche, and P. Poncelet, "Epidnews: An epidemiological news explorer for monitoring animal diseases," in *Proceedings of the 11th International Symposium on Visual Information Communication and Interaction*, ser. VINCI '18. New York, NY, USA: ACM, 2018, pp. 1–8.

[51] ProMED. (2018, Sep.) About promed-mail. http://www.promedmail.org/aboutus/.

[52] Y. AndreaWang and M. Barry, "Making online outbreak surveillance work for all," *Annals of global health*, vol. 83, no. 3-4, 2017.

[53] World Health Organization. (2018, Sep.) Epidemic intelligence - systematic event detection. http://www.who.int/csr/alertresponse/epidemicintelligence/en/.

[54] H. Chen, D. Zeng, and P. Yan, *Argus*. Boston, MA: Springer US, 2010, pp. 177–181.

[55] S. RALF, F. FLAVIO, V. D. G. ERIK, B. Clive, V. E. Peter, and Y. Roman, *Text Mining from the Web for Medical Intelligence*. IOS Press, 2008.

[56] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports," *Journal of the American Medical Informatics Association*, vol. 15, no. 2, pp. 150–157, 2008.

[57] S. Alanazi, "A named entity recognition system applied to arabic text in the medical domain," Ph.D. dissertation, Staffordshire University, 2017.

[58] D. Samy, A. M. Sandoval, C. B. Diaz, M. G. Salazar, and J. M. Guirao, "Medical term extraction in an arabic medical corpus," in *Proceedings of the 8th Language Resources and Evaluation Conference*, 2012.

[59] M. Alruily, N. Hammami, and M. Goudjil, "Arabic web-based surveillance system for monitoring infectious disease outbreaks," *2015 Internet Technologies and Applications (ITA)*, pp. 358–360, 2015.

[60] M.-T. Nguyen and T.-T. Nguyen, "Extraction of disease events for a real-time monitoring system," in *Proceedings of the Fourth Symposium on Information and Communication Technology*, ser. SoICT '13. New York, NY, USA: ACM, 2013, pp. 139–147.

[61] W. Yang, Z. Li, Y. Lan, J. Ma, L. Jin, S. Lai, Y. Liao, W. Lv, Q. Sun, and J. Wang, "Chapter 7 - china infectious diseases automated-alert and response system (cidars)," in *Early Warning for Infectious Disease Outbreak*, W. Yang, Ed. Academic Press, 2017, pp. 133 – 161.

[62] R. Wang, Y. Jiang, X. Guo, Y. Wu, and G. Zhao, "Influence of infectious disease seasonality on the performance of the outbreak detection algorithm in the china infectious disease automated-alert and response system," *Journal of International Medical Research*, vol. 46, no. 1, pp. 98–106, 2018, pMID: 28728470.

[63] H. Zhang, L. Wang, S. Lai, Z. Li, Q. Sun, and P. Zhang, "Surveillance and early warning systems of infectious disease in china: From 2012 to 2014," *The International journal of health planning and management*, vol. 32, no. 3, pp. 329–338, 2017.

[64] Z. Haddad, L. Madoff, E. Cohn, J. Olsen, A. Crawley, J. Brownstein, M. Smolinski, J. Shao, M. Pollack, and D. Herrera-Guibert, "The epicore project: Using innovative surveillance methods to verify outbreaks of emerging infectious diseases," *International Journal of Infectious Diseases*, vol. 45, p. 19, 2016.

[65] E. N. Osaghae, K. Okokpujie, C. Ndujiuba, O. Okesola, and I. P. Okokpujie, "Epidemic alert system: A web-based grassroots model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 3809–3828, 2018.

[66] N. Elhadad, L. Gravano, D. Hsu, S. Balter, V. Reddy, and H. Waechter, "Information extraction from social media for public health," in *KDD at Bloomberg Workshop, Data Frameworks Track (KDD 2014)*, 2014.

[67] B. Norris, C. Boicey, and M. Silverberg. (2012) Mappyhealth application. http://mappyhealth.com.

[68] A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: Tracking flu infections on twitter," in *NAACL*, 2013.

[69] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA 10. New York, NY, USA: ACM, 2010, pp. 115–122.

[70] M. A. Al-garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, and A. M. Al-Kabsi, "Using online social networks to track a pandemic: A systematic review," *Journal of Biomedical Informatics*, vol. 62, pp. 1–11, 2016.

[71] S. Pollett, B. M. Althouse, B. Forshey, G. W. Rutherford, and R. G. Jarman, "Internet-based biosurveillance methods for vector-borne diseases: Are they novel public health tools or just novelties?" *PLoS neglected tropical diseases*, vol. 11, no. 11, p. e0005871, 2017.

[72] N. Collier, "What's unusual in online disease outbreak news?" *Journal of Biomedical Semantics*, vol. 1, no. 1, pp. 1–18, 2010.

[73] N. Boudad, R. Faizi, O. h. t. Rachid, and R. Chiheb, "Sentiment analysis in arabic: A review of the literature," *Ain Shams Engineering Journal*, 2017.

[74] H. S. Ibrahim, S. M. Abdou, and M. Gheith, "Idioms-proverbs lexicon for modern standard arabic and colloquial sentiment analysis," *International Journal on Natural Language Computing*, vol. 4, 2015.

KASP: A Cognitive-Affective Methodology for Designing Serious Learning Games

Tahiri Najoua, El Alami Mohamed
National School of Applied Sciences of Tangier
Abdelmalek Essaadi University
Tangier, Morocco

***Abstract*—Many research studies agree on the existence of a close link between emotion and cognition. Actually, much research has demonstrated that students with learning disabilities (LD) experience emotional distress related to their difficulties. In this regard, this article proposes a new methodology of designing intelligent games called KASP Methodology, it's a new approach applied to the serious games (SGs) design field. It includes new decisive factors for designing SGs for children with LD.**

The proposed methodology is based on four pillars which are: Knowledge, Affect, Sensory and Pedagogy and it aims helping designers of serious games for building suitable serious learning games for children with LD taking into account the cognitive and emotional aspects of the child learner in order to improve his learning rhythm and foster his emotional state related to learning in a playful and interactive environment.

***Keywords*—Methodology; Affective Computing; Serious Games; Learning Disabilities; Game Design; Knowledge; Pedagogy**

I. Introduction

Although the term learning disability has been used since 1962, there is no universally accepted and consistent definition. The current descriptions and definitions of learning disabilities are found in the World Health Organization Disability Document, in legislation and policy related to education, disability issues, psychology, medicine and human rights.

According to the World Health Organization, the learning disability is a medical term that refers to a permanent disorder of neurological origin. A learning disability is an impairment of one or more neuropsychological functions, which disrupts the acquisition, comprehension, use and processing of verbal or nonverbal information.

Many international organizations have come up with their own definitions of learning disabilities with different visions. The NIH(National Institutes of Health), for example, defines LDs as disorders that affect the ability to understand or use spoken or written language, to make mathematical calculations, to coordinate movements or to attract attention. However, learning disabilities occur in young children, disorders are generally not recognized until the child reaches school age.

In addition, the National Advisory Committee on Children with Disabilities (CNAMC) developed in 1967 a widely cited definition of learning disabilities: Specific Learning Disorder means a disorder in one or more of the basic psychological processes involved in the comprehension or use of spoken or written language, which can manifest itself in an imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations. The term includes conditions such as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia.

In general, learning disabilities are neurological treatment problems. These treatment problems can interfere with learning basic skills such as reading, writing and / or mathematics. They can also interfere with higher-level skills such as organization, time planning, abstract reasoning, long-term or short-term memory, and attention.

In order to improve the learning potential of children with LD, we propose in this article a new approach to design serious games with educational interest for the benefit of children with learning disabilities.

A serious game is a tool that aims to train more than to be entertained. The serious games market has, and continues to be, an undeniable popularity in various fields such as education, health, advertising, politics, etc. The synthetic definition of a Serious Game is the one proposed by video game designers Michael and Chen [1]: Any game in which education (in its various forms) in the primary goal, rather than entertainment.

Indeed, a Serious Game is characterized by the presence of a dimension playful and a serious dimension explicitly desired by its designer [2]. However, a Learning Game is a computer application that uses fun springs to catalyze the attention of learners and facilitate their learning. It has explicit educational goals and can be used for training at all levels [3].

Moreover, the main objective of the design methodology proposed in this article is to allow the implementation of serious learning games tailored to take into account several specifications not only cognitive but also the emotional character of the learner by benefiting from the advantages of affective computing in an educational approach.

In this perspective, the article first presents the potential of Game-based learning and an overview of learning disabilities, their symptoms and types. Secondly, it establishes the difference between a serious game and a learning game. Then, it draws up a literature review related to methods of designing serious games. Fourth, it introduces a new methodology of designing serious learning games for children with learning disabilities. Finally, it presents a discussion of the results obtained as well as the strengths and limitations of the proposed methodology.

II. Game-Based Learning Potential

Learning is an active process of combining new information with old ones, a process of building networks and relationships between and in the fields of knowledge [4].

Despite the fact that learning theories ignored the importance of affective processes in learning and focused on cognitive aspects. Many specialists in psychology (Wilhem Wundt) and pedagogy (Montessori) have proved the close relationship between emotion and learning. Schutz et al.[5] defines emotions as ways of being and as holistic episodes that include physiological, psychological and behavioral aspects.

In addition, the neuroscientist Damasio explains how minds emerge from emotions and feelings, his insight confirms the essential role of emotions in cognition: Recent advances in the neuroscience of emotions highlight the links between cognitive and emotional functions that have the potential to revolutionize our understanding of learning in the context of schools [4].

In this regard, Hascher proposes a new framework for the dynamic interaction of learning and emotion in the school context where she stresses the importance of integrating the two cognitive and affective approaches, she experiences that the emotion plays a role parallel to that of cognition in any learning environment [6].

As a result, the links between emotion and learning are bidirectional and interdependent, because emotions influence thoughts, and infer new emotions. Therefore, building the self-esteem, interest, and intrinsic value of education that boosts enthusiasm for learning and reduces the negative emotions of the learner will improve his learning rhythm.

III. Learning Disabilities

A dysfunction in the neurological treatment of the child leads to the onset of learning disabilities, they are disabilities that disrupt as well as academic performance as daily interactions such as relationships with peers and that last a lifetime. Its a neurological disorder that comes from a difference in how a person's brain is wired.

Children with learning disabilities may have difficulty reading, writing, spelling, reasoning and organizing information, they are intelligent or smarter than their peers. So they need unconventional learning.

Learning disabilities appear as developmental delays or as difficulties in the aspects of functioning listed below:

- Limited attention capacity;
- Bad memory;
- Difficulty following directions;
- Inability to distinguish between / among letters, numbers or sounds; Bad reading and / or writing skills;
- Hand-eye coordination issues; poorly coordinated;
- Difficulties with sequencing, and / or disorganization and other sensory difficulties;

Relatively to these signs, each child displays one or more symptoms that indicate the type of disorder the child is suffering from as shown in Fig. 1. In fact, learning disabilities include:

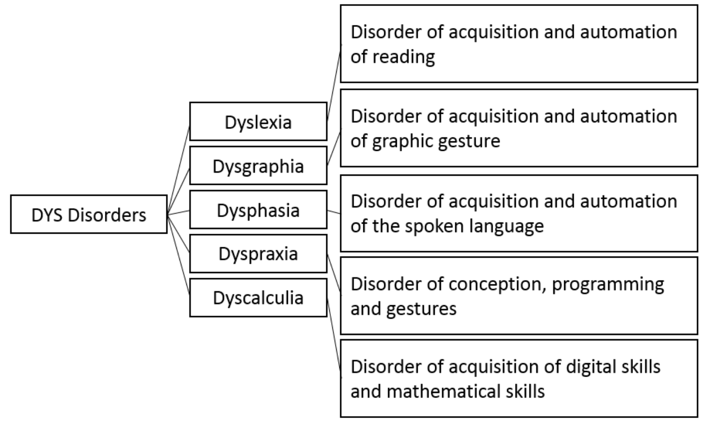


Fig. 1. Dys disorders

- Dyslexia or reading difficulties: Dyslexic children have difficulty reading correctly and fluently because they have a bad association between graphemes (written signs) and phonemes (sounds). They may also have problems with reading comprehension, spelling, and writing;
- Dyspraxia: it is often associated with visio-spatial deficiencies or difficulties in the muscular control which causes problems of coordination as well as disturbances of organization of the gaze which disturb the apprehension of the environment by the child;
- Dysgraphia: it refers to the physical difficulty forming words and letters. A dysgraphic child may have problems, including unreadable writing, irregular spacing, poor space planning on paper, bad spelling, and difficulty composing, writing, and thinking at the same time;
- Dyscalculia: Children with this type of learning disability may have a poor understanding of mathematical symbols, have difficulty remembering and organizing numbers, have trouble telling time, or have trouble counting;
- Dysphasia: it comes in various forms relating to difficulties in expressing oneself, or understanding what is said. Dysphasia affects more than one component of language in a variable way;

It is important to realize that learning disabilities are distinct from learning difficulties because these are temporary barriers closely related to socio-emotional conditions. Once the elements in question are eliminated, the person usually finds a normal learning rhythm. Unlike learning disabilities that are persistent and permanent and result from impairments in one or more learning-related psychological processes, in combination with normally average skills essential to thinking and reasoning.

IV. Serious Games and Learning Games

Despite the many treatises published on the term 'game', researchers still find it elusive. Its a structured, rules-based activity, usually done for fun. However, games can have another purpose other than fun, and can position themselves as a teaching tool bringing together playfulness and learning at a time. Also, Gaussot [7] describes the game as an educational means contributing to the emotional, sensory, motor, cognitive, intellectual and social development of the individual.

Golinkoff et al. [8] say that Play = Learn. As children move from sandbox to classroom, play should be the cornerstone of their education. The research is clear: Game pedagogy supports socio-affective and academic forces while instilling a love of learning. Many literature reviews were conducted on learning games, Abdul Jabbar and Felicia[9] established a study to discover how the design of play-based activities influences engagement and learning. They developed a set of general recommendations for the educational design of Game-Based Learning.

It should be mentioned that the word serious game exists since the 1960s when Clark Abt baptized his book Serious Game, where he explores the ways in which games can be used to educate and inform as well as to provide pleasure, he proposes its own definition: have an explicit and thoughtful educational purpose and are not intended to be played primarily for fun. This does not mean that serious games are not or should not be entertaining [10].

As for Egenfeldt-Nielsen et al.[11], they believe that Serious Games as digital games and equipment with a program of educational design and beyond entertainment. To the extent that learning is characterized by being an active, constructive, self-directed and emotional process.

Therefore serious games have the potential to support this sequence, as they present aspects of motivation, accuracy, challenge as well as information on request and just in time. Through play, the child-learner fully engages in training and becomes master of his learning in defiance of traditional passive learning.

V. Design Methodologies of Serious Games: Litterature Review

There are several methods of designing educational games, which vary according to the context used, the objectives to be achieved and the target audience. In the following a brief presentation of serious educational game design methods.

Before discovering the different methods of designing a Serious Game, first defining the term 'Game Design'. According to Salen and Zimmerman [12]: it is a process by which a designer creates a game, intended to be used by a player, so that a game experience is born. Alvarez et al.[2] defines it as: The process of inventing in a broader sense, game design refers to the idea behind a game.

Since a Game Design is a process, it means a continuous sequence of operations or activities leading to a predefined goal. Consequently, its necessarily a collaboration between several actors, hence the notion of 'collaborative'design method. The work that has been done to find a generic approach to design serious learning games is based on three pillars, such as the one proposed by Lloyd Buck [13], which confirms that these components must be attractive, responsive to objectives and specific educational applications.

Paquette et al. [14] proposes a learning systems engineering method called MISA, which aims to apply cognitive science principles to the field of instructional design. It is a very robust approach because it implements the most rigorous teaching practices. In the same vein, the Swedish Tele-Pedagogical Knowledge Center (1993) has created its own design method for developing about twenty standard documents describing all parts of the training. In addition, KTM Advance, a leading digital professional training company, offers an industrial design model specific to Serious Games [15].

Several approaches of designing serious games are inspired by pedagogical engineering, the DODDEL model (Document-Oriented Design and Development of Experiential Learning) proposed by Mcmahon [16], it has key features in that it promotes a theoretically inclusive approach to learning, a focus on game elements and an emphasis on documentation to provide the rigour necessary to be used as part of a broader project management model.

Furthermore, Marfisi-Schottman suggests a collaboration platform that offers a methodology and adapted tools to guide the various actors that participate in a Learning Games conception.

Although there is a plurality of methods that have a certain maturity in the field of e-learning, no approach concerns serious learning games for children with learning disabilities and which affect not only their ability to learning but also their interaction with their environment. The choice of educational activities related to this specific type of games should be carefully designed to encourage the learner to be a central player in the learning cycle.

We note that the different design methodologies of serious games previously cited all converge towards an approach that is based on the cognitive aspect of the learner where some rely on pedagogical engineering, others propose a collaboration between designers. However, it is necessary to mention that the design methodology of intelligent serious games must support the evolution of knowledge and involve the players profile in all stages of the design.

Indeed, the affective aspect is always isolated, although the factor related to the emotion occupies a crucial place in any process of learning, also the audio-visual and kinesthetic specifications are ignored despite their importance in the engagement of the player in any serious learning game.

On the other hand, the learning style is also considered as a major factor in the design of serious games because it makes it possible to adapt the proposed system to the preferences of the learner. It is defined by distinctive cognitive, affective, physiological and sociological behaviors; these behaviors serve as relatively stable indicators of how an individual perceives and processes information, interacts and responds to the learning environment [17].

Recommending serious learning games, according to the kind that best suits the characteristics of a learner, could improve learners motivation and lead to a better learning experience [18]. As a result, the adaptation of learning systems to learning styles and gamification seems to be a promising concept.

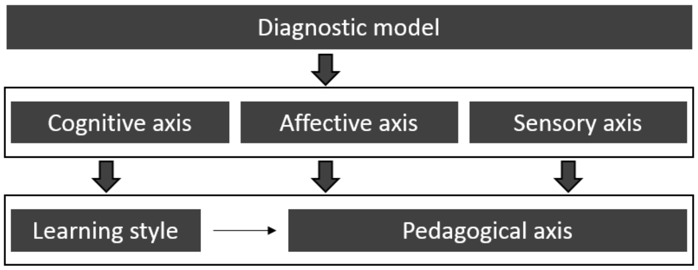


Fig. 2. The components of KASP methodology

VI. Design Methodology Diagnostic-Centered of Serious Game for Children with Learning Disabilities

In this regard, the authors propose a new methodology of designing serious games well adapted to the cognitive and affective needs of children with LD. Conceptualizing an intelligent game to develop the behavior, knowledge and reflections of a child presents a challenge especially when it comes to a player with learning disabilities taking into account basic criteria such as: the pedagogical environment, affective and cognitive reactions of the learner, the learning context, and also the learners profile.

In the interest of developing an effective intervention plan that will help children with LD to learn better, a first step of screening is therefore required, which is to gather information from various sources: parents, teachers, and speech therapists, so to determine the educational needs of the child and then indicate the appropriate educational strategies taking into account cognitive, affective and physical factors.

This early detection requires the coordination of all people revolving around the child, in order to formulate diagnostic hypotheses to improve the capacities considered to be deficient.

The next step is that of the diagnosis conducted by a clinical team, it consists of a list of tests to specify: personal history of the child, several examinations of type: psychological, cognitive, and somatic including a sensory examination.

In this perspective, we propose to use a multi agent platform called DIAUTIS [19] that offers a wide and variable collection of tests, including computerized interviews with parents, using sensors to analyze the interaction of the child through the monitoring of ocular activity, the observation of movements, as well as the interpretation of vocal responses.

Through these sensors, the DIAUTIS system is able to deduce the emotional state of the child diagnosed through its multi agent structure scalable and flexible.

The results of the tests performed via DIAUTIS will be analyzed in order to deduce the basic skills of the child by developing a deliverable containing five specifications: motor, behavioral, cognitive, socio-emotional, and sensory. An intelligent system for classifying these specifications is put in place using the techniques of Machine Learning that will give rise to the type of disorder that the diagnosed child suffers.

Machine learning is a domain derived from artificial intelligence that makes machines capable of learning, in simple terms, is the ability to identify models that can analyze larger and more complex data and to obtain faster and more accurate results, even on a very large scale, and to make decisions with minimal human intervention. The renewed interest in Machine Learning is due to the efficiency of the algorithms it understands in terms of accuracy and speed.

In this case, the authors propose to use one of the famous supervised type algorithms, SVM (Support Vector Machine), on the set of specifications selected from the diagnosis, in order to obtain the type of the disorder of the child in question thanks to the intelligent classifier. The proposed methodology consists of four axes, each axis includes an iterative process allowing its development through the experts intervention:

- The knowledge axis (KA): distinguishes different types of knowledge and links, which prepares the choice of educational activities.
  At this level, the serious game is assimilated to knowledge-based system, where unity 'knowledge 'is the key element since through this, we take cognitive deficits of the child with LD.
  Each concept is represented hierarchically in the brain of the child, at each new experience, the model of knowledge undergoes some modifications. So the first goal is to acquire the model corresponding to a given concept, to extract knowledge.
- The affective axis (AA): the omnipresence of the emotional aspect is essential, because the emotional state of the child can significantly change the pace of learning.
  This axis aims to raise the difficulties related to emotion and social interactions in order to develop them implicitly via the games on offer.
- The sensory axis (SA): it allows to specify the auditory, visual and kinesthetic parameters of the child to adapt the choice of the game to the player, the choice of interfaces, colors and sounds is done at this level.
- The pedagogical axis (PA): the objective through this one is to specify the basic pillars for learning from an institutional point of view, by making possible a mental activity of the learner in a learning situation.

The knowledge, emotional and sensory axes are derived from the diagnostic model, and the learner style and the pedagogical axis are developed by pedagogical experts by analyzing the three axes previously mentioned as presented in Fig. 2.

The KASP methodology is based on four pillars which are: Knowledge, Affect, Sensory and Pedagogy, as seen in Fig. 3,and it contains three main phases which are:

1) Preliminary analysis (PAP)
2) Development of conception phase (DCP) which is a prototyped cyclic process composed of three stages as follows:
   a) Elaboration of the conceptual architecture (CA)
   b) Development of the prototype (PD)
   c) Test and validation stage (TE)
3) Deployment phase (DP)

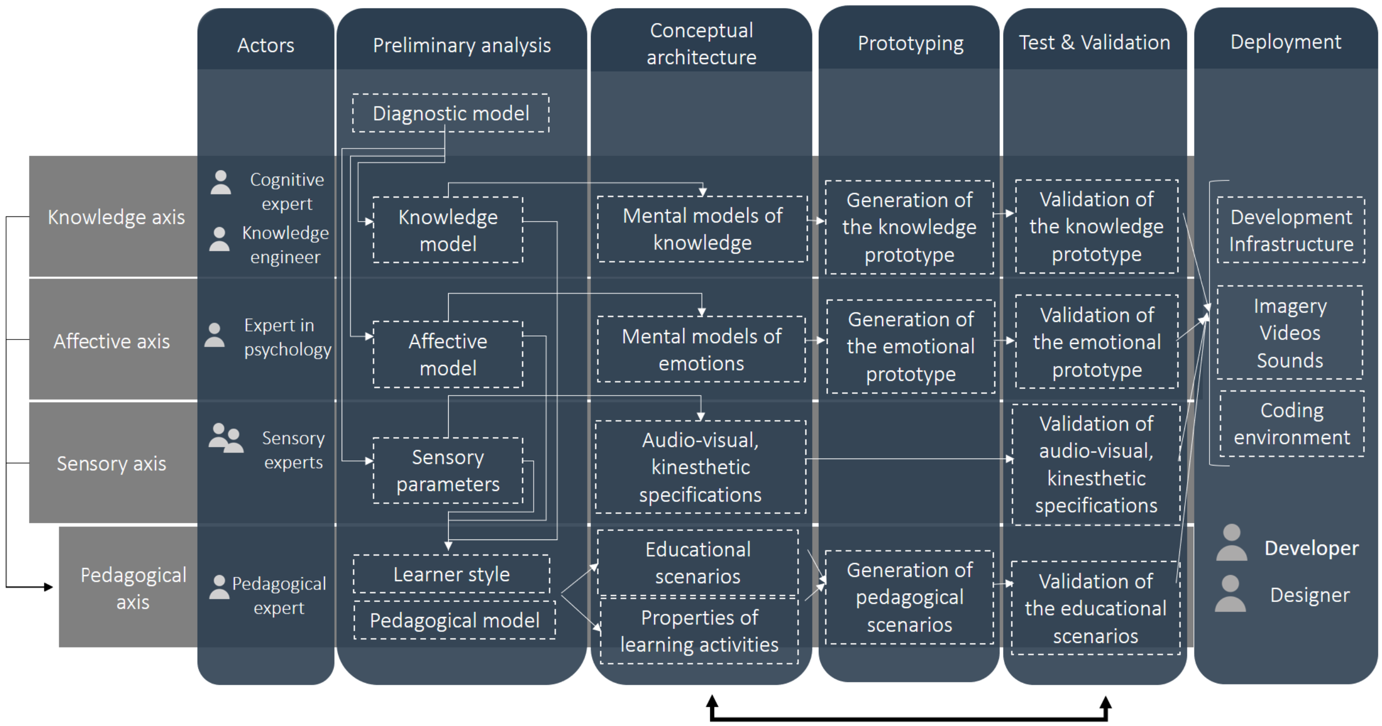


Fig. 3. KASP methodology

In the first phase PAP: It allows a better understanding of the system based on the understanding and the structuring of the needs.

