ITRU

Faculty of Information Technology

Proceedings
of

ITRU RESEARCH SYMPOSIUM 2020

"TOWARDS THE NEW DIGITAL ENLIGHTENMENT"

Co-Hosted with ICITR 2020

Information Technology Research Unit
Faculty of Information Technology
University of Moratuwa

TABLE OF CONTENTS PAGE

Automatic Labelling & Classification for Research Papers of Wildlife in Sri Lanka

Premisha Premananthan
Department of Computing &
Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
ppremisha@std.appsc.sab.ac.lk

Kumara BTGS
Department of Computing &
Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
kumara@appsc.sab.ac.lk

Enoka P Kudavidanage
Department of Natural Resources
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
enoka@appsc.sab.ac.lk

Banujan Kuhaneswaran
Department of Computing &
Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
bhakuha@appsc.sab.ac.lk

***Abstract*— Sri Lanka, is one of the global biodiversity hotspots, which contains a larger variety of fauna and flora. But nowadays Sri Lankan wildlife faced many issues because of poor management and policies to protect wildlife. The lack of technical & research support leads many researchers to retreat to select wildlife as their domain of study. Wildlife research results should be integrated into data-driven decisions on conservation and management, but the existing contribution is not sufficient. This study demonstrates a novel approach to data mining to find hidden keywords and automated labeling for past research work in this area. To model topics and extract the main keywords, we used the Latent Dirichlet Allocation (LDA) algorithms. Using the Topic Modeling performance, an ontology model was also developed to describe the relationships between each keyword. We classified the research papers using the Artificial Neural Network (ANN) using ontology instances to predict the future gaps for wildlife research papers. The automatic classification and labeling will lead many researchers to find their desired research papers accurately and quickly. Our model provides 83% accuracy in this labeling and classification using past research papers on the wildlife of Sri Lanka**

Keywords— ANN, LDA, ontology, topic modeling, wildlife

I. Introduction

Wildlife is critical for the sustenance of life on earth. Biodiversity conservation is crucial to preserving a stable global ecological balance.

Sri Lanka is a global biodiversity hotspot consisting of a large variety of fauna and flora. It is one of the main sources of income generation through tourism and other means. The diversity of ecosystems is primarily due to its topographical and climatic heterogeneity, as well as its coastal effect [1]. This rich biodiversity is threatened due to unplanned land use, pollution, overexploitation, etc.

Data from wildlife research can contribute to a large extent is proper conservation and management. However, there is a gap between research and application. Most of the existing research work is not converted into applications while there are many data gaps. Limited numbers of researchers are focusing on the actual research needs from conservation. The selection of research topics is often not compatible with the actual research needs due to multiple reasons. This is a disheartening scenario as there are plenty of opportunities for such work. Inadequate knowledge of the existing research and its applicability, inadequate use of technology, and the inability to locate some research are some of the contributing factors. Other than the research published in a known journal, some past research information available online cannot be found properly because they belong to conventional archives, unfortunately.

Increasing public awareness of the values of wildlife and the consequences of losing this heritage can assist conservation to a large extent. To achieve this, we have to simplify the gap between the public and the accessibility to information on wildlife. Technology can play a major role in filling the gap between them. Mostly wildlife studies aimed to understand species diversity, behavior, and habitat use, and ecology, the role of wildlife in disease transmission, species conservation, population management, and methods to control threats to diversity.

In our study, we focus on analyzing past research papers using data mining techniques to give potential research ideas to the future. To fill the data demands for conservation our solution focuses mainly on semi-automating the finding of research gaps via abstract analysis. Finally, the model includes the most commonly used keywords and question top. This will be an important milestone for researchers as well as wildlife activists to give tips on recent problems that need a solution urgently.

Our research mainly focuses to resolve the inadequate application of wildlife research and technologies in the decision-making process. This paper is arranged as follows. In Section II, we discussed the background study as a literature review. In section III, we describe the core theories used by the proposed methodology, and we discuss the results of the study Section III. In Section IV, we discuss the conclusion of our research and predict areas for future works to improve.

II. Literature Review

From a technical point of view, previous works[2][3] have shown hierarchical relationship-based latent Dirichlet allocation (hrLDA), hierarchical topics data-driven model to retrieve terminology ontology from a large number of amalgamated papers, has concentrated on many previous works to come up with a novel idea to solve past problems about automatic classification and labeling. In comparison to conventional topic models, instead of unigrams, hrLDA relies on noun phrases, deals with pattern and text structures, and enhances topic hierarchies with topic relations. Via a series of analyses, we found the Excellency of hrLDA over current topic models, especially for the construction of hierarchy.

Another paper [4] has shown an efficient approach for training a neural network model to measure moving objects in a video. For the network to simultaneously prepare a named dataset for the first one, object recognition, tracking, and counting are required item identification, pursuing and counting are threats for efficient Intelligence Transportation Systems (ITS) that used to decrease congestion and detect traffic offenders on highways and in cities. The labeled data creation to training a neural network is one of the essential prerequisites for the successful implementation of supervised machine learning.