- For the KA: An exploitation of the diagnostic model by the cognitive expert is implemented to clarify the deficits relating to the childs knowledge.
A good expression of the childs cognitive needs will allow a better choose of the pedagogical activities in order to correct any imperfection of the executive functions (EF), which are the cognitive skills necessary to control ones thoughts, emotions and actions.
- For the AA: Also the EF occupy a central place in the psychological development of the child, from where an intervention of the expert in psychology of the children is necessary in order to draw the emotional specifications relating to the identified disorder.
- For the SA: Three other experts come into play, hearing, vision and kinesthetic, the goal is to analyze the diagnostic model to specify if the child suffers from a hearing, visual or kinesthetic abnormality in order to carry out a needs analysis sensory appropriate to the child.
- For the PA: following the deliverables of the KA, AA and SA, the pedagogical expert determines the pedagogical objectives of the game to be chosen, it also describes the path that the learner must take to acquire the knowledge.

In the second phase DCP: The goal is to grasp the operational requirements, thus to focus on the main features and functions to be provided. The stage is composed of three iterative sub-steps, starting with the development of the conceptual architecture:

- For the KA: the cognitive expert elaborates the set of mental models of the child, which are cognitive representations illustrating the way in which the child perceives a given concept, in other words a mental model is a form of analogical representation of knowledge. Any inadequacy of the mental model is followed by a state of disorientation.
- For the AA: the psychological expert prepares the mental models of the childs emotional states in order to act in the direction to regulate the emotions that influence his learning in a destructive way.
- For the SA: in terms of sensory, the experts specify the audio-visual- kinesthetic parameters to take into consideration when elaborating the serious games.
- For the PA: the pedagogical expert defines the pedagogical scenarios which are the set of learning situations aiming at the appropriation of a precise set of knowledge taking cognizance of the cognitive and affective mental models as well as the sensory parameters.

Relative to the prototyping step, the four axes provide prototype deliverables for test and validation by the experts in each axis. The operation is repeated as and when to obtain the best prototype in order to move to the deployment phase. For the third step (Test and validation): it enables to identify the problems, to analyze concretely their causes and to propose solutions that are implemented in the future prototype.

In the third phase DP: Several actors intervene at this stage, namely: study and development engineers, graphic designer, expert in the design in order to implement a software prototype answering the needs and the conceptual architectures expressed in the steps PAP and DCP. The made prototypes are an implementation of the educational scenarios in a playful environment.

In brief, The KASP methodology begins with a preliminary analysis studying the diagnostic model delivered by the clinical team (as can be seen in Fig. 4).

In a first step, three models are concluded: knowledge, emotional, pedagogical, as well as a deliverable concerning the sensory specifications of the diagnosed child. At this level the objectives to be achieved are studied.

In a second step, the development of the conceptual architecture is set up by developing the mental models relating to knowledge, emotions, as well as the orientation of the pedagogical scenarios based on the learning style found, these orientations obviously take into account auditory, visual and kinesthetic specifications.

In a third step called prototyping, we proceed to the generation of prototypes of knowledge, emotional, and the generation of pedagogical scenarios, in other words it is at this level that the design of the scenes of the proposed game are built.

In the fourth step, the experts intervene to test and validate the prototype stage in order to guarantee a design adapted to the purposes set in the first step. These last three steps enter a loop to refine the result before moving to the deployment stage, in which the technical environment is set up.

KASP methodology serves the strengths of artificial intelligence that lies in the intelligent management of the knowledge collected, more precisely in the representation of knowledge through semantic networks [20] formed by several elements in interrelation, these elements are: nodes, links, and the set of operations constituting the reasoning mechanisms.

The semantic networks make it possible to structure the knowledge base thanks to the organizational axes (generalization, partition, aggregation). More precisely, the authors suggest using learning networks that build or extend their representations by acquiring knowledge from examples. New knowledge can change the old network by adding and removing nodes and arcs or by modifying numerical values, called weights, associated with nodes and arcs [21].

In addition, identifying the emotion of the learner through affective computing [22] is one of the strong points of the KASP methodology. Indeed, emotional learning is concerned with how learners feel while they are learning, as well as the way in which learning experiences are internalized so that they can guide the attitudes, opinions and behavior of learners [23].

Multiple research studies related to the integration of affective computing in intelligent learning systems (ILS) have been put in place, we quote that of de Arriaga et al.[24] presenting the implementation of a factory of ad hoc agents

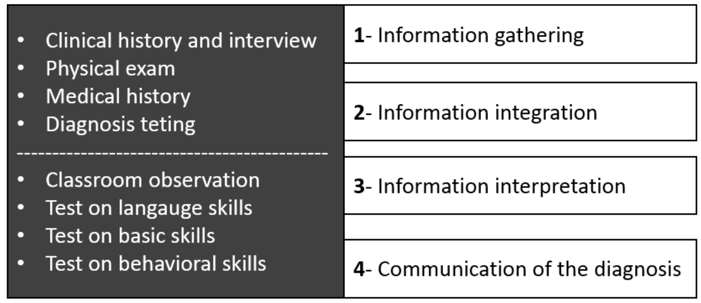


Fig. 4. Prior medical and psychomotor diagnosis

named NEOCAMPUS2 for the development of ILS, the latter is characterized by the co-existence of pedagogical, cognitive and affective agents.

In addition, the intelligent management of knowledge on the different cognitive, affective and pedagogical levels will make it possible to deduce the learning style of the learner. Providing learners with learning resources and activities tailored to their preferences and learning styles increases their satisfaction and improves learning performance.

Theoretically, there are several models of learning styles [25]; [26]. Extraction and follow-up of the learners learning style can be done through neural networks [27]. According to Simon[28], the neural network is defined asa massively parallel distributed processor consisting of simple processing units, which has a natural propensity to store experiential knowledge and make it usable.

The proposed methodology is diagnostic centered because it exploits the knowledge derived from a diagnostic model established by a clinical team. Although diagnosing a learning disability is not always easy, even if the symptoms seem clear. It is important that the child is tested and evaluated by qualified professionals. Recognizing that the diagnostic criteria for identifying different learning disabilities have been established by the American Psychiatric Association (DSM-IV-TR, 2000) and the World Health Organization (ICD-10, 2003). These criteria are internationally recognized by the scientific and medical authorities.

The diagnosis may sound like a medical term; and, in some ways, it is. It is primarily administered by a health professional and is often based on the medical symptoms of a child. Sometimes, several professionals coordinate team services to obtain an accurate diagnosis. They can ask for comments from the childs teachers and parents.

To carry out a diagnosis of learning disabilities, we must start from the analysis of basic skills, which are the basic skills necessary for the child to build and perform the correct behaviors of learning to read, write and calculate. Without the proper organization of these skills in terms of perceptual, linguistic and psychomotor expression, the child will not attain the development that allows him to acquire the knowledge and skills of the different learning.

For this, the diagnostic model includes 5 specifications (see Fig. 5) from a medical and technical diagnosis of high level, its aspects are as follows:

A. Motor Specifications:

Motor impairment can also occur for some children with LD and influence their academic performance and activities in their daily lives. The following parameters are the subject of the model motricity diagnosis:

- Inability to coordinate both sides of the body
- Difficulty following directions
- Difficulty controlling pencils, scissors, or coloring in lines
- Difficulty with buttons, zippers, learn to tie shoes
- Bad writing
- Difficulty using clay to make balls
- Difficulty feeding yourself with spoon and fork
- Move your head
- Difficulty discriminating shapes, texture size only by touch

B. Behavioral specifications:

Children with LD report being more disruptive, less co-operative, insensitive, and more tactical in school or family environments, always seeking to attract attention, indicating a disruption in behavior. The diagnostic model includes the behavioral indicators below:

- Clownerie
- Complete the tasks as quickly as possible "just to finish"
- Reject, leave an activity or task
- Not wanting to go to school
- To complain about the teacher
- To say that the work is too hard
- Say negative things about your academic abilities such as: "I'm stupid"
- Refuse to work at school or to resist homework
- Refuse to follow the instructions of the teacher to be sent from the room
- Use other delay tactics to avoid a mission
- Change places frequently

C. Cognitive specifications:

Historically, constructivist theories originated with the work of Jean Piaget, a researcher who has studied how children's intelligence progresses with their age, and how they progress intellectually into adulthood. Even if it is strongly criticized, Piaget's theory still remains a reference for the first apprenticeships. For the diagnostic model, we are interested in two major aspects of the child's cognitive development which are the spatio-temporal aspect and the pedagogical aspect. *Spatio-temporal deficit:*

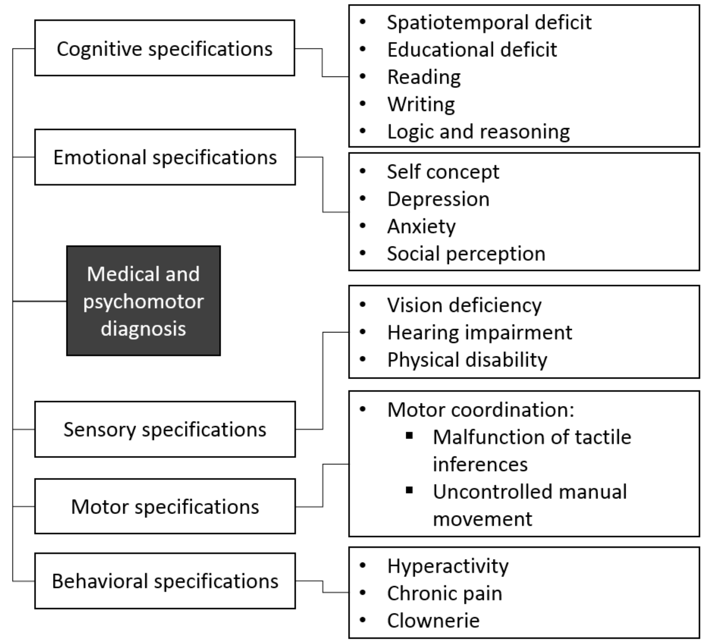


Fig. 5. Diagnostic model specifications

When there is a kind of anomaly in visual or auditory perception and spatio-temporal structuring, this manifests itself in the acquisition and expression of the signs used in basic learning, so it is essential to analyze the development and operation of these levels.

Educational deficit:

A psycho-educational battery of formal and informal tests is used to determine patterns of strength and weakness in relation to intellectual ability. This battery will provide answers to the following:

- Reading and listening:
  - Reading aloud
  - The understanding of certain words
  - The ability to understand what is said
  - The ability to follow a short story and memorize it
  - Is the child able to answer an interview?
- Spelling and writing:
  - Writing ability
  - The writing order
  - Capitalization and punctuation
- Math and logic:
  - The ability to differentiate symbols and operations
  - Does the child consider the number and its inverse in the same way?
  - The ability to count coins, bills and give change
  - Is the child able to find the duration, start time and end time?

D. Socio-emotional specifications:

Although communication plays a preponderant role in the social life of the child because it is a valuable tool for transmitting values and bases. Children with learning disabilities find it difficult to interpret social cues which influence their ability to adapt their behaviors or even exhibit inappropriate social behaviors.

As a result, negative interactions with others have been linked to the neglect of subtle social cues or a lack of ability to accurately perceive and read social cues.

In addition, the negative affect and emotional regulation problems that often occur in children with LD are likely to deflect perceptions and interpretations of others' behaviors toward them. Moreover, the diagnostic model covers the following socio-emotional elements:

- The concept of self in children (is it negative?)
- Anxiety
- Depression
- Loneliness
- How does the child perceive:
  - His intellectual abilities?
  - His school status?
  - His popularity
- Does the child have feelings of:
  - Insecurity?
  - Inadequacy?
  - Guilt?
  - Impotence?
  - Vulnerability?
- Does the child change mood in a remarkable way?
- How does he react to changes?

E. Sensory specifications:

Children with learning disabilities often experience impaired color vision. As a result, many institutional situations require a soothing environment. An experiment at the University of Alberta proves the impact of environmental colors on the behavior and blood pressure of 14 children with disabilities. Hence, color psychology has a significant impact on the learning abilities and behavior of a child with LD. Failure in auditory processing is often cited as a major or contributing cause of language and learning disabilities in children. Some indicators of hearing loss:

- Distracted by noise
- It's hard to stay focused or to remember a verbal presentation
- Can misinterpret or have difficulty remembering oral instructions; difficulty following directions in a series
- Say "what?" Many, even when heard a lot of what was said

In addition, the impact of emotion on the learning process of the child with LD cannot be a negligible factor because

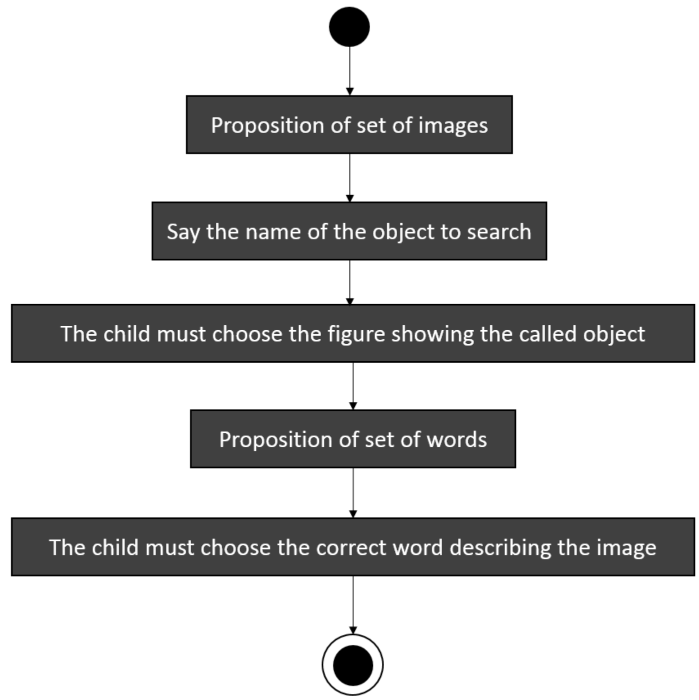


Fig. 6. Educational activity proposed by the pedagogue expert for a dyslexic child

emotion can play the role of leverage because it can be an obstacle to change. acquisition of a new knowledge. Strengthening the emotional state of the child and correcting the abnormalities that he / she experiences will help to improve his / her perception of learning and, as a result, promote learning.

VII. Case Study: DYSLEXIA

In this part, we develop all the deliverables of the KASP methodology for the case of a child with dyslexia disorder. Knowing that dyslexia is a lasting disorder of written language affecting reading, spelling and writing. These disorders are distinguished from a simple delay of acquisition, a mental retardation, a hearing problem, a visual problem, an emotional problem, a speech problem, bilingualism. The troubles persist in time, thats why we talk about lasting trouble.

The Case Study Diagnostic Worksheet contains the five specifications, indicating their severity and frequency on a scale of 0 to 3 as exposed in the Table. 1 :

TABLE I. Severity and Frequency Levels

Severity	
	0 does not exist 1 soft 2 moderate 3 severe
Frequency	
	0 never have the symptom 1 have the symptom from time to time 2 have the symptom often 3 have the symptom all the time

After having the diagnosis model worksheet presented in Table. 2, experts in cognition, pedagogy, hearing and vision intervene to develop the conceptual architecture.

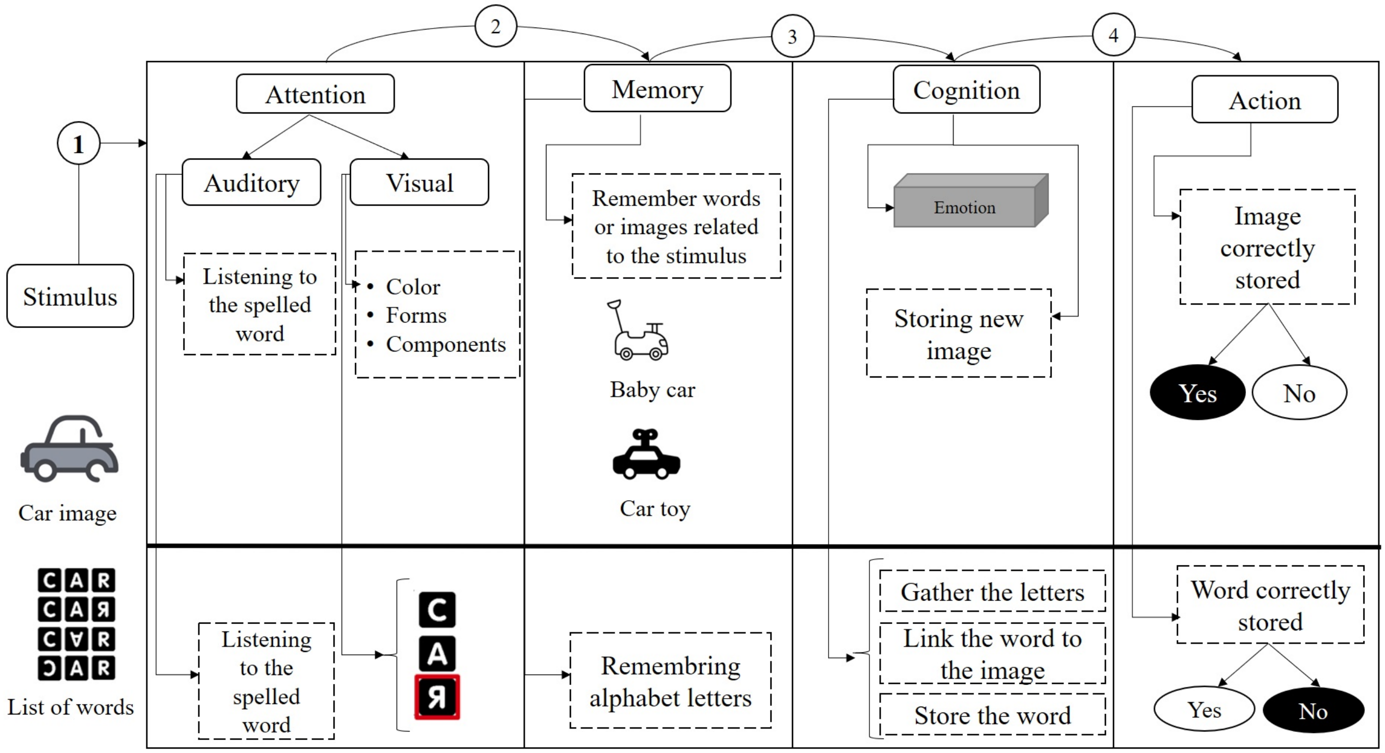


Fig. 7. Mental model of a dyslexic child

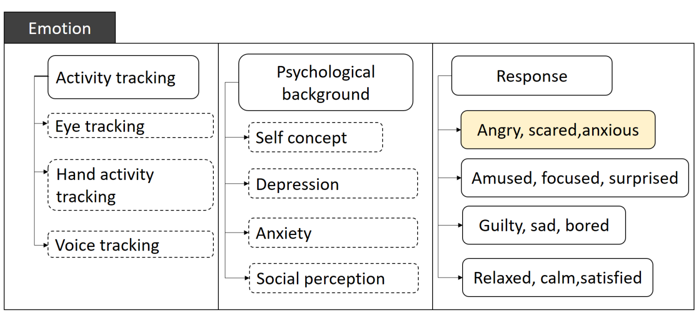


Fig. 8. Emotional Mental Model of a Dyslexic Child

TABLE II. DYSLEXIA CASE STUDY DIAGNOSTIC WORKSHEET

	Symptoms	Severity	Frequency	Notes
Cognitive specifications				
	- Spatiotemporal deficit	2	2	
	- Pedagogical deficit			
	* Reading	3	3	
	* Writing	3	3	
	* Logic and reasoning	2	1	
Motor specifications				
	- Motor coordination	2	2	
	- Mulfunction of tactile inferences	1	2	
	- Uncontrolled manual movement	1	2	
Emotional specifications				
	- Bad self-concept	3	3	
	- Depression	1	1	
	- Anxiety	2	2	
	- Social perception	3	3	
Behavioral specifications				
	- Hyperactivity	3	3	
	- Chronic pain	1	2	
	- Clownerie	2	3	
Sensory specifications				
	- Hearing impairment	2	2	
	- Vision deficiency	1	1	

The cognition expert analyzes the diagnostic model in order to build the mental models of the dyslexic child, the latter having a score of 3 at the level of severity of the writing deficit can be translated by a frequent permutation of letters during the writing. So the pedagogue expert proposes a game scenario to correct this deficit, the scenario is as illustrated in Fig. 6.

Here is an example of a mental model in Fig. 7 developed by the cognitive expert illustrating the deficit related to the pedagogy where the child permutes the letters, the mental model includes in itself the affective one because the emotion presents itself at the level of the cognitive treatment.

So the expert in child psychology studies the diagnostic record in order to build the emotional pattern of the child and it is presented in Fig. 8.

The proposed methodology is iterative trying to formalize a pragmatic and manageable approach, it helps to help children with learning disabilities to overcome these deficits by highlighting their emotional state that plays a key role in the process of learning. the child while integrating a playful environment generating entertainment.

VIII. DISCUSSION

Indeed, the methodology of designing serious learning games for the benefit of children with learning disabilities combines the conceptual and software processes, the novelty associated with this proposal lies in taking into consideration the emotional state of the child.

This can induce various emotional experiences in learners. Therefore, emotional influences need to be carefully considered in the design of educational courses in order to maximize learner engagement as well as improve learning and long-term retention of materials [29]. Numerous studies have reported that human cognitive processes are affected by emotions, including attention [30], learning and memory [31], reasoning [32] and problem solving [33]. Abrams [34] said,The vast majority of children with learning disabilities have an emotional problem associated with learning difficulty. Therefore, the proposed methodology gives the same level of importance to the knowledge as the emotion and it's a very robust approach because it implements the most rigorous teaching practices.

The KASP methodology proposes a design environment ensuring a good collaboration between the various stakeholders, including a well-detailed modeling of the pedagogical structure allowing a better choice of pedagogical activities.

Considering hearing ,sight and kinesthetic parameters helps enabling better interactive communication between the learner and the proposed game scene.

Moreover its integration of the emotional factor favors a better regulation of the emotional state of the learner and it guarantees intrinsic motivation.

It begins by defining the needs of the target audience, in other words the type of disorder to be treated, then it goes on to the conception of the cognitive and affective mental models as well as the sensory specifications in order to propose the appropriate pedagogical scenarios.

The development of the prototypes, which will be tested and evaluated in the next step. And finally, the implementation of the game through the deployment of prototype results.

In fact, the deployed serious learning game, promoting the involvement of the player-learner and a good connection between educational and fun elements, respects the four dimensions:

- Fun: its about fun, the taste of competition, creating commitment to efficiency and productivity.
- Epistemic: Where knowledge is the object of the game
- Ergonomic: presenting seamless, easy to use, flexible and interactive interfaces
- Contextual: based on the objectives designated by the experts to meet the needs initially expressed

Contrariwise,the KASP methodology lacks some performance measures regarding the validity of the mental models of knowledge and emotions made by the experts. Moreover, it depends closely on the human experts because thanks to them that the raw material is elaborated for generating the educational scenarios.

IX. Perspectives and Future Work

Serious learning games have shown great potential as an effective training methodology, but such a technique involves an extremely complex development process, with few guidelines available for effective evaluation.

The authors suggest using a fuzzy logic approach to establish an efficient evaluation of the KASP-based learning system, such as that proposed by El Alami and de Arriaga [35]. It is a fuzzy assessment of computing emotional and cognitive in intelligent e-learning systems that encompasses a set of fuzzy techniques acting on several levels especially the cycle of methodology as well as development and maintenance of the system designed.

X. Conclusion

In conclusion, the design of a serious learning game that helps children with LD in their learning curriculum is the subject of this article. Having a diagnostic model specifying the key parameters to identify the type of disorder is therefore a crucial step that obviously precedes the game design process.

In addition, exploiting the potential of gamification to improve learner engagement and increase motivation is the central idea of this article. Furthermore, understanding the emotion of the learner having a LD throughout the learning process through the techniques of affective computing will help to correct imperfections not only at the knowledge level but also at the socio-emotional level.

References

[1] D. Michael and D. Chen, *Serious games*, Mason, Ohio: Course Technology, p.17, 2011

[2] J. Alvarez, D .Djaouti and O. Rampnoux, *Introduction au serious game*, 2nd ed. Paris: Questions thoriques, 2012.

[3] I. Marfisi-Schottman, S. George and F. Tarpin-Bernard, *Tools and Methods for Efficiently Designing Serious Games*, In: 4th Europeen Conference on Games Based Learning ECGBL, 2010

[4] M. Immordino-Yang and A. Damasio, *We Feel, Therefore We Learn: The Relevance of Affective and Social Neuroscience to Education*, Mind, Brain, and Education, 1(1), pp.3-10, 2007

[5] P. Schutz, J. Hong, D. Cross and J. Osbon, *Reflections on Investigating Emotion in Educational Activity Settings*, Educational Psychology Review, 18(4), pp.343-360, 2006

[6] T. Hascher, *Learning and Emotion: Perspectives for Theory and Research*, European Educational Research Journal, 9(1), pp.13-28, 2010

[7] L. Gaussot, *Le jeu de l'enfant et la construction sociale de la ralit*, Le Carnet PSY, 62(2), p.22, 2001

[8] K. Hirsh-Pasek and R. Michnick Golinkoff, *Play — Why Play = Learning — Encyclopedia on Early Childhood Development* Available at: http://www.child-encyclopedia.com/play/according-experts/why-play-learning, 2008

[9] A. Abdul Jabbar and P. Felicia, *Gameplay Engagement and Learning in Game-Based Learning*, Review of Educational Research, 85(4), pp.740-779, 2015

[10] C. Abt, *Serious games*, Lanham: University Press of America, 1987

[11] S. Egenfeldt-Nielsen, B. Meyer and B. Srensen, *Serious Games in Education*, Santa Barbara: Aarhus University Press, 2011

[12] K. Salen and E. Zimmerman, *Rules of play*, Johanneshov: TPB, 2011

[13] T. Lloyd Buck, *Living the Case Study: Teaching Management and Leadership Ethics Online Through Serious Games*, In: Society for the Advancement of Information Systems, 2014

[14] G. Paquette, F. Crevier and C. Aubin, *Mthode d'ingnierie d'un systme d'apprentissage (MISA)*, Revue informations in Cognito, 8, 1997

[15] V. Lavergne Boudier and Y. Dambach, *Serious game*, Paris: Hermes Science, 2010

[16] M. Mcmahon, *The DODDEL Model*, Games-Based Learning Advancements for Multi-Sensory Human Computer Interfaces, 98-118, 2009

[17] J.W. Keefe, *Learning style: An overview*, NASSP's Student learning styles: Diagnosing and proscribing programs (pp. 1-17). Reston, VA. National Association of Secondary School Principles, 1979

[18] K. Rapeepisarn, K.W. Wong, C.C. Fung and M.S. Khine, *The Relationship between Game Genres, Learning Techniques and Learning Styles in Educational Computer Games*, Technologies for E-Learning and Digital Entertainment Lecture Notes in Computer Science, 497-508, 2008

[19] M. El Alami, N. Tahiri and F. de Arriaga, *DIAUTIS: A Fuzzy and Affective Multi-agent Platform for the Diagnosis of Autism*, British Journal of Applied Science & Technology, 21(4), 1-28, 2017

[20] S.S. Haykin, *Neural networks and learning machines*, Delhi: Pearson, 2016

[21] J.F. Sowa, *Semantic Networks*, In Encyclopedia of Cognitive Science, L. Nadel (Ed.), 2006

[22] R.W. Picard, *Affective computing*, Cambridge, MA: MIT Press, 2000

[23] S. Gano-Phillips, *Affective learning in general education*, Special Topic: Assessment in University General Education Program 6(1), 144, 2009

[24] F. de Arriaga, A. Arriaga and M. El Alami, *Intelligent E-Learning Systems: Affective and Cognitive Computing*, In C. Montgomerie & J. Seale (Eds.), Proceedings of ED-MEDIA 2007–World Conference on Educational Multimedia, Hypermedia & Telecommunications (pp. 1127-1137). Vancouver, Canada: Association for the Advancement of Computing in Education (AACE), 2007

[25] R. Felder and L. Silverman, *Learning and teaching styles in engineering education*, Journal of Engineering Education, 78, 7, 674-681, 1988

[26] A.Y. Kolb and D.A. Kolb, *Learning Styles and Learning Spaces: Enhancing Experiential Learning in Higher Education*, Academy of Management Learning & Education, 4(2), 193-212, 2005

[27] A. Ugena, F. de Arriaga and M. El Alami, *Neural Network Architectures: New Strategies For Real Time Problems*, Computing and Information Technologies, 2001

[28] H. Simon, *Neural networks: A comprehensive foundation*, New Delhi: Prentice-Hall of India, 2008

[29] L. Shen, M. Wang and R. Shen, *Affective e-Learning: Using "Emotional" Data to Improve Learning in Pervasive Learning Environment*, Educational Technology & Society, 12 (2), 176189, 2009

[30] P. Vuilleumier, *How brains beware: Neural mechanisms of emotional attention*, Trends in Cognitive Sciences, 9(12), 585-594, 2005

[31] E.A. Phelps, *Human emotion and memory: Interactions of the amygdala and hippocampal complex*, Current Opinion in Neurobiology, 14(2), 198-202, 2004

[32] N. Jung, C. Wranke, K. Hamburger and M. Knauff, *How emotions affect logical reasoning: Evidence from experiments with mood-manipulated participants, spider phobics, and people with exam anxiety*, Frontiers in Psychology, 5, 2014

[33] A. Isen, K. Daubman and G. Nowicki, *Positive affect facilitates creative problem solving*, Journal of Personality and Social Psychology, 52(6), pp.1122-1131, 1987

[34] J.C. Abrams, *On Learning Disabilities: Affective Considerations*, Journal of Reading, Writing, and Learning Disabilities International, 2(3), 189-196, 1986

[35] M. El Alami and F. de Arriaga, *Fuzzy Assessment for Affective and Cognitive Computing in Intelligent E-learning Systems*, International Journal of Computer Applications, 100(10), 40-46, 2014

Video Authentication using PLEXUS Method

Dr. Hala Bahjat Abdulwahab[1]
Computer Science Department
University of Technology
Baghdad, Iraq

Khaldoun L. Hameed[2]
Ibn Sina University for Medical and Pharmaceutical Sciences
Baghdad, Iraq

Nawaf Hazim Barnouti[3]
Al-Mansour University College
Baghdad, Iraq

***Abstract*—Digital Video authentication is very important issue in day to day life. A lot of devices have got the ability of recording or capturing digital videos and all these videos can be passed through the internet as well as many other non-secure channels. There is a problem of illegal updating or manipulation of digital video because of the development in video editing software. Therefore, video authentication techniques are required in order to ensure trustworthiness of the video. There are many techniques used to prevent this issue like Digital Signature and Watermarking, these solutions are successfully included in copyright purposes but it's still really difficult to implement in many other situations especially in video surveillance. In this paper, a new method called PLEXUS is proposed for digital video authentication on temporal attacks. In authentication process, the sender will generate a signature according to the method steps using a video and private key. In verification process, the receiver will also generate a signature using the same video and private key then each signature will be compared. If the two signatures are matched then the video is not tampered otherwise the video is tampered. This method is implemented using 10 different videos and proved to be an efficient method.**

Keywords—PLEXUS; video authentication; video tampering; temporal attacks

I. Introduction

Video authentication continues to be an important subject with significant attraction to researchers in last few years. By definition, Digital video authentication represents the technique of deciding whether the taken video is original and has not happened to be tampered with or not. Information can be transmitted quickly to thousands of kilometers with a few seconds. This will make a powerful influence on the growth and development in public. However, the significant improvement in information technology taken us to a new generation of effective information, it also has additionally added a few challenges related to information [1].

In digital generation, communication and compression methods help learning to share multimedia data including images and videos. Although, multimedia editing resources and tools may often use successfully modify the material of digital data, then straining the integrity of information [2]. The developing of computer systems and the equipment are making digital manipulation of video very simple and so easy to accomplish. Digital video trustworthiness and credibility has become really difficult because the copy of digital multimedia data acts similar to the original data.

A video might possibly be modified in a certain process to defame someone. In last few years, a number of situations are actually reported where some well-known individuals in the society was detected against the law in different actions in the video recordings done by some journalists. But unless they use reliable methods to find that the video is not easy to believe on these kinds of incidents. Meanwhile, criminals escape from getting arrested simply because this video that is used as facts against them indicating their crime will not completely confirmed in the court of law [1][3]. When it comes to surveillance systems, it is hard to guarantee that the digital video provided as facts, is just like it had been in fact recorded by the camera. For this reason, you can find an interesting require for video authentication.

Standard data authentication technology for message credibility is developed fully, but video authentication remains as early development step and several essential questions are still in mind. As an example, several different authentication methods designed during the last few years, it is difficult to acknowledge which method appear to be appropriate choice to make sure of credibility adapted to videos. There is certainly a good reason to use synthesizing literature to figure out the type of the condition, discover the probability of research difficulties, standardize different research subjects and evaluate the relative performances of the various methods [3].

II. The Need for Video Authentication

In certain products the credibility of video data is major concern including video surveillance, law enforcement, and forensic investigations. As an example, in court of law, trustworthiness must be established in this case for any video that is used as evidence. It is simple and easy to clear away a specific activity, individuals through removing some frames from video sequences. Meanwhile, it is simple to place some objects into the same video [4].

For that reason, video authentication is a technique that make sure that the information in the video is original and similar when taken. For confirming that video content is original, check whether the video was modified or not, and avoid different types of forgeries, video authentication methods are utilized [5].

These methods also detect and recognize the form of modification. In point of fact, several powerful processing tools are available for digital video. Considering that different video recording devices become much more convenient and reasonably priced choice in the private and public sections.

In criminal investigations, video evidences receive an important role as a result of their ability to acquire complete information and additionally have significant possibilities to help with investigations. Therefore, it will be important to obtain highest attention to ensure that the provided video evidence is original [3].

III. Video Authentication Applications

Digital video applications have a large number of advantages compared with custom analog video, better image quality, better color reproduction, and sharper images. Additionally, with the development in digital technologies, a video is often simple to carried over the internet and it produces easy editing and cropping [6].

Today the world is a video world from standard television broadcasting to modern communication media. In public individuals are more likely being video recorded. If you are walking on street, riding a city bus, entered a government building, etc. you probably being video recorded by a camera. Some civilians have setup mobile CCTV systems in their home and even their cars just in case of anything happen and they need to secure themselves or discover a crime that can happen any time. In fact, the police originally setup cruiser video recording systems to protect themselves and also to protect the citizens [6][7].

Forensic professionals have got several tools to decide scientifically whether the video is original or has been modified. In particular, it is difficult with digital video to discover which way a video was edited. This is where forensic investigation will become the only way to decide the video evidence credibility.

There are actually a number of situations in our everyday life exactly where video authentication is apparently necessary. In a situation of a well-known person was involved in illegal activities, it is a major interest to be able to determine whether or not the video was modified. In other situation, criminals can be set free simply because the video displaying their crime is not proved definitely in the court of law [1][7].

IV. Video Tampering

Video Tampering is a process of maliciously modification to the information material that going to be made by a video sequence. This process will be done for the purpose to hide an object or event. The seriousness and importance of video tampering depends on how and where these tampered videos have to be put to use. Since many advanced and low-cost video editing software tools are presented in the market which will make it an easy task to modify the video information material maliciously, it gives a significant challenge to researchers to be solved [8][9].

There are lots of potential attacks which will performed to modify the video data material. Whenever a risky modification is done on a video sequence, it either attacks on the contents of the video or attacks on the temporal dependency between the frames [10]. A large number of authentication methods are actually proposed but the majority of them are actually mainly dedicated to still images. However, the primary activity of video authentication system is to confirm whether or not the given video is tampered. Although, in a number of applications, because the ability to access information in video sequences, it will become better if the authentication system can find exactly in which part the modifications occurred and exactly how the video is tampered. Taking into consideration where and how, the video tampering attacks offers different classifications [11][12].

According to video sequences regional property, video tampering attacks can be classified to three basic types: spatial tampering, temporal tampering and the combination of these two, spatio-temporal tampering attacks. However, each category can even further be classified into their subcategories [10].

A. Spatial Tampering

In spatial tampering risky modifications are carried out on frames content. The operations which can be performed as spatial tampering are different methods to modify the frame just like copy, move, splicing, object adding and removing etc. Spatial tampering can be divided into three categories as shown in Figure 1. These kinds of attacks can be successfully performed by using video editing software [13][14].

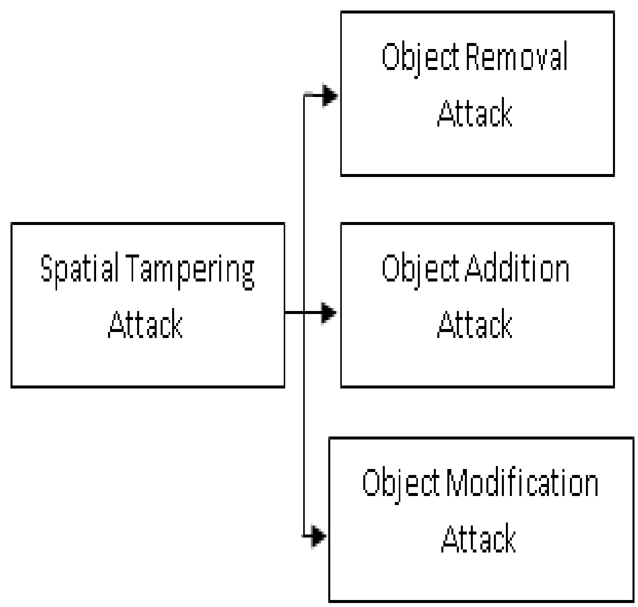


Fig. 1. Spatial Tampering Classification.

Object removal attack can be done with both foreground and background objects basically by hiding the occurrence of a person or object in a specific sequence of frames. Object addition attack can be done with both foreground and background objects basically by inserting any kind of objects in a frame or in a number of frames that belongs to a specific digital video that are available used as evidence fact. Additional object can easily paste in a frame or set of frames by using video editing software. Object modification attack can be done with both foreground and background objects basically by modifying any existing object of the frame in such a way that the actual identity of that object is misplaced, and a new object may occurred which is totally different from the original object [15][16].

B. Temporal Tampering

In temporal tampering malicious modifications is done on the video sequence of frames give attention to the temporal dependency. Time sequence of the frame that is recorded by a digital video camera generally influenced by Temporal tampering attacks [8][9]. Temporal tampering attacks can be divided into three types: frame addition, frame removal and frame shuffling as shown in Figure 2.

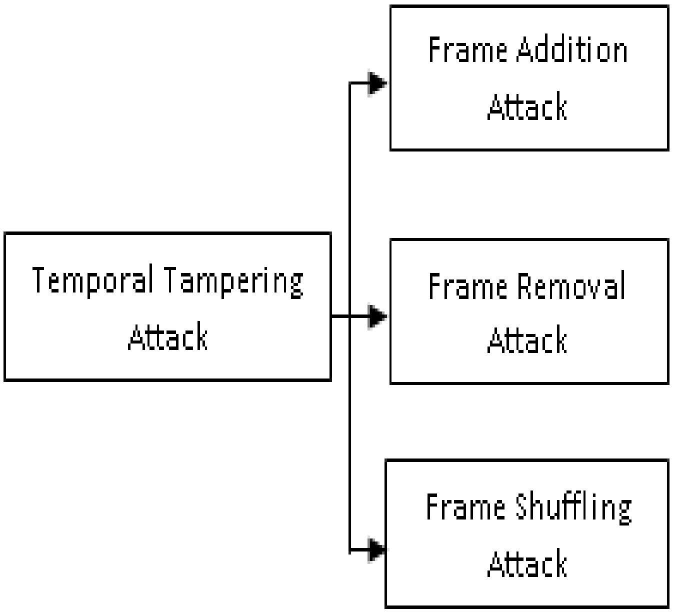


Fig. 2. Temporal Spatial Tampering Classification.

1) Frame Addition Attack

This kind of attack can be performed basically by adding additional frames from another video at some random locations in a video which has the same statistical properties as shown in Figure 3 [8][17].

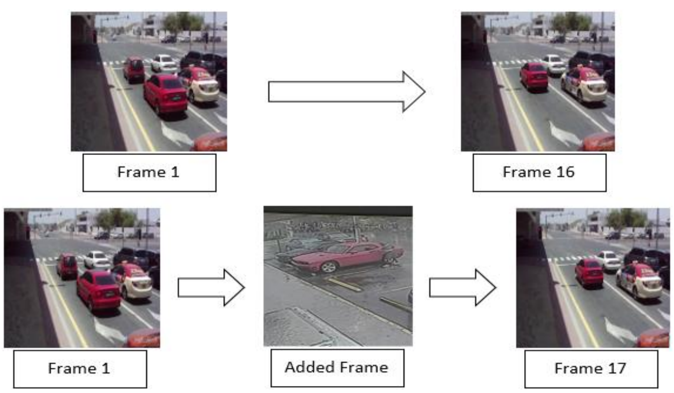


Fig. 3. frame addition attack example. In first row the original frame sequence from 1 to frame 16 has been shown. After attack, the second row of the frames shows the altered frame sequence in which a new frame is inserted between frame 1 and frame 16. And frame 16 become frame 17.

2) Frame Removal Attack

This kind of attack can be performed basically by removing one frame or a number of frames from the digital video at a certain location to a fixed location or removing set of frames from different locations as shown in Figure 4 [9][18].

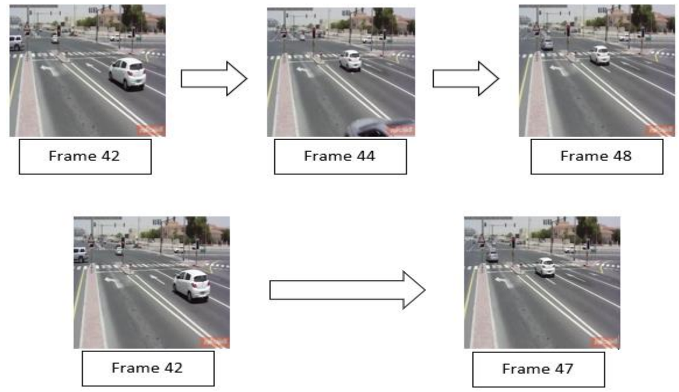


Fig. 4. frame removal attack example. In first row the original frame sequence with frame 42, frame 44, and frame 48. After attack, the second row of the frames shows the altered frame sequence with frame removal attack in which frame 44 is eliminated from the video and hence frame 48 become frame 47.

3) Frame Shuffling Attack

This kind of attack can be performed basically by shuffling the frame. The order of the frames will be changed and incorrect information is done by the digital video when compared with the original taken video as shown in Figure 5 [8][9].

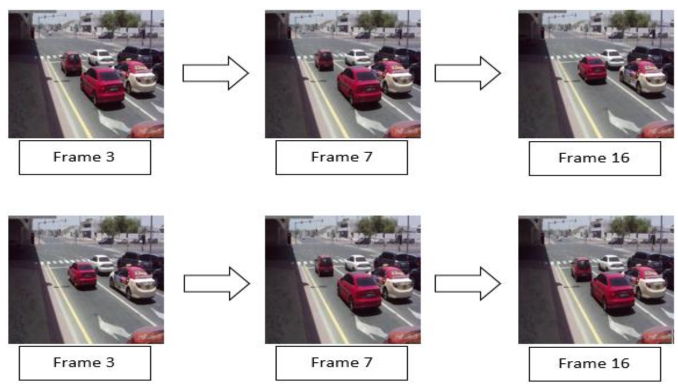


Fig. 5. Frame shuffling attack example. In first row the original frame sequence with frame 3, frame 7, and frame 16. After attack, the second row of the frames shows the altered frame sequence with frame shuffling attack in which the position of frame 3 and frame 16 have been changed.

V. Typical Video Authentication System

A video authentication system is consisting of two basic steps: Authentication process and verification process. Ideal and effective video authentication system have to follow the properties including sensitivity to changes, strength to benign operations, sensitivity against false alarm, self-recovery of modified regions, compactness of authentication data, localization and computational feasibility. In the authentication process, the authentication method processes the feature that taken out from the video and outputs the authentication data that is certainly encrypted by using the encryption key to form the signature [2][18].

In verification process, the video credibility is approved by determine the new authentication data through the use of the exact same authentication method. Then the new authentication data is compared with the original authentication data. The video will be treated as authentic if both authentication data are matched otherwise it is construed to be tampered [4][17].

VI. Previous Related Work

Researchers have done some work in the area of video authentication and several methods are presented on the subject of digital video authentication. The methods that have been done to match specific requirements but their techniques have one or more weak points. Basically, two techniques have already been implemented: digital signature and digital watermarking. This paper focus on the digital signature video authentication.

In [11] proposed a new method for digital video authentication that depends on video statistical local information. In this method, SVM (Support Vector Machine) classifier has been not used. However, this method approved to be efficient and trusted. The proposed method was evaluated on the dataset of videos, eight different attacks in each video were completely inserted and the method can successfully detect the attacks with overall classification accuracy 96.77%.

In [12] proposed a signature-based video authentication method to improve the digital video authentication in surveillance system using histogram of oriented gradient of the selected DCT (discrete cosine transform) coefficients in three dimensions. In this method, the result depends on optimal threshold that need a high threshold to ignore all tampered. The experiment results show that video with modification is ignored when using high threshold.

In [13] presents an algorithm which helps to determine whether the video is tampered or not. The algorithm is divided in two steps: computing the repeated frames and computing the tampering attack. Local information is determined and SVM classifier is successfully applied to classify whether the digital video is tampered or not.

VII. The Proposed Plexus Method

In this work, a new method called PLEXUS is proposed to address digital video authentication challenges and proved to be an efficient and high accuracy method. In the authentication part that done by the sender, the method generates a signature by multiplying first frame with the private key $f_1 * K$ to generate a new authentication image i_1 then multiplying second frame with the private key and the authentication image $f_2 * K * i_1$ to generate a new authentication image i_2 and this process will continue until reaching the last frame f_n and the signature can be generated by multiplying the last frame with the private key and the last authentication image $f_n * K * i_{n-1}$.

The embedding process between the frames, private key, and authentication images actually done between the three images by adding each color pixel-by-pixel and then divide the result by three which represent the number of images except first stage when the first frame only adding with private key and then divide the result by two. Embedding process diagram between first frame and private key is shown in Figure 6. Embedding process diagram between second frame, private key, and authentication image is shown in Figure 7

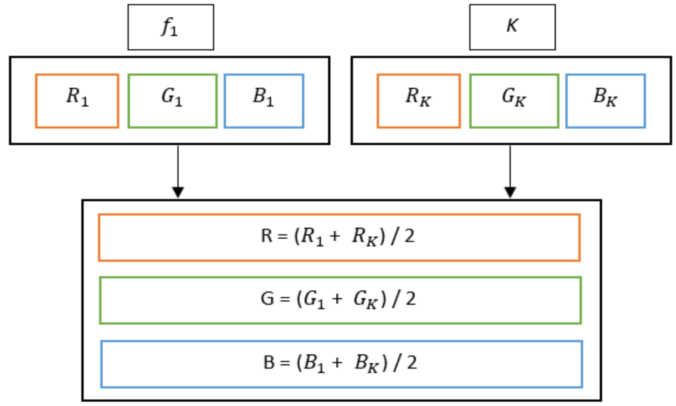


Fig. 6. Multiplication Process between First Frame and Private Key.

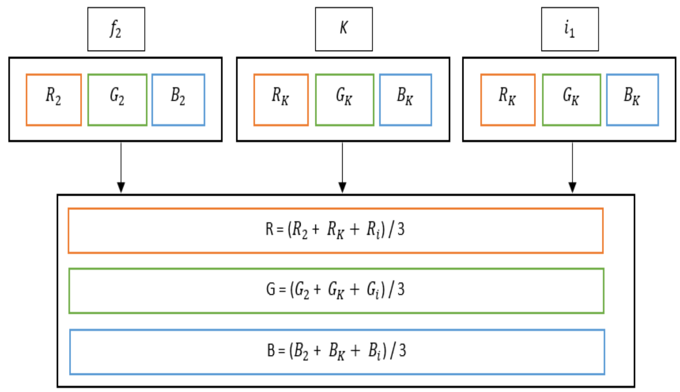


Fig. 7. Multiplication Process between Second Frame, Private Key, and Authentication Image.

In RGB color model, each pixel contains three colors which represent Red, Green, and Blue. The PLEXING process is the multiplication between these colors. This process will be divided into three steps: (1) Red color remains unchanged and the multiplication process will be done between green and blue colors (GB). The pixel three colors will be R GB GB. (2) Green color remains unchanged and the multiplication process will be done between red and blue colors (RB). The pixel three colors will be RB G RB. (3) Blue color remains unchanged and the multiplication process will be done between red and green colors (RB). The pixel three colors will be RG RG B. These steps will be applied to each pixel in the image while accessing the last pixel.

In verification part that done by the receiver, this process will be repeated to generate a signature. Each signature will be compared using image quality similarity measurement to determine whether the video is tampered or not. The authentication accuracy will be higher rate when increasing the number of frames. The overall PLEXUS method diagram is shown in Figure 8.

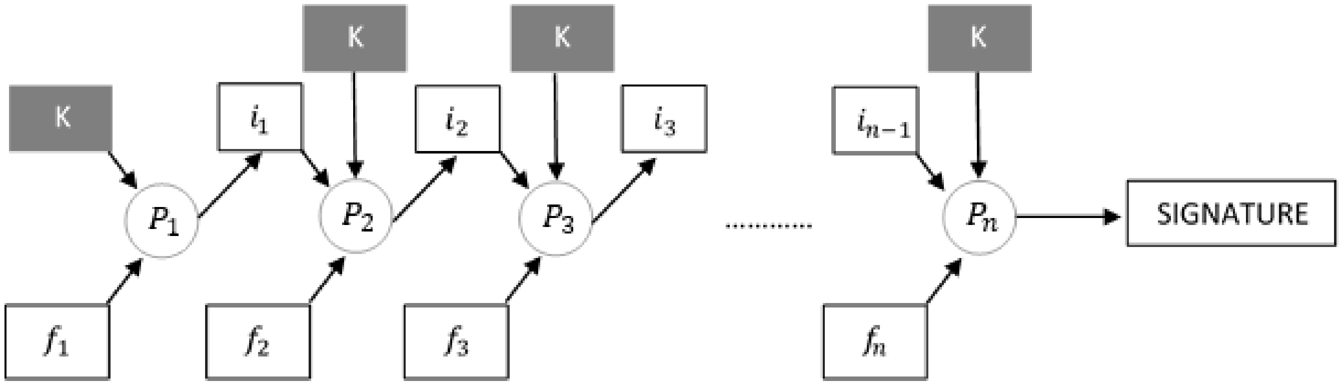


Fig. 8. Overall PLEXUS Method Diagram.

A. Frame Quality Measurement

The digital image can be affected by different types of distortions when passing through several processing stages. Image processing stages can lead to important loss of information or quality. Different metrics are utilized to estimate digital video quality. In image quality evaluation there are basically two methods: subjective and the objective methods. The subjective method quality evaluation being considered as time consuming because it is depending on human evaluation and work without references to specific considerations. The objective quality evaluation takes advantage of automatic algorithms to determine the quality of the image without human interfere [6][14].

The most popular are the objective methods: PSNR (peak signal-to-noise ratio) and MSE (mean-squared-error). The two measurements are based on pixel-by-pixel comparison and its parameters are frequently used for simple identification, but they do not reflect the perceptions of the recipient. To reach the best quality, PSNR should be the biggest and MSE should be the smallest [14][19].

1) MSE – Mean Squared Error

MSE is objective method represent the average of the squared differences between the luminance values of corresponding pixels in two different frames. It is possible to evaluate the degree of image reconstruction by a decoder and do not consider any peculiarity of HVS (human visual system) [14]. MSE is definitely non-negative, and should be as small as possible. Given a noise free $m * n$ monochrome image I and its noisy approximation I'. MSE mathematical representation can be shown in the equation bellow:

$$MSE = \frac{1}{M * N} \sum_{i=1}^{M} \sum_{j=1}^{N} ([I(i,j) - I'(i,j)])^2 \qquad (1)$$

2) SNR (Signal-to-Noise-Ratio) AND PSNR (Peak Signal-to-Noise-Ratio)

SNR measure used in science and engineering that compares the level of a desired signal to the level of background noise. SNR ratio is defined as the ratio of signal power to the noise power, often expressed in decibels. A ratio higher than 1:1 (greater than 0 dB) indicates more signal than noise [14][19].

PSNR is objective method that measure image quality based on the pixel difference between two different images. It is the most commonly used measurement metric describes the ratio of peak to noise [19]. The SNR measure is an estimation of quality of reconstructed image as compared with original image. PSNR mathematical representation can be shown in the equation below:

$$PSNR = 10.log_{10}\left(\frac{MAX_I{}^2}{MSE}\right) \qquad (2)$$

Where, MAX_I is the maximum possible pixel value of the image.

VIII. Result and Discussion

This work is done using a laptop with Intel(R) Core (TM) i5, CPU 2.40 GHz, 8 GB RAM. Visual Studio Community 2017 with Visual Basic programming language.

First, we would take the input video (37 second length) and then extract two frames per second (each frame 450x450 pixels with PNG format) as shown in Figure 9.

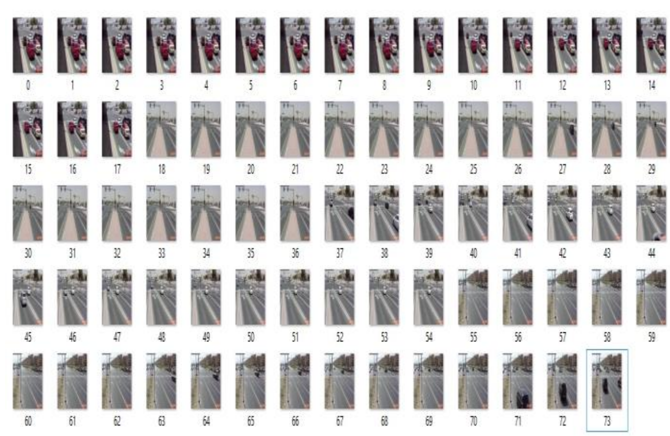


Fig. 9. Frame Extraction (74 Frames).

Second, the sender and receiver choose a private key. The private key is also an image (750x540 pixels with JPG format) as shown in Figure 10. Finally, PLEXUS method will work to generate a signature (sender signature) as shown in Figure 11.

The receiver should apply the previous steps for verification process. The receiver will generate a signature (receiver signature) and compare this signature with sender signature. The video is not tampered if both sender and receiver signatures are matched as shown in Figure 12. Otherwise, the video is tampered (Frame 70 is removed) as shown in Figure 13. This method achieves high accuracy and tested on 10 different digital videos. Then, this method evaluated using different evaluation measurements including: SNR and PNSR.

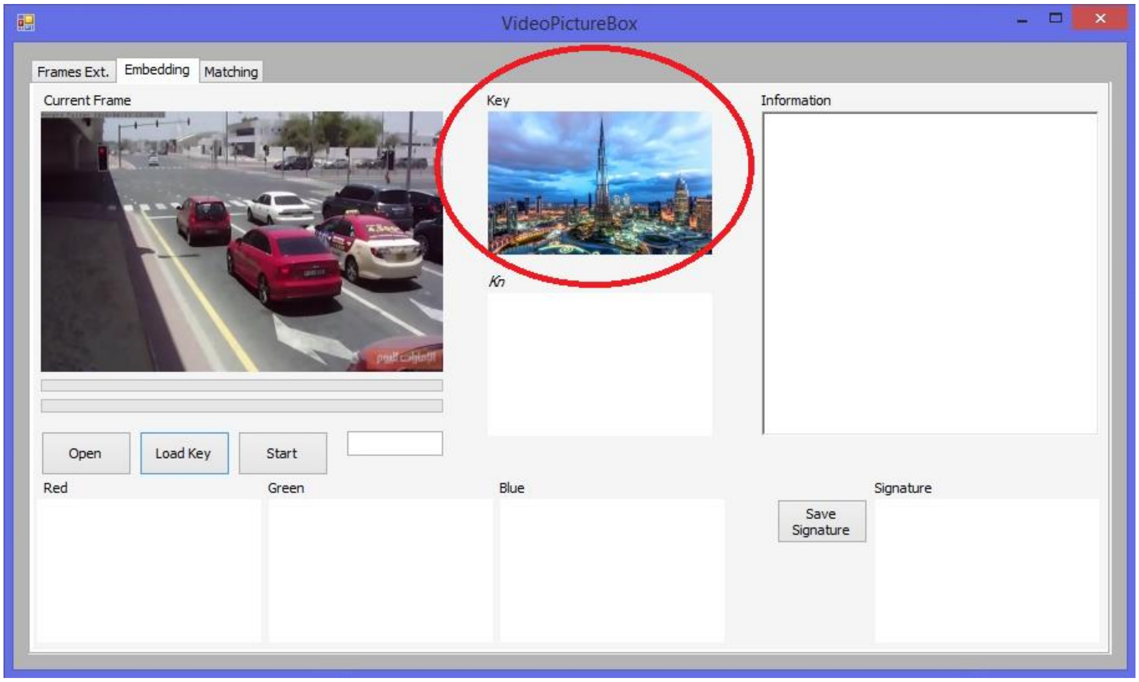


Fig. 10. Choosing Private Key.

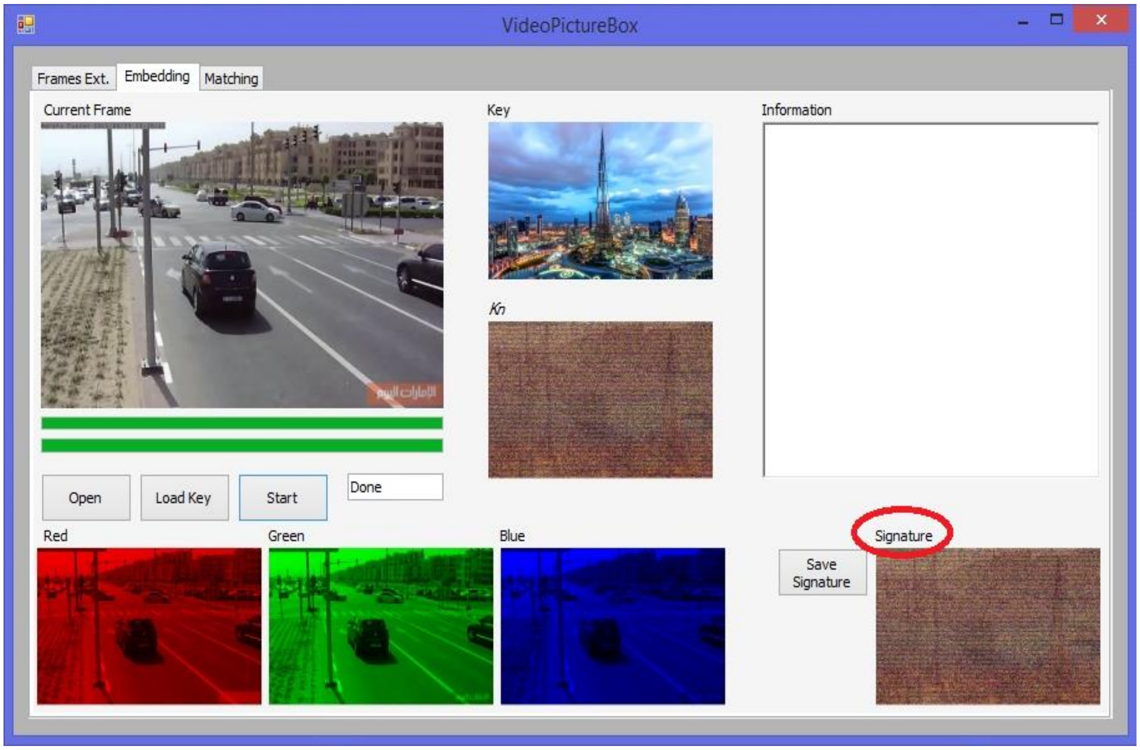


Fig. 11. Generate Sender Signature.

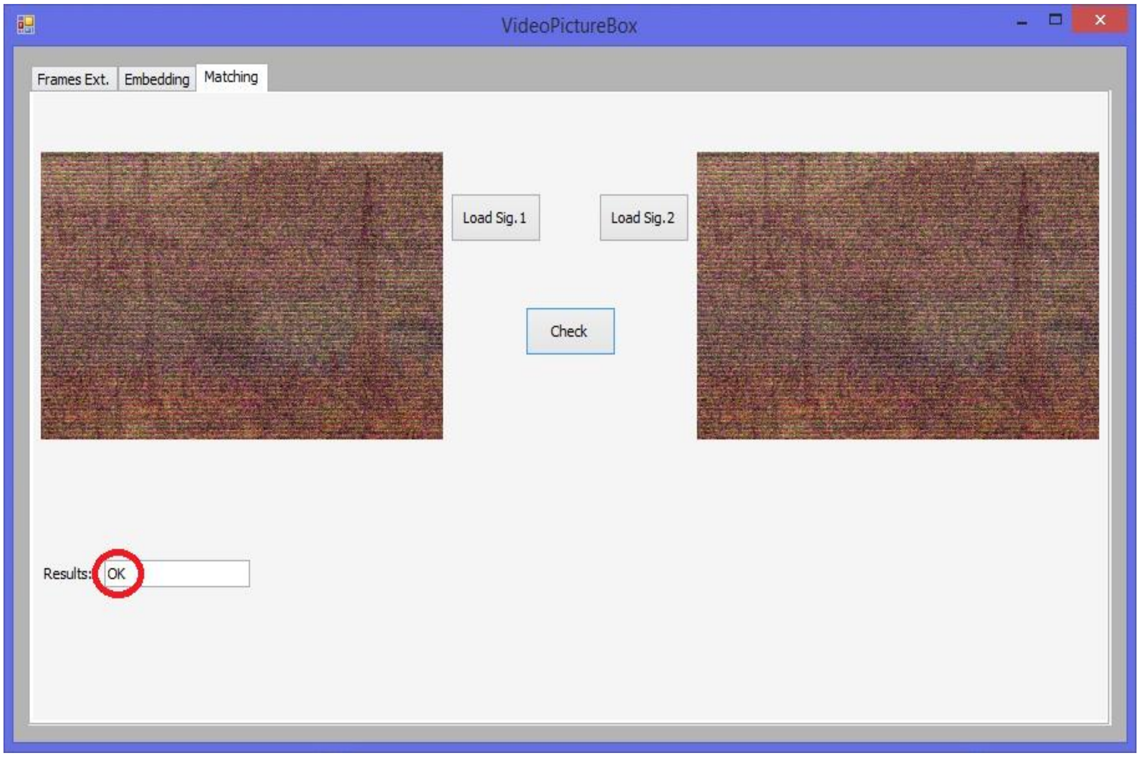


Fig. 12. Verification Process (Not Tampered Video).

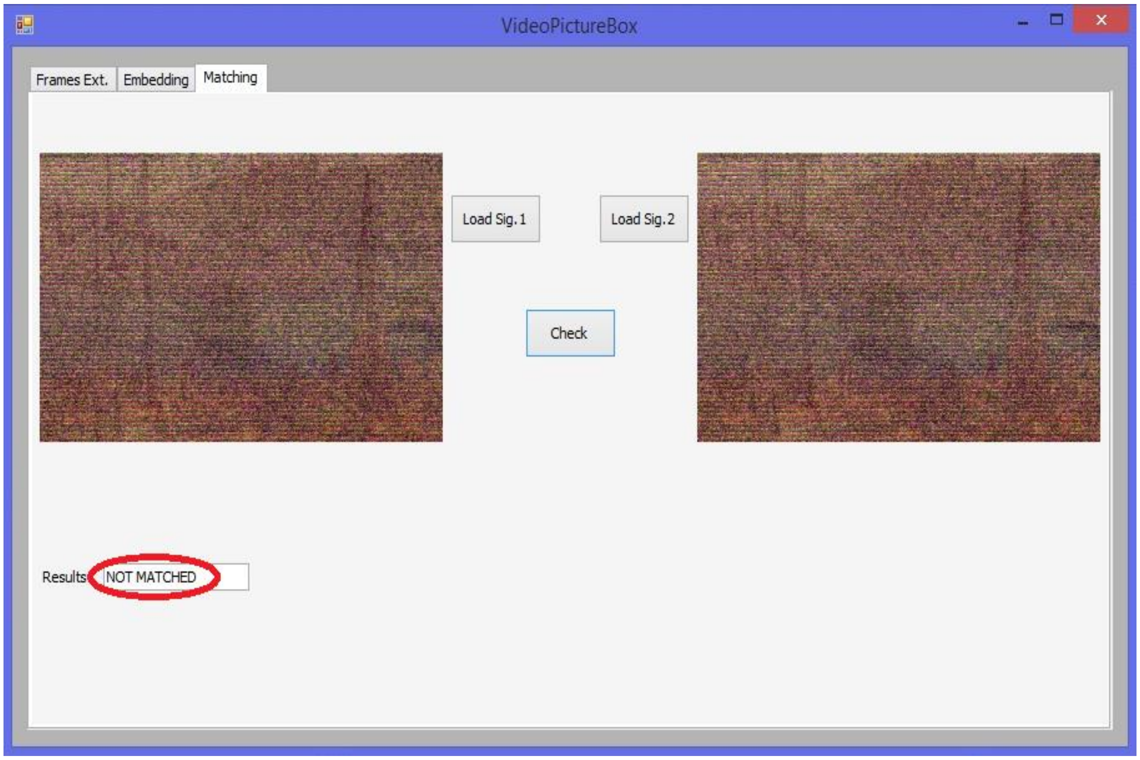


Fig. 13. Verification Process (Tampered Video).

The evaluation results including SNR and PSNR in the proposed method applied on two frames which are frame1 and signature as examples, because to find the final result we must apply this evaluation measurements on all video frames. The result of SNR= 3.4964 and PSNR= 8.9078.

IX. CONCLUSION

Video authentication is certainly really challenging issue in computer science area and very important subject in several applications specially with the growing of development tools which are available in video editing software. The digital video can be exposed to tampering attacks which means the video content is not trusted.

In this paper, we proposed a new video authentication method called PLEXUS to improve the reliability of digital video against temporal attacks. This method consists of two basic steps: authentication step and verification step. In each step a signature will be generated and then the two signatures will be compared and must be matched if the video is not tampered. This method is tested using 10 different videos and achieve high accuracy.

REFERENCES

[1] Sethulekshmi U. S, Remya R. S, Mili Rosline Mathews. "A Survey on Digital Video Authentication Methods." International Journal of Computer Trends and Technology (IJCTT) – volume 22 Number 1– April 2015

[2] Lo, Chun-Chi, and Yu-Chen Hu. "A novel reversible image authentication scheme for digital images." Signal processing 98 (2014): 174-185.

[3] Yeswanth.P, Pothumani.s, "Tampering Detection in Digital Videos." International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 3, March 2015.

[4] Atrey, Pradeep K., Wei-Qi Yan, and Mohan S. Kankanhalli. "A scalable signature scheme for video authentication." Multimedia Tools and Applications 34, no. 1 (2007): 107-135.

[5] Upadhyay, Saurabh, and Sanjay Kumar Singh. "Video authentication: Issues and challenges." International Journal of Computer Science 9, no. 2012 (2012).

[6] He, Dajun, Qibin Sun, and Qi Tian. "A secure and robust object-based video authentication system." EURASIP Journal on Advances in Signal Processing 2004, no. 14 (2004): 545469.

[7] Upadhyay, Saurabh, and Sanjay Kumar Singh. "Video authentication-an overview." International Journal of Computer Science and Engineering Survey 2, no. 4 (2011): 75.

[8] Christian, Aldrina, and Ravi Sheth. "Secured Digital Video Authentication System." (2017).

[9] Vartak, Reshma, and Smita Deshmukh. "Survey of Digital Image Authentication Techniques." International Journal of Research in Advent Technology 2, no. 7 (2014): 176-179.

[10] M. AL-ATHAMNEH, F. KURUGOLLU, D. CROOKES, M. FARID. "Video Authentication Based on Statistical Local Information." 2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing.

[11] Al-Athamneh, Mohammad, Fatih Kurugollu, Danny Crookes, and M. Farid. "Video authentication based on statistical local information." In Utility and Cloud Computing (UCC), 2016 IEEE/ACM 9th International Conference on, pp. 388-391. IEEE, 2016.

[12] Kroputaponchai, Teerasak, and Nikom Suvonvorn. "Video authentication using spatio-temporal signature for surveillance system." In Computer Science and Software Engineering (JCSSE), 2015 12th International Joint Conference on, pp. 24-29. IEEE, 2015.

[13] Gupta, Ankita, Shilpi Gupta, and Anu Mehra. "Video authentication in digital forensic." In Futuristic Trends on computational analysis and knowledge management (ABLAZE), 2015 International Conference on, pp. 659-663. IEEE, 2015.

[14] Komal, A., Shipra Khurana, and Amit Kumar. "Comparative analysis of image quality assessment using HVS model." IJIRCCE 2, no. 7 (2014): 5033-5038.

[15] Berge, L.A., Dangler, J.R., Doyle, M.S., Liang, T.W. and Orozco, M., International Business Machines Corporation, 2018. TAMPERING DETECTION FOR DIGITAL IMAGES. U.S. Patent Application 15/376,314.

[16] Sowmya, K. N., H. R. Chennamma, and Lalitha Rangarajan. "Video authentication using spatio temporal relationship for tampering detection." Journal of Information Security and Applications 41 (2018): 159-169.

[17] Sowmya, K. N., H. R. Chennamma, and Lalitha Rangarajan. "Video authentication using spatio temporal relationship for tampering detection." Journal of Information Security and Applications 41 (2018): 159-169.

[18] Gusev, Pavel D., and Georgii I. Borzunov. "The Analysis of Modern Methods for Video Authentication." Procedia Computer Science 123 (2018): 161-164.

[19] Mohsen, Asmaa Hasan, and Shaimaa Hameed Shaker. "Authentication of Digital Video Encryption." Iraqi Journal of Science 57, no. 4C (2016): 2954-2967.

An Automatic Cryptanalysis of Arabic Transposition Ciphers using Compression

Noor R. Al-Kazaz, William J. Teahan
School of Computer Science, Bangor University,
Dean Street, Bangor, Gwynedd
LL57 1UT, UK

***Abstract*—This paper introduces a compression-based method adapted for the automatic cryptanalysis of Arabic transposition ciphers. More specifically, this paper presents how a Prediction by Partial Matching ('PPM') compression scheme, a method that shows a high level of performance when applied to the different natural language processing tasks, can also be used for the automatic decryption of transposition ciphers for the Arabic language. Another well known compression scheme, Gzip, is also investigated in this paper with less efficient performance demonstrated by this method. In order to achieve readability, two further compression based approaches for space insertion are evaluated as well in this paper. The results of our experiments with 125 Arabic cryptograms of different lengths show that 97% of the cryptograms are successfully decrypted without any errors using PPM compression models. As well in a post-processing step, we can effectively segment the output that is produced by the automatic insertion of spaces resulting with only a few errors overall. As far as we know, this is the first work to demonstrate an effective automatic cryptanalysis for transposition ciphers in Arabic.**

Keywords*—*Automatic cryptanalysis; Arabic transposition ciphers; compression; PPM; word segmentation

I. Introduction

The Arabic language is one of the most widely spoken languages around the world, with as many as 300 million people in Asia and North Africa speaking Arabic. It is the fifth most spoken language [1]. The old manuscripts that have been recently discovered show that the origin of cryptology is much older than first thought, and Arab contributions to it are much more extensive than previously reported. These important findings are largely based on 8th century and later manuscripts written by the prominent Arab cryptologists such us al-Kindi. The newly discovered manuscripts push back the frontiers of the known history of cryptology by more than 500 years. The Arab scientist al-Kindi, the father of cryptology, is the author of the oldest book on cryptology [2].

Despite the large number of Arabic speakers and with Arabs being the first to use encryption systems as well as having success in breaking them, unfortunately encryption systems and cryptanalysis methods for Arabic are very few in comparison to other languages such as English. The purpose of this paper is to generate an Arabic transposition cipher and explore the use of compression models for the automatic cryptanalysis of Arabic text.

Text compression can be defined as a method of eliminating redundant information in a text aiming to minimize space that is required to store it, thereby reducing the time required to transfer this information without losing any information from the original text. There are two main approaches of constructing text compression models, which are dictionary and statistical approaches [3]. A well-known statistical based approach is Prediction by Partial Matching (PPM) while a well-known dictionary based approach is Gzip [4], which uses the Lempel-Ziv algorithm. PPM models can be applied to tasks with performance that emulates that of humans [5].

There are many different approaches used for cryptanalysis especially for the English language. However, using compression schemes adapted for the Arabic language as one way to tackle the Arabic plaintext recognition problem as we do in our paper is a novel approach and not tried before in the literature. The PPM compression system uses a modelling method that performs well at different language processing tasks and can be successfully applied to the automatic cryptanalysis of transposition ciphers when using the English language with a 100% success rate [6]. In this paper, a compression-based approach for the automatic cryptanalysis of transposition ciphers adapted for the Arabic language without any need for human intervention is proposed. In addition, a further algorithm is also introduced to automatically insert spaces into the decrypted texts in order to achieve readability.

Our paper is arranged as follows. Related work is presented in the next section followed by background to transposition ciphers. A description of compression based cryptanalysis and specific details of our method are provided in sections 4 and 5. The experimental results are reviewed in section 6 and the conclusions are provided in the last section.

II. Related Work

Many different cryptanalysis methods have been invented to break transposition ciphers for the English language, such as anagramming [7], using genetic algorithms [8], simulated annealing and Tabu search [9]. Alkazaz et al. [6] presented a new universal automatic cryptanalysis approach of transposition ciphers for the English language. A compression-based approach was used as a basis for breaking the cipher using the PPM and Gzip compression schemes. Ciphertexts with different sizes (from 12 to 600 letters) were tested. They managed to achieve a 100% decryption success rate by using PPM compression models, in addition to the high quality performance achieved in segmenting the decrypted texts.

Despite the fact that the science of cryptology was born among Arabs who also managed to break the mono-alphabetic substitution cipher after many years of its successful use [10], unfortunately only a few publications have investigated the use of the Arabic language for various encryption/decryption systems [11]–[14] and none for the problem of solving transposition ciphers.

In this paper, we further investigate the English-based work of Al-kazaz et al. [6] to examine if it is applicable to other languages and ascertain whether the Arabic language provides a greater challenge for decryption. We also choose this language specifically because it has characteristics that differentiate it from other languages and because Arabic is non-related to English. Specifically, it is written and read from right to left rather than vice versa. It consists of 28 consonant letters and the vowels are represented by marks below and above the letters. It also has distinctive variations to represent singular, dual and plural forms and to represent male and female forms. Written Arabic also often exhibits triglossia with classical, modern and mixed forms often appearing together. Each letter in the Arabic script denotes a unit of the language. Unlike the English language, there is no upper case or lower case in Arabic and each word comprising of more than one letter are written joined together (cursive writing), with an exception for some letters which are "ز ر ذ د", and "ا" if they come at the beginning. These characteristics, as well as its rich morphology [15], presents challenges for natural language processing and cryptographic systems.

In our approach, Arabic transposition ciphers are generated and then a modified compression based approach for the automatic cryptanalysis of Arabic transposition ciphers is proposed in this paper. Various 125 Arabic ciphertexts with different lengths, from very short ciphers (17 Arabic letters) to ciphers of length 800 letters, are then tested in our experiments described below.

III. Transposition Ciphers

This section will now review transposition ciphers specifically. In a transposition cipher, the content of a message is obscured by rearranging groups of symbols; a transposition is thus a permutation. The concept of transposition is an essential one and has been used in the design of modern ciphersystems [16]. Originally, the message was written into a matrix in row-major order and then read out in column major order. Encoding "أنظمةالتشفير", which means "encryption systems", using a 3×4 matrix is shown in Figure 1. This figure gives "أةشنافظليمتر" if the columns are read off in order.

م	ظ	ن	أ
ت	ل	ا	ة
ر	ي	ف	ش

Figure 1: Example of a matrix transposition cipher.

By reading the columns in a different order, say 3–2–1–4, different ciphertexts can be obtained, such as "ظلينافأةشمتر". The order to read the columns and the size of the matrix constitute the key. The technique can be extended to d dimensions. In transposition ciphers, a plaintext block is encrypted into a ciphertext block using a fixed permutation [17].

Transposition ciphers are a class of ciphers that in conjunction with substitution ciphers form the basis of all modern symmetric algorithms [17]. These algorithms, such as block and stream ciphers also use in conjunction to the above forms more complex transformations, notably for providing diffusion. Although the field of cryptology has undergone a revolution after the introduction of the asymmetric cryptographic cipher in 1976, symmetric ciphers still form the basis for secure data transmission today, because of their superior speed and efficiency [18]. Hence in this paper, we focus on a very basic building block that is inherently present in most symmetric cryptographic systems, namely transposition ciphers.

Transposition ciphers are generally considered as much more difficult to decrypt than the other basic ciphers such as simple substitution ciphers. A number of statistical tools have been developed aiding automated breaking of substitution ciphers while the automatic decrypting of transposition ciphers is notoriously difficult. In general, cryptanalysis of transpositions is highly interventionist and demands some knowledge of the probable contents of the ciphertext to give an insight into the order of rearrangement performed [6], [19].

IV. Compression as a Cryptanalysis Method

The ciphertext only cryptanalysis of simple cipher systems heavily depends on the statistical features of the source language, and it is not a trivial issue to get computers performing this analysis. Although computers have been routinely used for a variety of tasks in cryptanalysis since their invention, the automatic recognition of valid decryptions has been acknowledged as a taxing problem [20]. Several previously published cryptanalysis methods can not run without human intervention or they assume at least known plaintext because of the difficulty of quickly recognizing a valid decrypt in a ciphertext only attack [21], [22].

Having a computer model that is able to predict and model natural language as well as a human is critical for cryptology [5]. The idea of using text compression to break Arabic transposition ciphers stems from the observation that the best PPM models can predict language about as well as expert human subjects [5]. The main idea of our approach depends on using the PPM model as a method for computing the compression 'codelength', which is a measure of the information, for each possible permutation [20]. The smaller the codelength, the more closely the cryptogram resembles the model. In this paper, we have shown how to use this to quickly and automatically recognise the valid decrypt in a ciphertext only attack against transposition ciphers for the Arabic language. We have also examined the use of another well-known compression scheme, which is Gzip. In our experiments, we have determined that the PPM compression method is the most effective to use for the automatic cryptanalysis, with Gzip having significantly less success.

A. PPM Compression Codelength Metric for Arabic

In this section, we describe the PPM method that we devised for our cryptanalysis solution in order to encode Arabic text efficiently and how the codelength metric is calculated. In the PPM compression scheme, the previously transmitted symbols are used to condition the probability of the next symbol using a Markov-based approach. There are many variants of the basic approach depending on how many symbols are used to predict the next one, whether or not multiple predictions are used, and how shorter context models are used when necessary. The predictions are based on simple frequency counts of the transmission so far. The primary decision to be made is the context length modelled. This technique was first described by Cleary and Witten [23].

The order of a model is the maximum context length used to predict the next symbol. In practice, an order-o model will sometimes base its prediction on less than o symbols. This occurs when the model "escapes" down to the next lower order context if the higher order context has not seen the upcoming symbol that is being predicted. By convention the order(-1) model predicts each symbol with equal probability and the order-0 model predicts each symbol with probability proportional to the number of times it has occurred previously. An order 1 model predicts each symbol based on just the previous symbol, an order 2 model based on the two previous symbols and so on.

In 1990, Moffat [24] devised a simple mechanism that improved compression results further called "update exclusion". This mechanism is based on how the symbol counts for each context model are updated. When encoding with update exclusions, the predicted symbol count is incremented only if it is not already predicted by any higher order context. This means that the counts are updated only for the higher order contexts that are actually used to predict it. Thus, the counts reflect better which symbols are likely to be excluded by the higher-order contexts. This mechanism typically improves the compression rate by 1 or 2% [5]. On the other hand, when encoding without update exclusions, all the counts for all orders of the model are updated. The counts are incremented even if they are already predicted by a higher order context. The use of both of these mechanisms are investigated in our paper: PPM without update exclusions; and PPM with update exclusions (standard PPM).

Several PPM variations have been invented such as PPMA, PPMB, PPMC and PPMD. These variants differ in the way the escape probabilities are assigned. For example, PPMC uses escape method C, and PPMD uses escape method D. When a novel symbol is seen, the escape probability is sent, followed by the prediction using the next shorter context. Several escapes may be made before a context is reached which predicts the symbol. In the worst case the order(-1) model makes the prediction. Also, the maximum order of the context models may be included when the variant is described in the literature; for example, PPMD4 refers to a fixed order 4 PPM model using escape method D.

The PPMC variant was developed by Moffat [24] and has become the benchmark version. The probability estimates for this method (method C) is based on using the number of symbols that have occurred before, called the number of types:

$$e = \frac{t}{n+t} \quad \text{and} \quad p(s) = \frac{c(s)}{n+t}$$

where e represents the probability of the escape symbol, $p(s)$ denotes the probability for symbol s, $c(s)$ is the number of times the context was followed by the symbol s, n is the total number of times that the current context has occurred and t denotes the total number of types.

The PPMD variant was first developed by Howard in 1993 [25]. In most cases, experiments show that the PPMD performs better than the other variants. This variant is similar to the PPMC variant with the exception that each count is incremented by a 1/2:

$$e = \frac{t}{2n} \quad \text{and} \quad p(s) = \frac{2c(s)-1}{2n}.$$

An adaptive PPM compression model for Arabic language has been presented by Alhawiti [26]. This method is called Character Substitution of Arabic for PPM ("CSA-PPM"). It not only showed a considerable improvement in Arabic text compression but also for other texts that use Arabic script such as Persian and Kurdish. There are two main operations in this method, which are the pre-processing and post-processing. Each two-byte Arabic character is substituted with an equivalent number of the UTF-8 encoding scheme in the first operation, and as a result one output file is generated. Contrastingly in the post-processing operation, a reverse operation is performed by replacing the numbers with the original equivalent characters.

The fundamental idea of our approach depends on using a PPM compression method to compute codelength values of each possible permutation. The 'codelength' of a permutation for a cryptogram is the length of the compressed cryptogram, in bits, when it has been compressed using the PPM language model. The smaller the codelength, the more closely the cryptogram resembles the model. By using this metric, it can easily find the valid solutions automatically [6].

The most important step in our implementation is the training step. During the training phase, a large set of training texts is used to prime the models. Training text is chosen that is hopefully representative of the text being compressed and consequently better able to predict it. Experiments with English text show that training can improve performance dramatically as when skipping the training phase, there is not enough and sufficient data at the beginning to effectively compress the texts. In our experiments described below, we use a corpus of many different Arabic books and novels from different topics converted to 36 and 37 character Arabic to train our models (by case-collapsing non-alphabetic sequences to a single space). Classical and modern Arabic texts are used to produce this large corpus (which we have called the 'Mixed Arabic corpus' with details shown in Table I). With the training step, we in essence use static models, unlike standard PPM—that is, once the models have been primed using the training texts, they are not further updated when processing the ciphertexts.

Concerning the Mixed Arabic corpus, it is a large corpus of mixed classical and modern Arabic text. The corpus is a combination of the Bangor Arabic Compression corpus (BACC) [26], Corpus A [27] and a selection of files from the King Saud University Corpus of Classical Arabic (KSUCCA) [28]. The BACC is a 31-million words corpus collected from different sources such as magazines, books and websites. This corpus contains texts from a wide range of genres containing

Table I: Details of Corpora used to Construct the Mixed Arabic Corpus used to Train the PPM Models

Corpus	Bytes	List of topics
BACC	256,867,247	sports, culture, economics, education, art and music, political literature and heritage, history, religion
Corpus A	252,338,294	business, cinema, opinions, conferences, economics, politics
KSUCCA	4,086,066	religion, linguistics, literature, science, sociology, biography

both modern and classical Arabic. Corpus A is a modern corpus recently published with different genres which covers several current modern standard Arabic areas. This corpus is collected from the bilingual newspaper Al-Hayat website, and from the open-source online corpus, OPUS. It consists of 27-million Arabic words. KSUCCA is a corpus made up of classical Arabic texts with the purpose of studying the lexical semantics of the Holy Qu'ran. It is a 50-million word corpus covering several genres. The overall file size of our Mixed Arabic corpus is 513,291,607 bytes encoded using a UTF-8 scheme and consists of 58,068,493 words.

B. Calculating Codelengths Using the Gzip Compression Method

Gzip or GNU zip [4] is one of the most common lossless compression schemes used on the Unix operating system and on the Internet. It uses a dictionary based approach (using the Lempel-Ziv algorithm) whereas PPM is a statistical based approach.

The reason for experimenting with other compression schemes in our paper is to find out which is the most efficient scheme that can be applied to the problem of plaintext recognition using a compression based technique. In our paper, the calculation of the Gzip codelength metric will be based on using a relative entropy calculation which allows us to use 'off-the-shelf' software without the need to re-implement the method itself (as we have done with our PPM models). The codelength metric can be calculated using the relative entropy technique by the following formula [6]: $h_t = h_{T+t} - h_T$, where T denotes the training text (which will usually be large in size), t denotes the testing text, and h_T refers to the size of the compressed file T. Essentially, the codelength for a particular compression scheme is calculated as being the difference between the compressed size of some large training text with testing text concatenated after it compared to the compressed size of just the training text by itself.

We have also investigated using another well-known lossless compression scheme, Bzip2, using our relative entropy method. However, due to the block-sorting nature of the Burrows-Wheeler algorithm, the calculation of some of the relative entropy codelengths ended up being negative and therefore could not be used.

V. Our Method

A full description of our new method for the automated decryption of Arabic transposition ciphertext will be presented in this section. As stated, the fundamental idea of our method is based on using a compression scheme to calculate the codelength value to use as a metric for ranking the quality of each possible permutation [6]. PPMC, PPMD and Gzip are the compression schemes used for the codelength calculations.

First, we generate an Arabic transposition cipher in Unicode. Then, we perform our new cryptanalysis approach. Our new cryptanalysis method consists of two main phases. The first phase (Phase I) is based on trying to automatically decrypt an Arabic ciphertext using a transposition of specified size by exhaustively computing all possible transpositions, while the second phase (Phase II) is based on achieving readability (word segmentation). Achieving readability is gained by automatically adding spaces to the decrypted message that is produced from phase I, as the spaces are omitted from the ciphertext traditionally.

The pseudo code for the first part of our approach is presented in Algorithm 1. The first step in this algorithm focuses on removing all the non-alphabetical Arabic characters (including spaces) from the ciphertext (see line 1). At this stage, the text comprises just 36 Arabic letters. In order to calculate the codelength value for each Arabic permutation, a pre-processing operation is performed using an Arabic-specific encoding method using UTF-8 numbers (see line 4) where each Arabic character is substituted with an equivalent number in the UTF-8 encoding scheme. The algorithm then starts to try all possible key sizes and for each key size a permutation is performed over each cryptogram block trying to get a permutation with a smaller codelength value which represents the correct solution (lines 5 to 14). Each cryptogram is divided into blocks according to the key size (see lines 6 and 7). Then, a permutation is performed over each cryptogram (line 8). For each possible permutation, a codelength value is calculated (line 9 to 12). Two main compression-based models, PPM and Gzip, are used for calculating the codelengths (line 10). Permutations with smaller codelengths are kept in the priority queue as shown in line 11. We have found that the smaller the codelength, the more closely the cryptogram resembles the model.

Algorithm 1: Pseudo code of the main decryption phase 'Phase I'.

```
   Input  : ciphertext
   Output: decrypted text
 1 remove all non-Arabic alphabet characters and spaces from the
     ciphertext;
 2 maximum size of Q (priority queue) ← 3;
 3 maximum key-size of transposition ← 12;
 4 perform pre-processing operation by using Arabic-specific
     encoding method on the ciphertext;
 5 foreach key-size do
 6    if ciphertext-length mod Key-size = 0 then
 7       divide the ciphertext into blocks according to key-size;
 8       perform a permutation over each ciphertext blocks;
 9       foreach possible permutation do
10          calculate codelength value using the CSA-PPM or
              Gzip compression model;
11          store a permutation with smaller codelength value in
              Q;
12       end
13    end
14 end
15 return the priority queue 'Q' (the 'decrypted text');
```

The pseudo code for the second phase of our approach is presented in Algorithm 2. As stated, this phase focuses on achieving readability. The second assessment step is performed in this phase through adding spaces automatically to the decrypted messages that was produced from Phase I, and then assessing the quality of the solutions by computing the codelength values for each possible message. Our solution for Phase II uses the Viterbi algorithm to find the best possible segmentation (see Algorithm 2). For each of the text produced as output from Algorithm 1, the Viterbi algorithm is used to search for the best segmentation sequence and this then is stored in a priority queue (lines 2 to 5) which is used to return the result (line 7). A post-processing operation is performed afterwards, by using the Arabic-specific decoding using UTF-8 numbers. PPM and Gzip compression models are also used again as the method for calculating the codelength values.

Algorithm 2: Pseudo code for Phase II

```
  Input  : the priority queue 'Q' from Phase I
  Output: segmented decrypted text
1 maximum size of Q1 (priority queue) ← 1;
2 foreach text in Q do
3     use the Viterbi algorithm to search for the best segmentation
        sequences;
4     store the text that have the best segmentation which present in
        Q1;
5 end
6 perform post-processing operation by using the
    Arabic-specific encoding method on the text in Q1;
7 return the best Arabic segmented decrypted text from Q1;
```

Two modified variants of the PPM modelling system, PPMD and PPMC, have been investigated in our method. Also two forms of these schemes are examined, one with update exclusions (i.e the standard PPMD and PPMC) [5] and one without update exclusions. The use of the Gzip compression scheme is examined as well, in order to explore the most effective method that can be applied to the problem of automatically recognizing the valid decryption.

Experiments with a full range of variations have been conducted (PPMC, or PPMD, with and without update exclusions; Gzip). However, for the purposes of this paper, only the results for three variants are shown in order to illustrate either the best performing schemes or to illustrate interesting results for comparison. In the first variant, called Variant I, order 5 CSA-PPM without update exclusions is used for Phase I and Phase II (using the Viterbi algorithm). For the second variant, Variant II, order 5 CSA-PPM with update exclusions (the standard PPM) have been used for Phase I and Phase II. The Gzip compression scheme is used in the last variant, Variant III.

VI. Experimental Results

The experimental results are discussed in this section. In our experiments, the order-5 CSA-PPM models have been trained on a corpus of many different Arabic books and novels from different topics using 36 and 37 Arabic character (when space is included). As explained above, classical and modern Arabic texts are used to produce this large Mixed Arabic corpus. After this training operation and during cryptanalysis, these different models remain static. Regarding the cryptograms test corpus, 125 different cryptograms were chosen at random from different resources to be used as testing texts. Cryptogram lengths ranged from 17 to almost 800 letters.

A sample of decryption is shown in Table II for the Arabic cryptogram 'ميكقنصدن دكعوكهاو اأرغخمنك'. Compression codelengths are listed in bits with the lowest three results presented for Phase I.

Table II: Output Produced for the Sample Cryptogram 'ميكقنصدن دكعوكهاو اأرغخمنك'.

Phase I	Phase II
Variant I	
108.64 صديقكمننهاكوعدوكمنأغراك 110.62 نكمنصديقكعدوهاكوراكمنأغ 111.37 يكمننصدقكعدوكهاوأراكمنغ	صديقك من نهاك وعدوك من أغراك يك من نصدقك عدوكها وأراك منغ نك من صديقك عدوها كورا كمن أغ
Variant II	
108.15 منكنصديقدوعكهاكواكرمنأغ 110.76 قكمنصدينوعدوهاككغراكمنأ 110.85 صديقكمننهاكوعدوكمنأغراك	صديقك من نهاك وعدوك من أغراك قك من صدين وعدوها ككغراكمنأ منكن صديق دوعك هاك واكر من أغ

We have created a transposition cipher for Arabic language that depends on using the Unicode encoding. For each run, a random key is generated to encrypt the original message (plaintext). Then, we have applied our cryptanalysis method to automatically break this message (ciphertext). Different permutation key sizes (from 2 to 12) and different cryptograms with different lengths have been examined in our experiments.

The experimental results of Phase I showed that for the first variant I, when the CSA-PPM methods without update exclusions is used, 97% of the cryptograms are successfully decrypted without any errors. For variant II, when the standard CSA-PPM method is used, the results showed that 95% of the ciphertext are solved with no errors. For the last variant, Variant III, using the Gzip scheme, a success rate of only 90% was achieved.

The main idea of the second phase is based on inserting spaces automatically into the cracked messages produced from Phase I. The codelength compression metric is also used in this phase to automatically rank the quality of the solutions to find the correct message. In order to measure the differences between the original texts and the decrypted texts that resulted from our decryption method, the Levenshtein distance metric is used. It is a string metric that is commonly used to calculate the differences (either deletions, insertions or substitution) between any two messages [29]. The results show that for variant I, 97% of the cryptograms are successfully decrypted and segmented. For variant II, which resulted in a 95% decryption success rate from Phase I, now 97% of the ciphertexts are effectively decrypted and segmented as a result of Phase II.

It is important to note that the second phase of our approach has the ability to find better solutions not found by the first phase. For example, referring to Table II, the best two solutions as output from using Variant II shown in column one are not correct (the third solution is the correct one). However, after the second phase has been applied, the best segmentation of the third solution in column one now appears as the best solution in column two which is correct.

Arabic word segmentation is considered a difficult problem [30]. As Arabic language is a rich morphological language, this characteristic represents a real challenge for identifying the

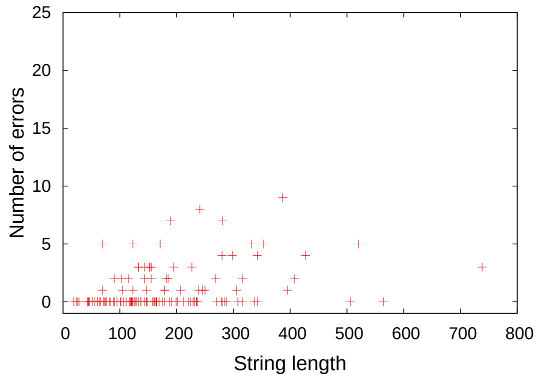


Figure 2: Segmentation errors produced from Variant I.

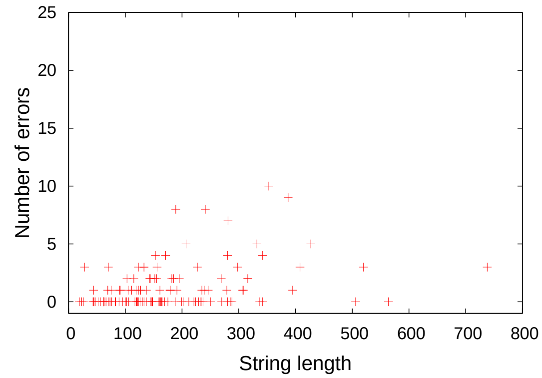


Figure 3: Segmentation errors produced from Variant II.

correct segmentation among many possible corrected alternatives. For example, the word “ ورده” could be segmented into “ ورده ” and “ و رده”, which means “rose” and “and his reply”. Both of these segmented forms are correct, but with completely different meanings and contexts. However, according to our approach and results, in almost all cases the proper readable solutions were obtained for variants I and II, but not for Variant III (using gzip).

Output examples (with spaces) produced from different variants are exhibited in Table III. For example, this table shows how ‘إناللهلاينظرإلصوركموأموالكمولكنينظرإلقلوبكموأعمالكم’, which is the best result produced from Phase I for a sample cryptogram, is effectively segmented with no errors using both variants I and II, while Variant III shows poor performance with nine segmentation errors.

Table III: Example of Solved Ciphertexts with Spaces by the Three Different Variants

Variants	Number of errors	Decrypted message (with spaces)
Variant I	0	إن الله لا ينظر إلى صوركم وأموالكم ولكن ينظر إلى قلوبكم وأعمالكم
Variant II	0	إن الله لا ينظر إلى صوركم وأموالكم ولكن ينظر إلى قلوبكم وأعمالكم
Variant III	9	إن اللهلاينظرإلصوركموأموالكمولكنينظرإلقلوب- كموأعمالكم

Figures 2, 3 and 4 present scatter-plots for the number of errors (on the y axis) versus the string lengths of cryptograms (on the x axis) for the different variants. For example, Figure 2 presents the number of word segmentation errors for each decrypted message that was output from Variant I (when PPMD is used). This figure shows that Variant I offered better performance than the other variants. More than 63% of the testing texts have been correctly segmented with no errors and a quarter of them with three errors or less. Just four cryptograms ended up with seven, eight or nine errors. These errors can be attributed to the rich morphological characteristic of the Arabic language with unusual words being used for various topics such as names of places and people, that were not included in the training text.

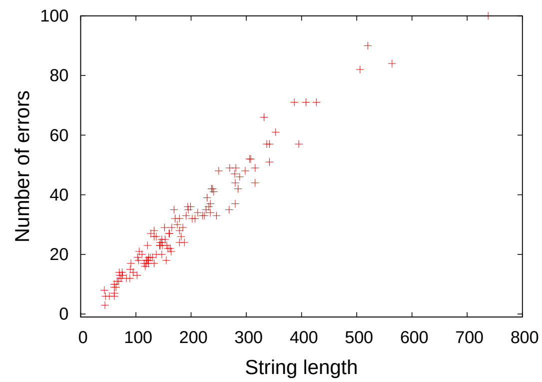


Figure 4: Segmentation errors produced from Variant III.

Variant II showed a slightly worse result than Variant I. Fifty three percent of the cryptograms were effectively segmented with no errors and 40 percent were segmented with four errors or less. Only one cryptogram consisting of 353 letters ended with ten errors. Figure 3 outlines Variant II results. The PPMD model is used in this variant. The number of errors produced from Variant III is shown in Figure 4. It is clear that the number of errors for each decrypted ciphertext in this case is much higher, with most of the errors being greater than 20. Also none of the cryptograms finished with no errors, and just eight of the solved cryptograms were segmented with the number of errors being less than ten.

The average number of space addition errors for the 125 testing texts for each variant using both main models, PPMD and PPMC, is presented in Table IV. The results show that PPMD produces slightly better results than PPMC and that the number of errors for the variants, except Variant III, is quite low with Variant I performing best.

Table IV: Average Number of Errors for the Phase-II Variants. (The PPMD and PPMC Models are used for the First Two Variants)

Variants	I	II	III
PPMD	1.10	1.23	30.43
PPMC	1.18	1.27	30.43

To investigate more about the accuracy of the compression methods that we used in segmenting the 125 Arabic decrypted texts, three main metrics are used: recall rate, precision rate and error rate. The first metric, recall rate, is calculated by dividing the number of successfully segmented words over the number of words in the original testing texts. The second metric, error rate, is calculated by dividing the number of unsuccessfully segmented words by the number of words in the original testing texts. The last metric, precision rate, is calculated by dividing the number of successfully segmented words by the number of words which are correctly and incorrectly segmented [6].

Table V: Three Different Accuracy Rates for Evaluating the Quality of the Word Segmentation

Variants	**Recall rate%**	**Precision rate%**	**Error rate%**
Variant I	96.35	95.94	3.65
Variant II	95.50	95.67	4.50
Variant III	1.53	10.02	98.47

The results presented in Table V indicate that the variants which depend on the CSA-PPMD compression methods were able to achieve very high recall and precision rates, reflecting the quality of the Arabic word segmentation. The error rates resulted from unusual words not included in the training text.

The average time that is needed to recognize the correct solutions for the different Arabic transposition cryptograms with different block sizes is shown in Table VI. This table presents the average elapsed time for the automatic cryptanalysis of three different length cryptograms, for both Phase I and Phase II. For each cryptogram, the average elapsed time in seconds is shown after running the approach ten times. The results show that the average time increases as the key size increases but overall the method is reasonably efficient.

Table VI: Average Time Required (Across 10 runs) to Automatically Break Cryptograms with Different Lengths and Different Keys Size

Ciphertext length (Letter)	**Key size**					
	Time (in seconds)					
	5	**6**	**7**	**8**	**9**	**10**
40	0.85	0.87	0.90	1.08	2.60	12.30
150	1.23	1.27	1.42	1.90	13.93	48.20
300	3.54	3.77	3.89	4.94	23.13	95.50

VII. Conclusions

The Arabic language is one of the most used languages in the world, but unfortunately the techniques used to encrypt and decrypt information in Arabic language are rare with few publications. In particular, we have not been able to find any cryptographic system for transposition ciphers in Arabic. In this paper, an Arabic transposition cipher is investigated and an automatic cryptanalysis of this cipher is also introduced. A success rate of 97% was achieved by using different Arabic compression models. PPM compression schemes were used as a basis for computing the codelength value (which we have found is an accurate way of measuring information in the text). Various Arabic ciphertexts with different lengths have been successfully decrypted and effectively segmented. This efficient method eliminates any need for human intervention.

The PPM compression variants showed better performance than Gzip in both decryption and segmentation processing. According to the main decryption phase, PPM models with no update exclusions, performed better than the standard PPM models with a 97% success decryption rate comparing to 95%. Concerning the second decryption phase, a success rate of 97% was achieved for the different PPM models with a high level of performance in segmenting words.

By using our method for automatically breaking Arabic transposition ciphers, various Arabic ciphertexts with distinct lengths (from 17 to almost 800 Arabic letter) have been examined. Different block lengths were also experimented with (from 2 to 12) and 97% of these ciphers were successfully decrypted with no error. This compares with 100% for the automatic cryptanalysis of transposition ciphers for the English language [6]. Although the success rate for Arabic is slightly worse than English, it still leaves open the question whether Arabic is a more challenging language for cryptanalysis or not. Further investigation needs to be performed regarding other Arabic ciphers and the automatic decryption of these ciphers to ascertain whether this language really provides a greater challenge for cryptanalysis.

References

[1] S. F. Gary and C. D. Fennig, *Ethnologue: Languages of the world (21st edition)*. Dallas, Texas: SIL International, 2018.

[2] I. A. Al-Kadit, "Origins of cryptology: The Arab contributions," *Cryptologia*, vol. 16, no. 2, pp. 97–126, 1992.

[3] T. C. Bell, J. G. Cleary, and I. H. Witten, *Text compression*. Prentice-Hall, Inc., 1990.

[4] Gzip. (2012) The gzip home page. (http://www.gzip.org), (accessed 1 March 2016).

[5] W. J. Teahan, "Modelling English text," Ph.D. dissertation, University of Waikato, New Zealand, 1998.

[6] N. R. Al-Kazaz, S. A. Irvine, and W. J. Teahan, "An automatic cryptanalysis of transposition ciphers using compression," in *the 15th International Conference on Cryptology and Network Securit*. Springer Int. Publishing, 2016, pp. 36–52.

[7] J. Seberry and J. Pieprzyk, *Cryptography: an introduction to computer security*. Prentice-Hall, Inc., 1989.

[8] K. P. Bergmann, R. Scheidler, and C. Jacob, "Cryptanalysis using genetic algorithms," in *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation (GECCO '08)*. New York, USA: ACM, 2008, pp. 1099–1100.

[9] A. Dimovski and D. Gligoroski, "Attacks on the transposition ciphers using optimization heuristics," in *Proceedings of International Scientific Conference on Information, Communication & Energy Systems & Technologies (ICEST)*, Sofia, Bulgaria, 2003, pp. 1–4.

[10] S. Singh, *The code book: the science of secrecy from ancient Egypt to quantum cryptography*. Anchor, 2000.

[11] Y. Alqahtani, P. Kuppuswamy, and S. Shah, "New approach of Arabic encryption/decryption technique using Vigenere cipher on mod 39," *International Journal of Advanced Research in Information Technology and Engineering*, vol. 2, no. 12, pp. 1–9, 2013.

[12] S. D. Rihan and S. E. F. Osma, "Arabic cryptography technique using neural network and genetic algorithm," *International Research Journal of Computer Science, vol. 3, pp. 35-42*, 2016.

[13] R. S. Habeeb, "Arabic text cryptanalysis using genetic algorithm," *Iraqi Journal for Electrical and Electronic Engineering*, vol. 12, no. 2, pp. 161–166, 2016.

[14] A. Muanalifah, "Construction of key echange protocol over max-plus algebra to encrypt and decrypt Arabic documents," *Journal Of Natural Sciences And Mathematics Research*, vol. 1, no. 2, pp. 51–54, 2017.

[15] T. Ng, K. Nguyen, R. Zbib, and L. Nguyen, "Improved morphological decomposition for Arabic broadcast news transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 4309–4312.

[16] M. Stamp and R. M. Low, *Applied cryptanalysis: breaking ciphers in the real world*. John Wiley and Sons, 2007.

[17] J. Giddy and R. Safavi-Naini, "Automated cryptanalysis of transposition ciphers," *The Computer Journal*, vol. 37, no. 5, pp. 429–436, 1994.

[18] W. Grundlingh and J. H. Van Vuuren, "Using genetic algorithms to break a simple cryptographic cipher," 2003, retrieved 31 March 2003 from (http://dip.sun.ac.za/˜vuuren/abstracts/abstrgenetic.htm), submitted 2002.

[19] R. Matthews, "The use of genetic algorithms in cryptanalysis," *Cryptologia*, vol. 17, no. 2, pp. 187–201, 1993.

[20] S. A. Irvine, "Compression and cryptology," Ph.D. dissertation, University of Waikato, New Zealand, 1997.

[21] B. Schneier, *Applied cryptography (2nd edition)*. John Wiley and Sons, Inc., New York, 1996.

[22] M. J. Wiener, "Efficient DES key search," in *Advances in Cryptology: CRYPTO'93, Lecture Notes in Computer Science*, vol. 733. Berlin. Springer-Verlag, 1993.

[23] J. G. Cleary and I. H. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Transactions on Communications*, vol. 32, no. 4, pp. 396–402, 1984.

[24] A. Moffat, "Implementing the PPM data compression scheme," *IEEE Transactions on Communications*, vol. 38, no. 11, pp. 1917–1921, 1990.

[25] P. G. Howard, "The design and analysis of efficient lossless data compression systems," Ph.D. dissertation, Brown University, Providence, Rhode Island, 1993.

[26] K. M. Alhawiti, "Adaptive models of Arabic text," Ph.D. dissertation, Bangor University, 2014.

[27] S. Alkahtani, "Building and verifying parallel corpora between Arabic and English," Ph.D. dissertation, Bangor University, 2015.

[28] M. Alrabiah, A. Al-Salman, and E. Atwell, "The design and construction of the 50 million words KSUCCA," in *Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL)*. The University of Leeds, 2013, pp. 5–8.

[29] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet Physics Doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[30] A. M. Zeki, "The segmentation problem in Arabic character recognition the state of the art," in *the 1st International Conference on Information and Communication Technologies*. IEEE, 2005, pp. 11–26.

Systematic Analysis and Classification of Cardiac Rate Variability using Artificial Neural Network

Azizullah Kakar[1], Naveed Sheikh[2],
Department of Mathematics
University of Balochistan Quetta

Bilal Ahmed[3]
Department of Computer Science
Virtual University of Pakistan

Saleem Iqbal[4], Abdul Rahman[5]
Department of Mathematics
University of Balochistan

Saboor Ahmad Kakar[6], Arbab Raza[7]
Department of Mathematics
University of Balochistan Quetta

Samina Naz[8]
Department of Mathematics
University of Balochistan Quetta

Junaid Babar[9]
Department of Computer Science
University of Balochistan Quetta

***Abstract*—Electrocardiogram (ECG) is acquisition of electrical activity signals in cardiology. It contains important information about the condition and diseases of heart. An ECG wave, pattern, size, shape and the time interval between different peaks of P-QRS-T wave provide useful information about the diseases which afflict heart. Heart rate signals vary and this variation contains important indicators of cardiac diseases. To assess autonomic nervous system, heart rate variability is popular and non-invasive tool. These indicators contained in ECG wave appear all the day or occur randomly in the day. So, computer based information is much useful over day long interval to diagnose heart disease. Thus, this paper deals with classification of heart diseases on the basis of heart rate variability using artificial neural network. Feed forward neural network is considered to be almost correct 85% of the test results.**

Keywords—Electrocardiogram (ECG); cardiology; P-QRS-T wave; autonomic nervous system; heart rate variability; artificial neural network; time and frequency domain; pattern recognition; diseases classification

I. Introduction

Electrocardiogram (ECG) is the study of electric activities coming from the heart through cardiac muscles. An ECG records the electric potential, its origin and propagation through cardiac muscles. For cardiac physiology, ECG is the representative signal [1]. The ECG in cardiac physiology plays the role of finger print for cardiac state. A useful way to understand the state and the health of heart is using electrocardiogram (ECG). An ECG defines the measurement of the electric activity of the heart [2]. A small electric impulse is produced when a heart beats. These impulses move in cardiac muscles when a heart contracts. So, ECG is responsible to detect these impulses. The ECG test has no pain and harm and that it does not put any extra activities on subject body. A healthy heart does not beat regularly it depends upon many factors like exercise, physical and mental stress [3]. R-R interval gets changed even in rest around its mean value [4].

A. Heart Rate Variability

Heart rate variability measures variation of time interval between two consecutive intervals of time. Heart rate variability is non-invasive marker which is used to measure sympathy-vagal activities. Heart rate variability is also used to measure beat-to-beat variation of two consecutive R-R intervals.

B. Structure

An ECG consists of wires, electrodes and a reading machine. Fig. 1. Electrodes are connected to arms, legs and chest of the patients, the wires from the electrodes are connected with ECG. The machine records electric impulses coming from the heart through the cardiac muscles. The recorded beat is shown on the screen or paper strip.

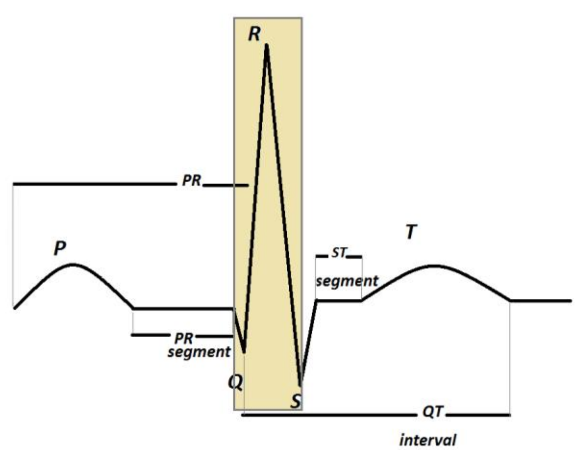


Fig. 1. An ECG waves with different peaks.

The ECG wave pattern determines the condition of heart [5]. To have proper diagnosis, it requires several hours, using ECG and heart rate variability.

For clinical diagnosis, doctors often ask the patient to perform number of tests for the diagnosis of disease. It was observed that the tests results did not lead to effective diagnosis of the disease. So, the classification approach of neural network is useful to classify the diseases in different classes. Thus, Neural Network approach for diagnosing heart diseases is more feasible, faster and accurate.

Artificial neural network is a mathematical model, inspired from human brain; it models human brain, functioning and activation. Thus, it is very useful in classification and pattern recognition. In recent years, ANN has produced accurate results in medical field like in lungs cancer, heart diseases and also in other fields like weather forecasting, marketing, stock markets and computer sciences.

In section II, a review of the literature is presented, while section III contains the Methodology adopted for this work. Section IV is about the result and discussion and then finally we conclude in section V.

II. Literature review

An artificial neural network which interconnects artificial neurons in a network is an inspiration of biological neurons with a task to classify information in different classes. The natural neurons and natural neural network system are considered imaginary while modeling ANNs. ANNs consist of inputs, hidden layers, outputs, sigmoid function, synaptic weights, and perceptron. These perceptrons are basically responsible to transfer information. ANN has many advantages as compared to conventional computers. An ANN is an experienced learning processing function which learns from experiences. ANN is used for classification of data in different classes; it puts information into its weights.

The heart rate is the study of variations of R-R interval in time [6]. The rhythms of healthy heart are not same. It depends on many factors; if someone is exercising its heart beats fast as compared to beating in normal conditions. Also heart beat depends on physical and mental stress. So the interval between normal sinus beats varies in normal condition. One of the world's most important and unsolved problems is sudden cardiac death [7].

The process of respiration, blood pressure, regulation, thermoregulation, circadian rhythms and other factors are the main sources of heart rate variability.

Basically the heart rate variability can be measured using two methods

A. Time Domain

B. Frequency Domain

Time domain measures the "total variability occurred in heart rhythms" while frequency domain states the "the underlying rhythms" [8].

An ECG consists of many significant points to state the condition of a heart. To analyze the heart state, spectral analysis is of great importance [9]. An ECG consists of important points when afflicting the heart. It has different results in different interval of time. Sometimes it is not able to show the symptoms of the diseases. For proper diagnosis, ECG and HRV may require several hours which is time consuming and there is a chance of missing of important points, because of large data and excess of information volume. The HRV is measured between the R-R intervals of PQRST waves. The useful frequencies measured by ECG often lies between 40Hz. While plotting HRV, we draw heart variability against time axis. In this paper we classify the heart diseases by artificial neural network into four classes and the heart rate variability (HRV) as base signal. Two parameters obtained from heart rate signals were used for the required classification.

III. Methodology

The data for the proposed study was collected from "Civil hospital" and "BMC hospital Quetta". The collected data contained 500 ECG papers of four different kinds of heart diseases (normal ECG, left bundle branch block, complete heart block, ischemic and sick sinus syndrome). The heart diseases in this paper are classified by artificial neural network and back propagation algorithm as methodology classifies using heart rate variability as base signals.

A. Neural Network Classifier

Artificial Neural network is basically a mathematical concept which is used to model the human brain. As the name suggests, the neural network is an inspiration of human brain. The study of human brain has been an interesting phenomenon in every time for human itself. The development and advancement in modern electronics have a natural effect to develop a machine which could work like human brain. ANN is used to solve many problems and multilayer perceptron is used for solution of problems and modeling [10]. When we model ANN the structure of biological neuron is very abstracted [11]. To predict a disease is one of the most interesting and challenging phenomena for modern world [12]. ANN is one of the powerful subjects for predicting and forecasting [13]. ANN is also used to identify cancer cells and the best cure of cancer is to identify cancer cells in its early stage [14]. ANN has many useful applications in medicines [15]. A biological brain contains billion of cells of different type and length [16].

Artificial Neural Network (ANN) is an attempt to develop an artificial network system that could work like human brain. As in biological neurons, a highly interconnected network of neurons performs the task of processing information from one place to another. In the same way ANN is also working in parallel to process the information. For weather forecasting ANN has much accurate results [17]. ANN is very useful in areas such as pattern recognition and classification. In ANN, input of the network determines activation of the function that determines if the function has an output or not. This is the result of total aggregate of total input given to the system. Physicians make decisions on the tests related to the diseases or on the conditions observed on patients earlier [18].

ANN has self-learning capabilities that enables it to produce better results. The more it is trained the more it can produce correct results.

The problems that has no algorithmic solution, ANN is very useful to solve those problems. An ANN is very useful for the phenomena like predictions, pattern recognitions, image analysis and medical diagnosis. A layer in ANN is the subgroup of processing elements. The ANN consists of two or more than two layers. The first layer which receives input from some source is called input layer, while the last layer which is dependent of the input layer and the layers lying between input and output layers are called hidden layers.

The structure of ANN is parallel, if some element of the ANN fails to function then it does not affect the whole system of perceptron and it has ability to learn from experience. The input data given to the system is classified into accurate and fair classifications. The information received from inputs are stored in connection weights (synapse) which makes decisions on total aggregate. It has a result when comparing the output with a threshold. When the output value is greater than threshold than the function is activated and the mentioned function is not activated when output value is less than threshold value. ANN has different number of layers in different situations. ANN may have only feedback or feed forward structure for the classification of data.

ANN is the network from which we can configure an output of our desire. Multilayer perceptron feed forward neural network with the back propagation algorithm and multilayer is a supervised learning process. In this process the desired output is trained and the input data is transformed into desired output. The linear combination of weights and inputs can be calculated as

$$x_1 w_1 + x_2 w_2 + \dots + x_n w_n$$

In summation form it can be written as

$$\sum_{i=1}^{n} x_i \, w_i \tag{1}$$

Where x_i is n-number of inputs and w_i are the corresponding weights.

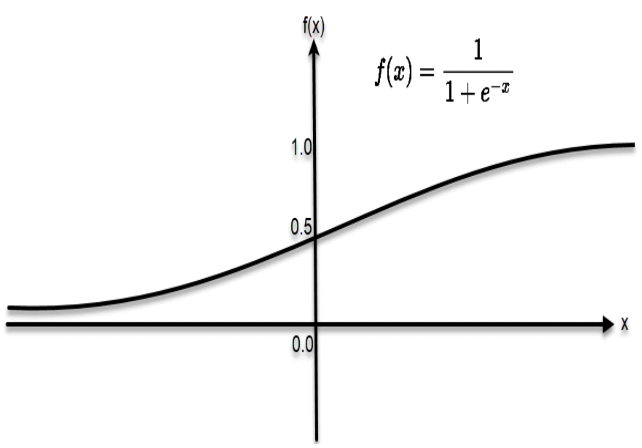


Fig. 2. Sigmoid Function.

An ANN contains input layer, one or more than one hidden layers and an output layer. A multilayer network with the j^{th} neuron with t iterations having actual output y and desired output d with an error 'e' is given by

$$e_j(t) = d_j(t) - y_j(t) \tag{2}$$

And the sigmoid function is given by Fig. 2.

$$\frac{1}{1+e^{-x}} \tag{3}$$

To the input layer, data is given in the form of numbers. The hidden layers process the data with the help of an activation function, Fig. 3. and in this way the data is forwarded to output layer. The hidden layer q_j^h and output layer b_k can be calculated as

$$q_j^h = \sum_{i=1}^{4} (w_{ji} \, q_i + c_j^h) \tag{4}$$

And

$$b_k = \sum_{j=1}^{4} (w_{kj}^o \; q_j^h + c_k^o) \tag{5}$$

Where w_{ji}^h and w_{kj}^o are the weights of connection and c_j^h and c_k^o are called biases terms.

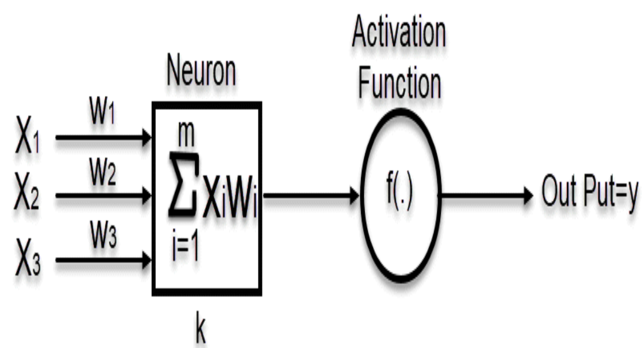


Fig. 3. Model of Artificial Neuron with Activation Function.

The weights are updated by the following equations

$$w_{kj}(\text{Updated}) = w_{kj} + \mu s_j^h e_k \tag{6}$$

And

$$w_{ji}\ (\text{Updated}) = w_{ji} + \mu s_i e_k \tag{7}$$

And partial derivative of total error can be calculated as

$$\frac{dE}{dw_{ij}}(t) = \frac{1}{2} \sum_{p=1}^{p} \frac{dE_P}{dw_{ij}}(t) \tag{8}$$

Where μ is the learning rate, its value is 0.9 and w_{kj} is the weight from neuron j to neuron k.

Input layer receives data through nodes or units, the subsequent layers process the data with the help of an activation function. The output layer consists of four neurons. The network has the ability to identify only four outputs. i.e. (0001, 0010, 0100, 1000) decoded in binary number system.

B. Back Propagation

It is the method in which input data are fed from input layer to hidden layer, from hidden layer to output layer and then from output layer to input layer through hidden layers, in the form of adjusted weights. Initially, the weights are assigned random numbers and then modified to minimize the error. The weights are updated from output layer towards input layer.

C. Diseases Classification using ANN

For the current paper we classify the heart diseases into four classes as

- Normal
- Complete heart block
- Dilated cardiomyopathy
- Sick sinus syndrome

The ANN was used to classify the input data in different classes feed by Spectral entropy and Poincare plot geometry.

D. Spectral entropy

The variation in time interval between R-R peaks of an ECG is referred as heart rate variability (HRV) [19]. Spectral entropy determines the complex form of time series. The Fourier transformation is one of the basic tool to transform time domain into frequency domain and determines the power of heart rate at the given frequency f. Fourier transformation converts time domain into frequency domain [20]. It expresses the power as function of frequency f. The spectral entropy H is given by

$$H = -\sum_{f} p_f \log (p_f)$$

Here p_f is the power density function at frequency f and $0 \leq H \leq 1$

E. Poincaré Plot Geometry

Poincare plot geometry is consisting of R-R intervals of an ECG. Fig. 4.

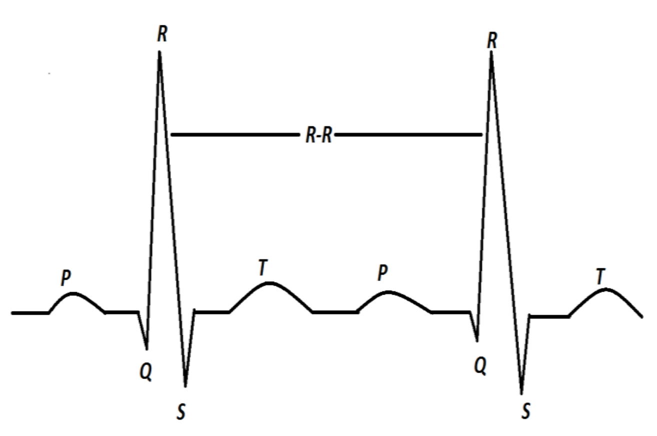


Fig. 4. Two Consecutive R-R Intervals.

It represents the ($RR_{i,} RR_{i+1}$) intervals at the co-ordinate axis and each point on the co-ordinate axis are corresponding to two consecutive intervals of time recorded at ECG [21]. The cloud of points can be determined by its length SD_2 which is parallel to the identity line y = x and its breadth can be determined by across the line SD_1.

SD_1 And SD_2 are the short and long term heart rate variability respectively, Fig. 5.

The expression SD_1/SD_2 gives the relation between these components [22].

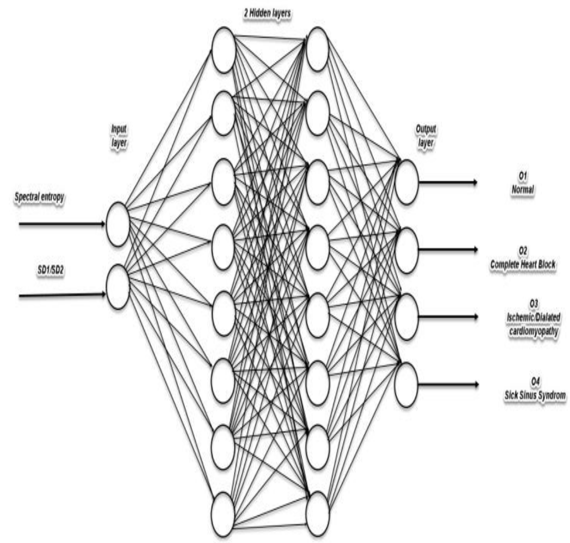


Fig. 5. Four -layered feed forward neural network.

IV. Results and Discussion

An ECG wave pattern consists of important information about the condition of heart. Different shape and size of an ECG wave with different time interval between different peaks provide significant information like whether the heart is healthy or not. To study pinpoint cardiovascular disease in large quantities of data consisted on several hours is time consuming and sometimes there is chance for misreading of important points about diseases contained in ECG wave. So computer based technology is very useful in predicting and diagnosing disease. A well trained and well-designed neural network is used to predict and classify cardiac diseases into four different classes as shown in the table 1. Multilayer neural network with input set of data derived the results related to non-linear parameter of heart diseases by increasing the number of hidden layers. The more hidden layers we have the less mean square error we obtain. Precisely by increasing the number of hidden layers we can successfully train neural network and an accurate prediction about classification is possible. The classification of cardiovascular diseases by back propagation algorithm feed by input data (spectral entropy and Poincare plot geometry) is predicted in table 1. To have better predictions about diseases the algorithm was trained and weights were modified and updated.

This research is going to be a pioneer one in application of Artificial Intelligence particularly the Artificial Neural Network in classification of heart diseases. This area of the subject is not much studied by the researchers. So, this research is a kind of corridor that opens the new avenues to further studies in the application of ANN especially in the diagnosis and classification of heart diseases.

TABLE I. RANGE FOR INPUT DATA FOR ANN FOR CLASSIFICATION

Class	Spectral entropy	SD_1/SD_2
More table copy[a]	1.63±0.024	0.80±0.16
Complete heart block (CHB)	00.86±0.054	00.64±0.025
Ischemic	1.11±0.10	0.64±0.024
Sick sinus syndrome (SSS)	1.27±0.135	0.96±0.32

V. CONCLUSION

The artificial neural network classifier is the new human brain inspired mathematical modeling having very useful results in medical while diagnosing diseases and their classifications. It enhances the patient abilities in diagnosing diseases. To predict disease with certainty and accuracy rate of 100% is not possible yet. The accuracy rate depends on many factors including input data's size and quality of training data and parameters selected for input data. So the results evaluated in table 1 by back propagation is effective the range of 80-85%.

REFERENCES

[1] U. Rajendra Acharya, P. Subbana Bhat, S.S Iyengar, Ashok Rao, Summet Dua. (2003). Classification of heart rate data using artificial neural network and fuzzy equivlance relation. Pattern recognition, 61-68.

[2] R. Acharya U, A. Kumar, P.S.Bhat, C.M. Lim, S.S. Lyenger, N. Kanathal, S.M. Krihan. (2004). Classification of caridiac abnormalities using heart rate signasl. Medical & Biological Engineering & Computing 2004, 288-293.

[3] Phyllis K. Stein,PhD, Matthew S. Bosner, Robert E. Kleiger, MD, and Brooke M. Conger,BS St. Louis, Mo. (1994). Heart rate variability: A measure of cardiac autonomic tone. CURRICULUM IN CARDIOLOGY, 1377-1381.

[4] Alberto Malliani, Federico Lombardi, Massimo Pagani and Sergio Cerutti, "Power Spectral Analysis of Cardiovascular Variability in Patient at Risk for Sudden Cardiac Death", Journal of Cardiovascular Electrophysiology 5(3):274-86, April 1994

[5] Akselrod. S, Gordon, D., Ubel. F.A., Shannon, D.C., Berger A.C., Cohen, R.J., "Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control", science, 1981 July 10; 213(4504):220-222

[6] P. T. Ahamed seyd, V.I. Thajudin Ahamed, Jeevamma Jacob, Paul Joseph K. (2008). Time and Frequency Domain Analysis of Heart Rate Variability and their Corrleation in Diabetes Mellitus. World Acadmy of Science, Engineering and Technology International Journal of Medical and Health Sciences , 85-68.

[7] Federico Lombardi, Timo H Makikallio, Robert J Myerburg, Heikki V Huikuri, "Sudden cardiac death: role of heart rate variability to identify patient at risk", cardiovascular research, volume 50,issue 2,1 May 2001, pages 210-217

[8] Phyllis K. Stein, PhD, Mathew S.Bosner, MD, Rovert E. Kleiger, MD,and Brooke M. Conger, BS St. Louis, M. (n.d.). Heart rate variability: A measure of caridac autonomic tone. Curriculum in Cardiology, 1376-1381

[9] Weisssman, M.W., Markwoitz, J.S., Ouelette, R., Greenwald,S., and Hahn,. (n.d.). Panic disorder and cardiovascular / cerebrovascular problems.

[10] M. Akhil Jabbar, B. D. (2013). Classifkcation of Heart Diseas Using Artificial Neural Network and Feature Subset Selection. GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY NEURAL & ARTIFICIAL INTELLIGENCE

[11] Saumya Bajpai, Kreeti Jain and Neeti Jain. (2011). Artificial Neural Network. IJSCE, 27-31.

[12] Mevlut Ture, Imran Kurt, A. Turhan Kurum, Kazim Ozdanar. (2005). Comparing classification technique for predicting essential hyprtension. ELESVIER, 583-588.

[13] Holger R. MAIER, Gracine C. Dandy. (2000). Neuial network for predicting and forecasting of water resources variables: a review of modeling issues and applications. ELSEVIER, 101-124.

[14] Zhi-Hua Zhou, Yaun Jiang, Yu- Bin Yang, Shi - Fu Chen. (2002). Lung cancer cells identification based on Artificial Neural Network Ensembles. ELESVIER, 25-36.

[15] Qathera Kadhim Al- Shayeaand Itedil S.H. Bahia. (2010). Urinary Syetem Diseases Diagnosis using Artificial Neural Network. IJCSS, 118-122.

[16] I.A. Basheer and M. Hajmeer. (2000). Artificial neural network: fundamentals, computing, design and applications. ELESVIER, 3-31.

[17] Meera Narvekar, Priyanca Fergos. (2015). Daily weather forecasting using artificial neural network. International Journal of computer applications, 9-13.

[18] Resul Das, Ibrahim Turkoglu, abdulKadir Sengur. (2009). Effective diagnosis of heart diseases through neural netwok ensembles. ELESVIER, 7675-7680

[19] P.T. Ahamed Seyd, V.I. Thajudin Ahamed, Jeevamma Jacob, Paul Joseph K, "Time and Frequency Domain Analysis of Heart Rate Variability and their Correlations in Diabetes Mellitus", International Journal of Medical and Health Sciences, Vol:2, No:3,2008

[20] S.M. DEBBAL, F. BEREKSI-REGUIG. (2008). Frequency analysis of heartbeat sounds. IJBSCHS, 85-90.

[21] R. Acharya U, A. Kumar, P.S. Bhat, C.M. Lim, S.S. lyengar , N. Kannathal and S.M. Krishnan. (2004). Classification of Caridac abnormaities uising heart rate signals. Medical & Biological Engineering & Computing 2004, 288-293

[22] J Piskorski and Guzik. (2007). Geometry of he Poincare plot of RR intervals and its asymmetry in healthy adults. Physiol. Meas., 287-300.

(IJACSA) International Journal of Advanced Computer Science and Applications,
Vol. 9, No. 11, 2018

A New Steganography Technique using JPEG Images

Rand A. Watheq, Fadi Almasalha, and Mahmoud H. Qutqut
Faculty of Information Technology
Applied Science Private University
Amman 11931 Jordan

***Abstract*—Steganography is a form of security technique that using ambiguity to hide a secret message within an ordinary message between senders and receivers. In this paper, we propose a new steganography technique for hiding data in Joint Photographic Experts Group (JPEG) images as it is the most known type of image compression between the lossy type compressions. Our proposed work is based on lossy compression (frequency domain) in images. This type of compression is susceptible to change even for the smallest amount of change which raises a difficulty to find a proper location to embed data. This should be done without affecting the image quality and without allowing anyone to notice the hidden message. From the senders side, first, we divide the image into 8*8 blocks, then apply a Discrete Cosine Transform (DCT), Quantization, and zigzag processes respectively. Second, the secret message is embedded at the end of each selected zigzag block array using the best method of our experimental results. Third, the rest of the code applies the Run Length Code (RLC), Different Pulse Code Modularity (DPCM) and Huffman encoder to obtain the compressed image that includes the embedded message. From the receiver's side, we will reverse the previous steps to extract the secret message using an encrypted shared key via a secure channel. Our experimental results show that the best array content size of zigzag computed coefficients are between 1 to 20. This selection allows us to utilize more than half of the image blocks to embed the secret message and the difference between the cover image that holds the secret message and the original cover image is very minimal and hard to detect.**

Keywords*—*Steganography; hide secret message; JPEG image; lossy compression; frequency domain; zigzag

I. Introduction

We live in the era of technology, which involves in growing use of the Internet; so we need a secure transfer for a secret message through the Internet. Data hiding techniques have lately become important in several application fields. The importance of data hiding techniques stems from the transmission medium is being not secure. Hence, some methods are needed to make it difficult for the unauthorized user to extract information from the images. The primary motivations to hide data are to protect personal, sensitive and private confidential data [1].

A data hiding technique is an innovative type of secret communication technologies. It is a form of embedding data into digital media (image, video, audio, text) with a minimum amount of perceivable degradation to the host signal, for the aim of identification, control access to digital media and copyright. The techniques used for data hiding differ depending on two factors; the quantity of data needs to be hidden, and the required invariability of these data to manipulation [2]. Any loophole to fill data in a host signal, either statistical or perceptual are possible targets for removal by lossy signal compression [2]. The answer to a successful data hiding is to find holes that are suitable for utilization by compression algorithms. There are different techniques of data hiding like watermarking, steganography, and cryptography.

To this end, we propose a new steganography technique for hiding data in JPEG images. Our proposed work is based on lossy compression (frequency domain) images; which raise a difficulty to find a proper location to embed data without affecting image quality and without allowing anyone to notice the hidden message. Our work focuses on how to embed and extract secret data without affecting the original image. After a comprehensive study of the JPEG encoding process, it was found that the best location to embed data is in a zigzag order based on several experiments. From the senders side, the work of this paper firstly divides the image into 8*8 blocks. Then, we apply the Discrete Cosine Transform (DCT), Quantization, and zigzag processes in order to calculate the required blocks for the secret message. After that, the process of embedding the data in the selected blocks will start based on a specific range. The rest of the code applies the Run Length Code (RLC), Differential Pulse Code Modulation (DPCM) and Huffman encoder as before to obtain the compressed image that includes the embedded message. During the previous process, the secret message is embedded at the end of each selected zigzag block array using the best method of our experimental results. The sender sends the compressed image to a communication channel, and the receiver will reverse the previous steps to extract the hidden data using a Shuffled Block Candidate Array (SBCA) array that has been sent via a secure channel. However, the receiver can use the shared private key to generate the same SCBA generated on the sender. Sharing the private keys allows both sender and receiver to exchange messages without the need to send the SCBA of each embedding process. In this paper, we choose to share the SCBA array for the sake of simplicity and to focus more on the proposed embedding technique.

The remainder of this paper is organized as follows. In Section II, we overview the background topics of steganography and image compression. Section III overviews related work of our paper. In Section IV, we introduce and describe our proposed technique and its procedures. Our experiments and results are presented in Section V. Section V describes the conclusion of the paper.

II. Background

In this section, we provide background topics of steganography and image compression.

A. Steganography

The word steganography derived from the Greek language and means ”covered writing” [1]. Steganography is a science of hiding message, file, image or other data types within another data file, in a way to prevent anyone to know if there is message hiding in the original message. Fig. 1 shows the steganography steps. In the following, we describe the different types of steganography.

1) Steganography based on Protocols: There are three main types of steganographic protocols as shown below [3].

- Pure Steganography: This type assumed the sender and receiver must have access to the process of embedding and extraction.
- Secret Key Steganography: In this type, the sender and the receiver use the same key to hide and extract the secret message in a cover media.
- Public Key Steganography: The sender and the receiver use the different key to hide and extract the secret message in a cover media.

2) Steganography based on Cover Media: There are five types of Steganography based on the carrier object that is used for hiding the secret data. These types are Audio Steganography, Text Steganography, Image Steganography, Video Steganography, and Network Steganography [4].

3) Steganography based on Domain: Steganography techniques can be divided into two domain types as described below.

- Spatial Domain: In this type, the secret message bits are embedded in a cover image by directly changing the pixel's color values. An example of this type is the Least Significant Bit (LSB) technique.
- Frequency Domain: In this type of techniques is tries to embed the secret message bits in the frequency domain coefficient of the cover image. An example of this is the Discrete Cosine Transform (DCT) technique.

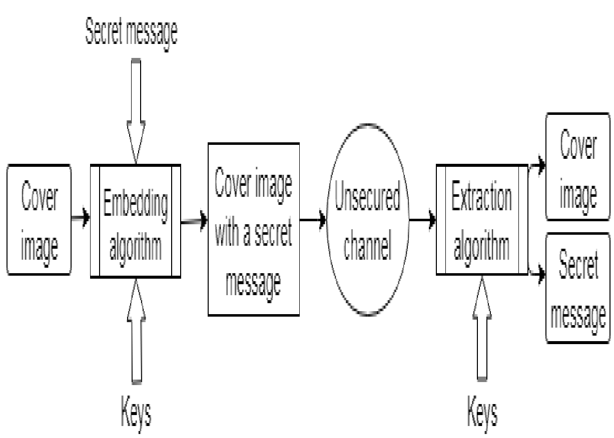


Fig. 1. Steganography steps overview (reproduced from [4]).

B. Image Compression

Image compression is the science/art of efficiently encoding digital images to minimize the number of bits that are required to represent an image [5]. Image compression divided into two types as follows [6].

- Lossy Compression: The concept of this type is that when data is compressed, it loses a portion of itself. Therefore, there became a slight difference between the image before compression and the image after decompression. In this type of compression, the amount of loss can be adjusted to achieve the desired compression ratio. Examples of this type are Joint Photographic Experts Group (JPEG), Moving Picture Experts Group (MPEG), and Moving Picture Experts Group Layer-3 Audio (MP3).
- Lossless Compression: The concept of this type is that when the data is compressed, it does not lose any data. In other words, the reconstructed data after the compression is identical to original data. The most common techniques of this type are Huffman coding, RLC, and Lempel–Ziv–Welch (LZW).

An image compression process contains two phases [7]; encoder and decoder.

1) Encoder: In this phase, we take an original image and apply some steps to obtain a compressed image. It contains multi-steps as shown in Fig. 2.
2) Decoder: In this phase, we use the result from the encoder phase and apply inverse the steps (the steps that applied in the encoder phase) to obtain a decompressed image. It contains multi-steps as shown in Fig. 3.

III. Related Work

There are many research work on data hiding techniques in multimedia data. We overview the latest research work related to our paper in this section. H. Lu *et al.* propose an algorithm for binary images that can embed a watermark in DC component [8]. They combine the embedding watermarks in the DC components of DCT and employing a biased binarization threshold. The results show that the embedding algorithm provides some degree of robustness against conventional image processing. M. Kaur *et al.* [9] propose a method using two watermarks in a cover image. They embed the first watermark in the middle frequency of the blue component. The second watermark is embedded into magnitude coefficients of the Discrete Fourier transform (DFT) in the form of local peaks. They concluded that the Red Green Blue (RGB) model is more suitable in case of repeating watermark than the YCbCr, and the blue component is more suitable for embedding the watermark. The work in [10] proposes a method using a secret data and Laplacian sharpening method. The author compresses the secret data using Huffman coding. After that, she embeds it in a cover image using Laplacian sharpening method to determine the useful hiding places based on a threshold value. S. Sujatha *et al.* [11] propose a method to generate watermark from the original image. First, they used low frequency, a rescaled version with the help of Arnold transform to generate a watermark. Second, they used a high frequency of Discrete Wavelet Transform (DWT) to embed the watermark. This method provides good robust against attacks such as JPEG compression, scaling and rotation.

R. Preda proposes a method for image authentication using a semi-fragile watermark in [12]. In this method, the author used a semi-fragile watermark to detect malicious tampering in the image and embedded a bit of watermark in the coefficient by means quantization. He selected wavelet coefficient with random permutation using a secret key to provide a higher level of security. This method achieved decent image quality and tampering detection resolution with a low watermark payload. In work [13], the authors propose a method to embed a binary watermark in a compressed color image. They embedded a binary watermark in middle bands of (Y) luminance. This method provided good watermark in colored images. M. Khan *et al.* [14] propose a method that combined between DWT and DCT. They applied DWT to an original image after that applied DCT to High-High (HH) band to obtain matrix H. They converted H matrix into four quadrants using zigzag. Finally, they applied Singular Value Decomposition (SVD) in each quadrant to embed the watermark. In research work [15], the authors propose a method using DCT. They applied DCT to a middle frequency of B plane and selected DCT (4, 3), DCT (5, 2) to embed the watermark. This method provides fair robust against different types of attack. M. Mundher *et al.* [16] they propose a method using the preprocessing stage and Discrete Slantlet Transform (DST). They used a preprocessing stage to find the best channel. After that applies DST to the selected quadrant to find the best frequency sub-band to embed the watermark in the best frequency. This method employed to ensure the imperceptibility and robustness of watermarked images.

In the research work [17], the authors propose a method

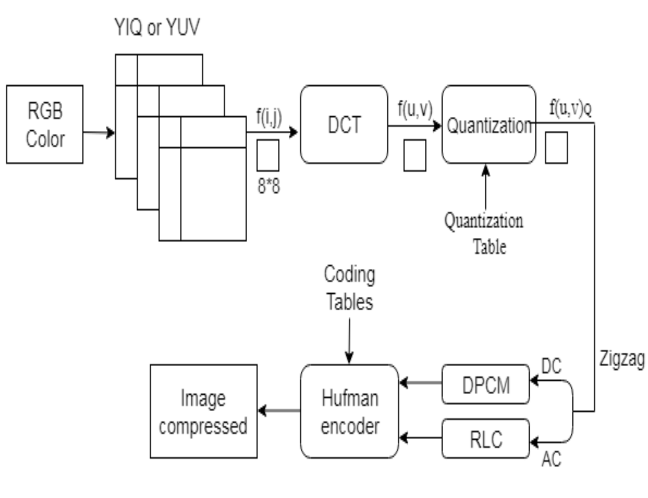


Fig. 2. JPEG Encoder Compression (reproduced from [7]).

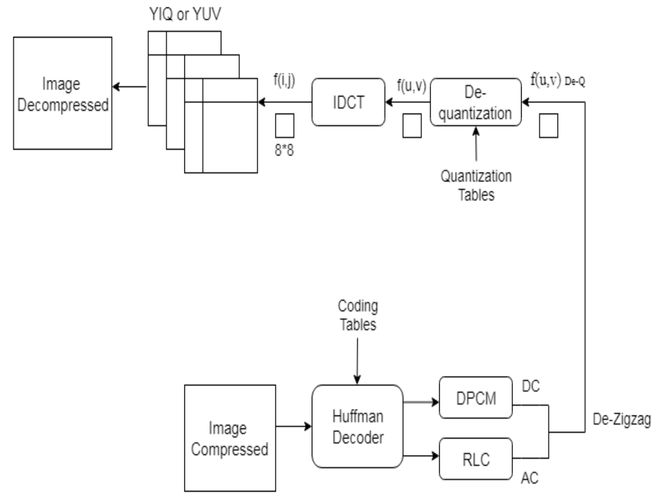


Fig. 3. JPEG Decoder Compression (reproduced from [7]).

using DWT in Hue, Saturation, and Intensity (HSI) color space. First, selecting I plane from the original image (in HSI form). Then, they apply DWT to it and select low-frequency. Second, dividing the watermark image into 8*8 blocks. Finally, comparing both images based on entropy values, then multiply the scaling factor. This method provides fair, robust watermark to noise attacks. In [18], the authors propose a method using 3D chaotic cat map, DWT and lifting scheme. The irregular output of the chaotic cat map is used to embed the secret message in a cover image. Then, they embedded the secret message in the mean coefficient of DWT. Finally, they applied lifted scheme to guarantee lossless extraction of hidden information. The method provides good performance and imperceptibility according to two measures (Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM)). The work in [5] proposes a method called Compressed Encrypted and Embedded Technique (CEET). In this method, they compressed a secret image using JPEG compression to reduce the size of a secret image. After that, they selected key randomly and applied encryption to the secret image to increase security. Finally, they embedded a secret image in a cover image using LSB by inserting at least 2 bits. This method collected different techniques to reach the goal of steganography which is embedding highest possible rate while remaining undetectable to steganalysis. J. Mazumder *et al.* [19] propose a method using DWT and optimized message dispersing. They used a high frequency of all color components (R, G, B) to embed a secret message. They start from the last column of each of the components from top to bottom based on the length of the message. Finally, they used Mean Squared Error (MSE) and PSNR to measure the imperceptibility of the method that shows it is acceptable compared to other methods.

G. Swain in [20] proposed a method using the LSB but in new method Group of Bits Substitution (GBS). Firstly, they appends length of message in the beginning of the binary message, after that embedding one bit of the secret message in the cover image if size of cover image $\geq$ size of secret message or embedding two bits of secret message in one byte of the cover image if size of cover image size of

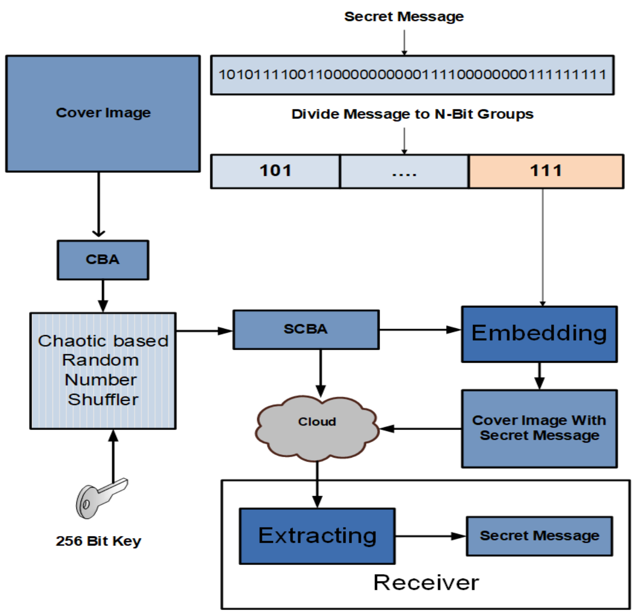


Fig. 4. Overview of Our Proposed Technique.

secret message divided by 2. Finally, this method provides the security to a higher level. In [21], the authors propose a method using "LSB substitution" and "Arnold transform". They used "Arnold transform" to encrypt the secret images using different keys. Then, embedding the first Most Significant Bit (MSB) (in secret image 1) in last three LSB once using red pixel of the cover image, again using green pixel and a blue pixel. Results reveal that the method successfully secures the high capacity data keeping the visual quality of transmitted image satisfactory. S. Kumar *et al.* [22] propose a method using index based on a chaotic mapping. They used a chaotic map to generate pseudo-random numbers (index). The index is used as a position to embed a message in a cover image, and they used LSB substitution to embed bit pixels in the cover image. In [23], the authors propose a system using three methods Huffman, zigzag and Pptimal Pixel Adjustment Process (OPAP). They used Huffman to compress the secret message to reduce the size of the secret message and to provide high embedding capacity then, used Zigzag scanning to select the pixels that the secret message be hidden in them and used OPAP to enhance the quality of the Stego-images to keep minimizing embedding error.

We test the effect of modification in the zigzag order using three methods then selected the best method that was adopted in this paper to hide data; the selected method provides better security with high embedding capacity and a better Stego-image quality than the existing system.

IV. Our Proposed Technique

The visual quality of Stego-image (imperceptibility), the security level (robustness) and embedding capacity are three basic principles that are used to evaluate the performance of the steganographic scheme. This section presents our proposed solution illustrated in Fig. 4, which provides a new method for embedding and extracting secret messages in zigzag order using JPEG image format. Our proposed technique consists

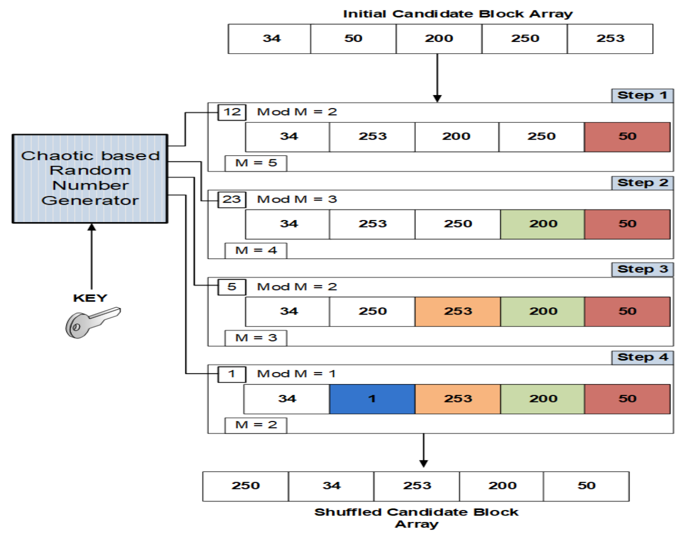


Fig. 5. Random number generation phase (reproduced from [24]).

Algorithm 1 Generating Shuffled Candidate Block Array

Input:
Cover Image M.
Private key K.
Zigzag array threshold value ZT

```
Create empty Candidate Block Array (CBA).
Create empty array Shuffled Candidate Block Array (SCBA).
Get No of Micro Blocks N from cover Image M.
for each B of N Blocks do
    Retrieve Zigzag Array Z of B Block.
    if Index of last Coefficient value of Z >= ZT then
        Save the location of this block into CBA.
    else
        Skip it
    end if
end for
SCBA = PRNG_Shuffle(CBA) [24].
```

Output: Shuffled Candidate Block Array (SCBA)

three phases; namely, **1)** Random Number Generator phase, **2)** Generating the Shuffling Array phase, **3)** Embedding or Extracting Secret message in cover image phase. In the following subsections, we will describe our proposed technique phases in details.

Phase 1: Random Number Generator

A Random Number Generator (RNG) is a computational or physical device or a piece of software code which is designed to generate a sequence of numbers or symbols that cannot be reasonably predicted better than by a random chance. The input range for generating every random number depends on the bit length of the keys and nature of data and operation to be performed on data. We use a Chaotic based random number generator proposed by [24] to shuffle the secret message location within the cover image. We use a 256 key to increase the security level.

Phase 2: *Shuffle Array Generation*

In this phase, we generate an array to represents the secret message in order to build a corresponding location inside the cover image. Firstly, the secret message is converted to a sequence of ASCII codes. Then, we store the ASCII code into a binary representation. The size of the latter array is divided by 3 bits to calculate the required Number of image micro-Blocks (NOB). Second, the cover image is decoded to find the locations of the micro-blocks that passes the validity condition which will be discussed in later sections. Third, we initiate the Candidate Block Array (CBA) array (where its size equals to NOB) that contains the micro-blocks locations obtained from the previous step. Finally, we pass the CBA array into the Pseudo-Random Number Generator (PRNG) to obtain the Shuffled Candidate Block Array (SCBA). Fig. 5 illustrates the process of shuffling the CBA.

Algorithm 1 represents the generation of the SCBA. The algorithm receives the cover image, secret key and a threshold value of the maximum number of non zero coefficients in the zigzag array. Steps 1 to 3 initialize the arrays CBA and SCBA. Steps 4 to 11 iterate through each block to examine the threshold value condition. After step 11, the CBA will contain the sequence location of each zigzag array that passes the threshold condition which selected to hold the secret message. In step 12, the CBA array is passed to the proposed shuffler in order to shuffle the location. Step 12 will randomize the location of the data which makes the possibility of assembling the secret message very hard without the availability of the SCBA.

Example: Generating SCBA for the secret message "Computer" using the image of Fig. 6.

- Cover image: Lena
- Secret massage: Computer
- Data array: 01000011 01101111 01101101 01110000 01110101 01110100 01100101 01110010 00100000
- Total blocks in image: (512*3*512)/64=12288 blocks.
- Using blocks that contain values between 1 to 20
- NOB = 24
- Initialize CBA
- CBA = {1,2,4,7,10,13,16,19,21,22,23,28,29,31,34, 37,40,43,46,49,52,55,58,61}
- SCBA = {13,21,55,52,31,22,2,43,49,46,61,4,19, 16,7,34,23,37,40,29,1,10,28,58}

Phase 3: *Embedding secret message in JPEG Image*

This phase works to embed the secret message in the cover image; without affecting the compressed image. This phase is delicate as JPEG image format (lossy compression), and the frequency domain is complicated and hard to find a location to embed the secret message without affecting the compression process. We present in this paper two different proposed methods to be evaluated based on the following two targets before giving the final algorithm.

Fig. 6. Lena Cover Image

1) We need to find the maximal number of bits to be embedded at the end of each selected blocks in the cover image. Our experiments include 2 bits, 3 bits, and 4 bits chunks.
2) We need to find how to add the data chunks to the last value in the selected block in a way that guarantee minimal effects on the encoding process.

In order to achieve our goals, we proposed a set of experiments using two different methods of adding the data chunks at the end of the zigzag array. First, it worthy to explain the structure of the zigzag order to better understand the proposed methods. As we see in Fig. 7, the 2d array of the calculated coefficients is converted into a 1D array in a way that accumulates the zero value coefficient at the end of the array. This structure allows us to add our data chunks after the last non-zero coefficient value without affecting the encoding or decoding process. Now, based on the structure of the zigzag order we proposed the following two methods in order to find the best way to add our data chunks without affecting the final compressed image and obtaining the most available space to hide the data in.

1) First Method: The first method puts the secret data after the last non-zero coefficient using three data chunk sizes as follows.

1) We divide the binary form of the secret message into a group of 2 bits and represent each group with a decimal value of -1, 0, 1, and 2. The corresponding

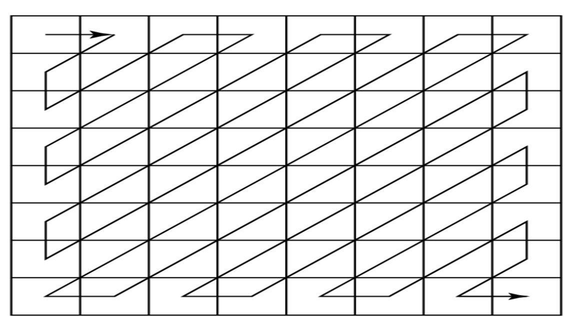


Fig. 7. Zigzag scan process in JPEG [7].

decimal values are then embedded after the last value of the non-zero coefficient.

2) We divide the binary form of the secret message into a group of 3 bits and represent each group with a decimal value from -4 to 4 excluding the zero. The corresponding decimal values are then embedded after the last value of the non-zero coefficient.
3) We divide the binary form of the secret message into a group of 4 bits and represent each group with a decimal value from -8 to 8 excluding the zero. The corresponding decimal values are then embedded after the last value of the non-zero coefficient.

2) Second Method: The second method adds the value of the data chunk to the last non-zero coefficient value also using the same previous three data chunk sizes as follows.

1) We divide the binary form of the secret message into a group of 2 bits and represent each group with a decimal value of -1, 0, 1, and 2. The corresponding decimal values are then added to the value of the last value of the non-zero coefficient.
2) We divide the binary form of the secret message into a group of 3 bits and represent each group with a decimal value from -4 to 4 excluding the zero. The corresponding decimal values are then added to the value of the non-zero coefficient.
3) We divide the binary form of the secret message into a group of 4 bits and represent each group with a decimal value from -8 to 8 excluding the zero. The corresponding decimal values are then added to the value of the non-zero coefficient.

After the evaluation process explained in Section V, we found that the second case in the first method is the best approach in term of minimal effect on the final compressed image and the larger space to hide data.

Algorithm 2 shows the process of embedding the secret message in a cover image while Algorithm 3 shows the steps of extracting the secret message.

At this point, an example is always good to illustrate the process of our proposed algorithms. The following example is divided into two parts. The first part will illustrate the process of embedding a secret message. The second part illustrates the process of extracting the same message from the cover image.

Algorithm 2 Embedding secret message in a cover image.

Input:
Cover Image M.
Hidden data size in each block N
Secret Message SM.
Shuffled Candidate Block Array(SCBA)

1: Convert SM into Binary Representation B
2: Initialize counter to Zero
3: **for** each D of size N from B **do**
4: Location = SCBA[counter].
5: Retrieve Zigzag Array Z of M[Location].
6: Insert the D value after the last value in Z.
7: **end for**

Output: Steganographic Image

A) Embedding the secret data in the 8*8 block pointed with a rectangle in Fig. 8: .

The secret message in this example is ”OLA”. The compression algorithm will generate the quantized DCT coefficients shown in Fig. 9 A of the selected blocks. The later 2D array is then converted into a 1D array using the zigzag order as shown in Fig. 9 B.

The binary representation of our secret message is:
O = (01101111)
L = (01101100)
A = (01100001)

The groups of binary representation using the 3 bits data size chunks will be as follow:
Group 1 = 000
Group 2 = 001
Group 3 = 010
Group 4 = 011
Group 5 = 100
Group 6 = 101
Group 7 = 110
Group 8 = 111

The previous group is mapped to each decimal values as follows:
Group 1 = -4
Group 2 = -3

Algorithm 3 Extracting secret message in a cover image.

Input:
Steganographic Image M
Shuffled Candidate Block Array(SCBA)

1: Initialize B array
2: Initialize counter to Zero
3: **for** each I from SCBA **do**
4: Retrieve Zigzag Array Z of M[I].
5: Extract the D value from the last value in Z.
6: Insert D to B
7: **end for**
8: Convert B from Binary to ASCII representation SM

Output: Secret Message (SM)

Fig. 8. Lena's image in Y sub-band

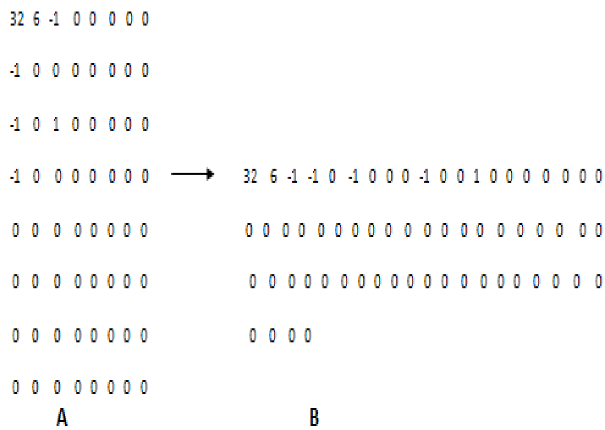


Fig. 9. A) Quantization of DCT coefficients, B) zigzag order.

Group 3 = -2
Group 4 = -1
Group 5 = 1
Group 6 = 2
Group 7 = 3
Group 8 = 4

Now, using the first three bits of the letter O. We place the corresponding decimal valuec (-1) after the last non-zero value within the zigzag array showed in Fig. 9 B). The resulted Zigzag array will be:
32 6 -1 -1 0 -1 0 0 0 -1 0 0 1 -1 0 0 0 . . . 0 0 0

Finally, the processed of encoding will continue as usual by applying RLC on DC component and DPCM on AC value then to Huffman encoding.

B) *Extracting the secret message from the cover image:* At the receiver side, the last value in the Zigzag array of the retrieved block based on the SCBA will be extracted and mapped back to the original bit representation. The extracted value will be added to a buffer that accumulates the binary representation of the completed secret message. The receiver will obtain the SCBA array through a secure channel.

V. Performance Evaluation and Results

In this section, we show and discuss the results of our experiments to illustrate the impact of hiding the secret message based on our proposed technique. We first provide a brief description of the experiment setup followed by the conditions and results of the experiment.

A. Implementations

Our proposed algorithm is implemented entirely using Java Eclipse. We calculate the difference between the cover image and cover image after embedding the secret message using MATLAB to measure the effect of the embedding method on the compression process. We also developed a Graphical User Interface (GUI) that allows the user to input a secret message to be embedded in a cover image using the proposed technique.

B. Experiments

In order to choose the best zigzag array of the cover image blocks, we need to answer the following questions.

1) Does the number of elements in the zigzag array before the last non-zero coefficient tolerate the inserted data chucks differently?
2) Does the size of the inserted chucks affects the compression process differently?
3) Which proposed method of adding the data chunk affect the compression process and what is the best threshold to use?

The first information we need is to find how many blocks are available in each image of our test data set and for each block how many are the coefficient values in the zigzag array. Our data set includes four popular images that many researchers use in order to compare the results. In TABLE I, it shows the maximum number of blocks in each range using 17 different intervals based on the number of coefficient values in each zigzag array before the last non-zero values of each image block. The results in TABLE I show the following information.

1) The number of blocks in each image.
2) The average number of the block in each range.
3) The ratio between the number of blocks in each interval, and the number of blocks in each image.

Based on TABLE I, we can observe that the number of blocks is higher on the 1-30 and 1-20 intervals. All the other intervals have less than 30% average which limits the size of the secret message to be embedded. Now, the question of which interval is better to use in term of less impact on the compression process. To answer the latter question, we generated 1000 different secret messages that occupy at least 50% of the selected blocks and measured the difference of pixel values before the embedding and after the embedding process. The results lead that using the interval 1-20 of non-zero coefficient values before the last non-zero coefficient value always resulted in less difference. Thus, we continue our experiments using the interval 1-20; although it has 7% fewer blocks than the 1-30 interval. This results answered the first question of our experiments cause.

Now to answer question two and three, we used the previous interval as our threshold value to choose which zigzag array to embed our data in. In TABLE II we show the experiments of each of our proposed methods in order to find the best method and the best size of the inserted data chunk. TABLE II shows the following information:

1) The used method and case.
2) The number of pixels changed during the embedding process with a value greater or equal to 50.
3) The average of difference using all values pixel.

The reason for showing the number of pixels changed with values equal or larger to 50 is to emphasize what human can detect. Usually, human eyes cannot detect changes on pixels with values less than 50. Thus, the less no of changed pixels means that human eyes will not notice any changes on the picture after the embedding process. This test is not enough, as most of the detection tools are computerized. In this case,

TABLE I. NUMBER OF BLOCKS BASED ON THE NUMBER OF COEFFICIENTS UNTIL THE LAST NON-ZERO VALUE

Number of blocks							
Total blocks	12288	6500	108000	108000	17856	30720	
Zigzag array contents interval	gg (1024*633)	r (510*737)	Flower (1920*1200)	Girl (1920*1200)	Animal (512*512)	Lena (512*512)	Average
1 - 30	(29%) 8783	(46%) 8253	(53%) 56968	(52%) 55854	(48%) 3142	(47%) 5783	23130
1 - 20	(25%) 7806	(42%) 7459	(48%) 51945	(44%) 48013	(43%) 2809	(40%) 4886	20486
1 - 15	(21%) 6349	(26%) 4721	(38%) 40747	(26%) 27808	(36%) 2348	(33%) 4061	14339
1 - 10	(20%) 6222	(25%) 4469	(36%) 38714	(22%) 23703	(34%) 2248	(32%) 3886	13207
10 - 30	(8%) 2561	(21%) 3784	(17%) 18254	(30%) 32151	(14%) 894	(15%) 1897	9923
15 - 30	(8%) 2434	(20%) 3532	(15%) 16221	(26%) 28046	(12%) 794	(14%) 1722	8791
10 - 20	(5%) 1584	(17%) 2990	(12%) 13231	(23%) 24310	(9%) 561	(8%) 1000	7279
30 - 60	(6%) 1778	(27%) 4771	(12%) 2359	(14%) 14725	(14%) 922	(9%) 1081	4772
20 - 40	(5%) 1567	(9%) 1607	(6%) 6868	(14%) 15010	(11%) 716	(14%) 1703	4578
30 - 50	(5%) 1635	(25%) 4435	(2%) 2348	(13%) 14308	(12%) 771	(9%) 1081	4096
20 - 30	(3%) 977	(4%) 794	(5%) 5023	(7%) 7841	(5%) 333	(7%) 897	2644
40 - 60	(4%) 1188	(22%) 3958	(0.48%) 514	(7%) 7556	(8%) 539	(2%) 275	2338
40 - 50	(3%) 1045	(20%) 3622	(0.47%) 503	(7%) 7139	(6%) 388	(2%) 275	2162
30 - 40	(2%) 590	(5%) 813	(2%) 1845	(7%) 7169	(6%) 383	(7%) 806	1934
30 - 45	(4%) 1213	(17%) 3079	(2%) 2314	(11%) 12314	(10%) 691	(9%) 1068	1394
45 -60	(2%) 565	(9%) 1692	(0.04%) 45	(2%) 2411	(4%) 231	(0.11%) 13	826
50 - 60	(0.004%) 143	(2%) 336	(0.01%) 11	(0.39%) 417	(2%) 151	(0%) 0	176

the number of the pixel value has no effect of the detection process. Thus, the second information which is the average difference is introduced to fulfill this test. As we see in TABLE II, using the first method and 2-bits data chunk size has the less average of difference. However, the first method using the 3-bits data chunk size also have the same average difference. In this case, using larger data chunk will provide more space to hide the secret message. Thus, we can conclude that using Method 1 is always better than Method 2. Also, using case 2 is better for the sake of space availability for hiding the data.

TABLE II. RESULTS OF ALL CASES OF RANGE FROM 1 TO 20 ZIGZAG VALUE

Number of non-black points (≥ 50)								
No. of method	Case No.	Animal	Flower	gg	Girl	Lena	r	Avg.
Method 1	Case 1 (2 bits)	1	0	1	0	0	0	0.33
	Case 2 (3 bits)	1	0	1	0	0	0	0.33
	Case 3 (4 bits)	1	0	1	0	4	1	1.16
Method 2	Case 1 (2 bits)	5	0	76	3	0	3	14.5
	Case 2 (3 bits)	462	34	0	43	30	22	98.5
	Case 3 (4 bits)	454	28	0	49	22	28	96.83

C. Results

In this section, we present the results based on the experiments that we discovered in the previous subsection. The following examples show the cover images in JPEG format as input, the cover image after embedding a secret message and the difference image between the cover image and the cover image after embedding the secret message. We use the following values for the following examples:

1) The specific range is from 1 to 20.
2) The secret message embedded in the selected block by using the second case of the first method (3 bits from the secret message).
3) The pixel value is greater than or equal to 50.

Example One: Using the animal image in Fig. 10 (a) as a cover image, the number of non-black points that represent the difference between Fig. 10 (a) and Fig. 10 (b) is illustrated in Fig. 10 (c) and is equal to 1. The MSE is 9.88, and the PSNR is 38.22.

Example Two: Using the flower image Fig. 11 (a) as a cover image, the number of non-black points that represent the difference between Fig. 11 (a) and Fig. 11 (b) is illustrated in Fig. 11 (c) and is equal to 0. The MSE is 1.89, and the PSNR is 45.41.

Example Three: Using the Lena image Fig. 12 (a) as a cover image, the number of non-black points that represent the difference between Fig. 12 (a) and Fig. 12 (b) is illustrated in

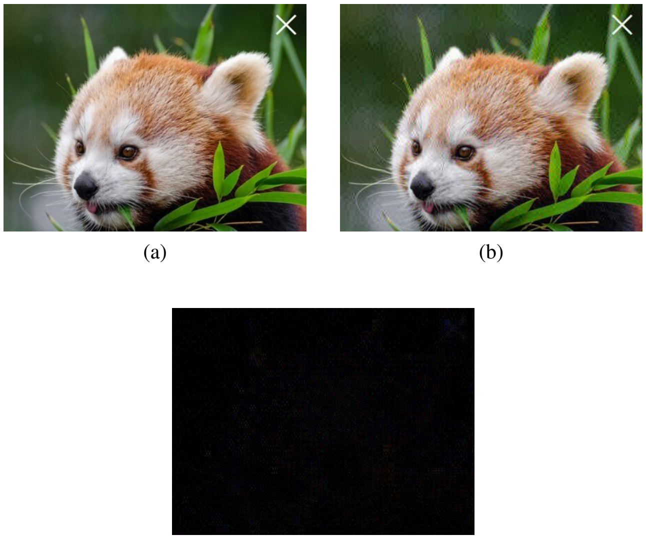


Fig. 10. A) Animal cover image, B) Animal cover image after embedding the secret message, C) The difference image between A and B.

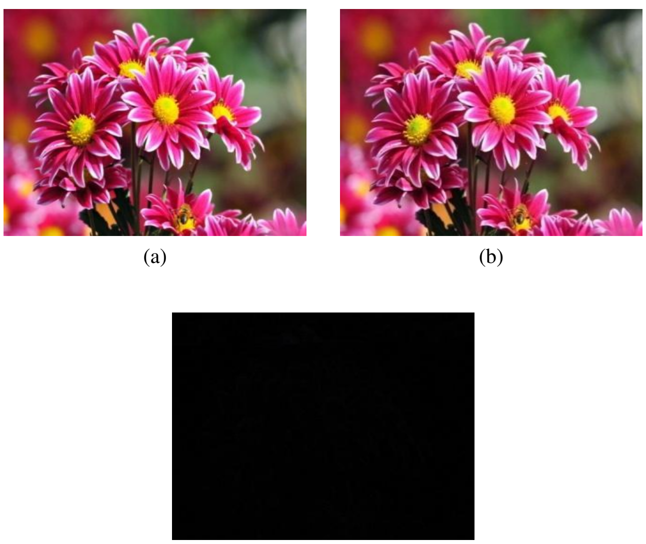


Fig. 11. A) Flower cover image, B) Flower cover image after embedding the secret message, C) The difference image between A and B.

Fig. 12 (c) and is equal to 0. The MSE is 4.86, and the PSNR is 41.36.

***Example Four**:* Using the the girl image Fig. 13 (a) as a cover image, the number of non-black points that represent the difference between Fig. 13 (a) and Fig. 13 (b) is illustrated in Fig. 13 (c) and is equal to 0. The MSE is 6.45 and the PSNR is 40.07.

The last four examples show that the difference between the cover image and the cover image containing the secret message is hard to be detected by human eyes. However, in our proposed algorithm we assume that the attacker cannot obtain the original image. Thus, finding the locations of the secret message cannot be detected without the use of SCBA or the secret message.

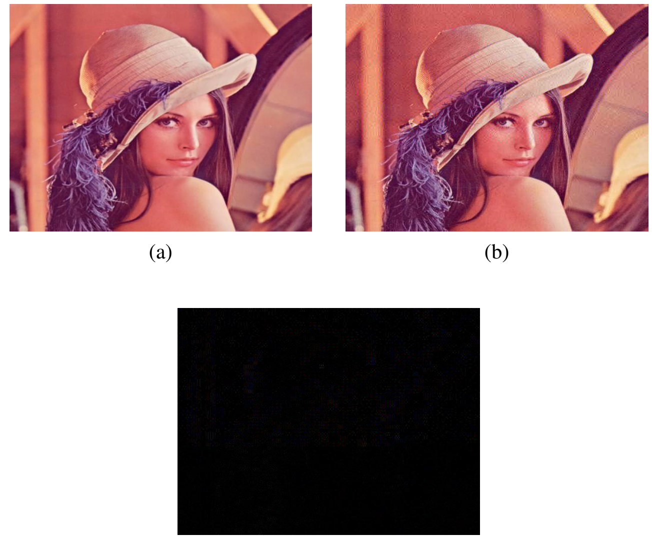


Fig. 12. A) Lena cover image, B) Lena cover image after embedding the secret message, C) The difference image between A and B.

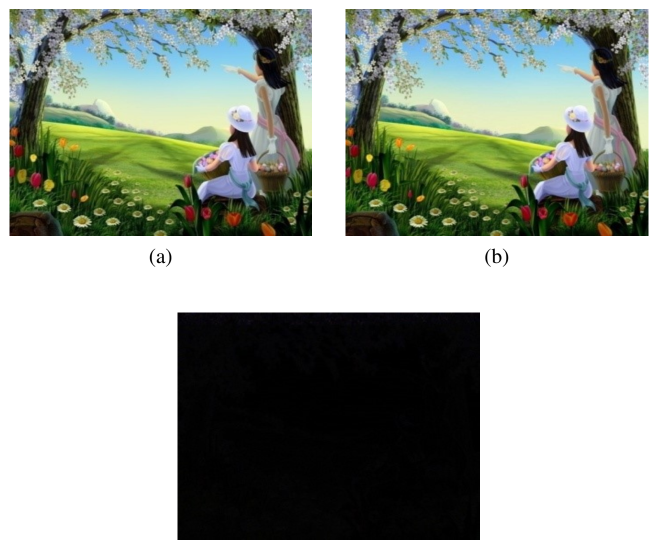


Fig. 13. A) Lena cover image, B) Lena cover image after embedding the secret message, C) The difference image between A and B.

VI. Conclusion

In this paper, we proposed a new technique to hide data in JPEG images using zigzag coefficients array to embed secret messages. The aim of using this method is to insert a secret message within a JPEG image without affecting the image quality and compression ratio. Therefore, nobody can detect or reveal the secret message hidden in the image. In this paper, we proposed two methods to embed the secret data using three sizes of data chunks that will be embedded within the zigzag coefficients array. Our experiments show that our proposed method one utilizing 3-bits data size outperform the other proposed second method in term of secret message space availability and the minimal difference after embedding the secret message. Based on our proposed selected method, we can utilize the use of almost 40% to 50% out of the total number of blocks to hide the secret message. The sender and the receiver can either share the SCBA or the private key to embed or reveal the secret message securely. We implemented the concept of shuffling the location of the embedded data using a chaotic based random number generator for each embedding process to increase the security level in our proposed algorithm. The algorithm was implemented entirely using JAVA libraries to produce an application that is used to run our experiments and to evaluate the proposed algorithm performance.

ACKNOWLEDGMENT

This work made possible by financial support from Applied Science Private University in Amman, Jordan.

REFERENCES

[1] R. Gupta, S. Gupta, A. Singhal, Importance and techniques of information hiding: a review, *International Journal of Computer Trends and Technology (IJCTT)*, vol. 33, pp. 260–265, Mar 2014.

[2] W. Bender, D. Gruhl, N. Morimoto, A. Lu, Techniques for data hiding. *IBM System Journal*, vol. 35, pp. 313–336, Feb 1996.

[3] Z. AL-Ani, A. Zaidan, B. Zaidan, H. Alanazi, "Overview: Main Fundamentals for Steganography", *Journal of Computing*, Vol. 2, pp. 158-165, Mar 2010.

[4] M. Jain, S. Lenka, "A Review of Digital Image Steganography using LSB and LSB Array", *International Journal of Applied Engineering Research*, Vol. 11, pp. 1820-1824, 2016.

[5] P. Mahajan, A. Koul, "CEET: A Compressed Encrypted and Embedded Technique for Digital Image Steganography", *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 16, pp. 44-52, Apr 2014.

[6] U. Shankar, "Image Compression Techniques, *International Journal of Information Technology and Knowledge Management*, Vol. 2, pp. 265-269, Dec 2010.

[7] A. Raid, W. Khedr, M. El-dosuky, W. Ahmed, "Jpeg Image Compression Using Discrete Cosine Transform-A Survey", *International Journal of Computer Science and Engineering Survey (IJCSES)*, Vol. 5, pp. 39-47, Apr 2014.

[8] H. Lu, X. Shy, Y. Shit, *et al.*, Watermark embedding in DC components of DCT for binary images, *In Proc. of IEEE Workshop on Multimedia Signal Processing*, St.Thomas, VI, USA, Dec 2002, pp. 300-303.

[9] M. Kaur, Robust watermarking into the Color Models based on the Synchronization Template, *In Proc. of International Conference on Information and Multimedia Technology*, Jeju Island, South Korea, Dec 2009, pp. 296-300.

[10] M. Lafta, "Image Watermarking based on Huffman Coding and Laplace Sharpening", *Journal of the college of education for women*, vol. 22, pp. 173-184, 2011.

[11] S. Sujatha, M. Sathik, "A Novel DWT Based Blind Watermarking for Image Authentication", *International Journal of Network Security*, vol. 14, pp. 223-228, July 2012.

[12] R. Preda, "Semi-fragile watermark for image authentication with sensitive tamper localization in the wavelet domain", *Journal of the international Measurement Confederation (IMEKO)*, vol. 46, pp. 367-373, 2013.

[13] M. Yesilyurt, Y. Yalman, A. Ozcerit, "A new DCT based watermarking method using luminance component", *Elektronika ir Elektrotechnika*, vol. 19, pp. 47-52, 2013.

[14] M. Khan, M. Rahman, M. Sarker, "Digital Watermarking for Image Authentication Based on Combined DCT, DWT and SVD Transformation", *International Journal of Computer Science Issues*, vol. 10, pp. 223-230, May 2013.

[15] J. Jeswani, T. Sarode, "Improved Blind Color Image Watermarking using DCT in RGB Color Space", *International Journal of Computer Applications (IJCA)*, vol. 92, pp. 50-56, Apr 2014.

[16] M. Mundher, D. Muhamad, A. Rehman, *et al.*, "Digital watermarking for images security using discrete slantlet transform", *International Journal of Applied Mathematics and Information Sciences*, vol. 8, pp. 2823-2830, Jan 2014.

[17] M. Haribabu, H. Bindu, V. Swamy, "A Secure and Invisible Image Watermarking Scheme Based on Wavelet Transform in HSI Color Space", *Computer Science Journal*, vol. 93, pp. 462-468, 2016.

[18] M. Ghebleh, A. Kanso, "A robust chaotic algorithm for digital image steganography", *Communications in Nonlinear Science and Numerical Simulation*, vol. 19, pp. 1898-1907, June 2014.

[19] J. Mazumder, K. Hemachandran, "Color Image Steganography Using Discrete Wavelet Transformation and Optimized Message Distribution Method", *International Journal of Computer Sciences and Engineering*, vol. 2, pp. 90-100, July 2014.

[20] G. Swain, "Digital image steganography using variable length group of bits substitution", *Procedia Computer Science*, vol. 85, pp. 31-38, 2016.

[21] P. Das, S. Kushwaha, M. Chakraborty, "Multiple embedding secret key image steganography using LSB substitution and Arnold Transform", *in Proc. 2nd International Conference on Electronics and Communication Systems (ICECS)*, 2015, pp. 845-849.

[22] S. Kumar, S. Kumari, S. Patro, *et al.*, "Image Steganography using Index based Chaotic Mapping", *IJCA Proc. on International Conference on Distributed Computing and Internet Technology (ICDCIT)*, Jan 2015, pp. 1-4.

[23] K. Bhaskar, M. Bakale, P. Chaure, P. Shirke, "Image Steganography for data hiding Using Huffman code, Zigzag and OPAP", *International Journal of Emerging Trends and Technology in Computer Science (IJETTCS)*, Vol. 4, pp. 91-93, Dec 2015.

[24] F. Almasalha, R. Hasimoto-Beltran, A. Khokhar, "Encryption of Entropy-Coded Video Compression Using Coupled Chaotic Maps", *Entropy*, Vol. 16, pp. 5575-5600, Oct 2014.