Medical text classification is considered a special variety of text classification. Health history and medical literature are kept in health documentation, these are the vital instruments for clinical information. In this paper, a unified neural network method [5] is proposed. In the sentence representation, the convolutional layer extracts characteristics from the phrase, and both the previous and subsequent sentence characteristics are accessed by a bidirectional gated recurrent unit (BIGRU). A phrase representation scheme with the essential word weights is used. The process used BIGRU to encode the sentences received in the phrase representation and decode them so that meaningful phrase weights are achieved through the attention procedure in document presentation. A medical group of text is obtained via a classifier. The experimental examinations are performed in four medical text data sets, including two medical history datasets, two medical records.

Another study [6] has shown the concept of subject relationship and leveraging knowledge theory with the probabilistic topic models studied, two algorithms were proposed for learning terminological ontologies. Experiments were carried out with various model parameters, and the domain experts evaluated studied ontology statements. They had also compared the outcomes of our approach on the same dataset with two existing concept hierarchy learning methods. The analysis shows that in terms of recall and accuracy tests, our approach outperforms other techniques. The level of accuracy of the learned ontology is adequate for it to be deployed in digital libraries for browsing, navigation, and information search and retrieval purposes.

We found another study [7], combining LDA and ontology, to overcome the weakness of the LDA labeling problem. This research used databases from an online news portal for 50 news documents. The experiment found the best word count representation for each subject to make the right mark name for the subject. "The ontological method used in this study is based on the field dictionary “Kamus Besar Bahasa Indonesian (KBBI)”. In the measurements on a label with a specific word representation, which has more than 41 percent of kappa meaning, Cohen's kappa’s coefficient is used to determine the trademark reliability based on the approval of two experts in the language field. The results of this research show that the highest kappa value is 100 percent value in the five words representation of each subject, while the highest mean significance rate is 80 percent value in the 5 words and 30 words representation of each subject.

A paper described a proposed neural network classificatory of propagation [8], which executed cross-validations in the Neural Network. Also, it solved the original neural network classification problem. Less training time reduced the accuracy of classification expansion. The viability of the benefits of the proposed solution can be seen in 5 data sets including contact lenses, CPUs, symbolic temperatures, negative data about the weather, and labor. It is shown that when the data set is larger than the CPU or the network is equipped with many hidden units. Also, it reduced training time more than 10 times faster compared to the output of the neural network model by CPUs, symbolic temperatures, negative data on weather, and labor. This happens when the data set is larger than the CPU or the network is equipped with many unknown units, which is the only missing attributes. On average, the accuracy of the proposed Neural Network for touch lances was about 0.3 percent lower than that of the original Neural Network. To ensure that there can be many ideas and solutions that can be transferred to various classification paradigms, this algorithm is independent of these data sets.

In LDA there are inefficiencies in automatically labeling each paper separately, so some ideas for automatic labeling have been seen in previous RNN research. Artificial neural networks [9] have been suggested, and recurrent neural networks are strongly given efficient results which used in text classification because of their natural syntax structure, which is most suitable for the processing of natural language. But since there is a problem with recurrent neural networks, for example, the model is vulnerable to gradient disappearance or gradient explosion when the duration of the text series is too long.

So we plan to come up with a full and full technology-based solution. Then the topic modeling, there were plenty of studies to automatically label but they mostly worked with a small number of texts, not like our study. So this is quite a challenge for us. After that in the LDA part, we found many inefficiencies while connecting with main topic classes. So we came up with prior studies on Ontology. The ontology part of our research also gave trouble because in our case we got more than one label to a research paper. So we seek for automatic Classification Finally we got the ANN model to solve that problem.

So we have to vary past research techniques to find our final solution. Some trending techniques are used here to improve the outputs. The final study revealed the automatic labeling model to research papers on wildlife of Sri Lank

III. Methodology

We used integrated technologies in our methodology which shows in Fig 1. This methodology developed using Artificial Neural Network, Latent Dirichlet Allocation (LDA), and Ontology in this study. The text data of the defined domain were collected and pre-processed for the input to LDA algorithms then compared with the ontology graph to label the dataset the final output. After that Artificial Neural Network was used to classify. The steps of our methodology are defined below.

A. Data Collection

We collected information about past wildlife researches in Sri Lanka from 2006 to 2019, with the aid of the Department of Natural Resources, Sabaragamuwa University of Sri Lanka, and an extreme literature survey. After that, we accessed full research papers of selected papers from each domain. We've selectively applied the title and abstract data to the CSV file from those research papers.

B. Data Cleaning

Data cleaning is the method of preparing data for review by deleting or altering data that is inaccurate, incomplete, obsolete, duplicated, or incorrectly formatted. Typically, this data is not necessary or helpful when it comes to analyzing the data because it can complicate the process or provide incorrect results.

We performed the following steps:

- Tokenization: Divide the text into sentences, and the sentences into words. Lower case the words and smooth punctuation
- Stop word removal: Delete all stop words. Natural Language Toolkit has a set of stop words modules.
- Lemmatizing: Words in the third person are shifted to first-person and verbs shifted to present from past and future tenses.
- Words are stemmed — words are reduced to their root form.

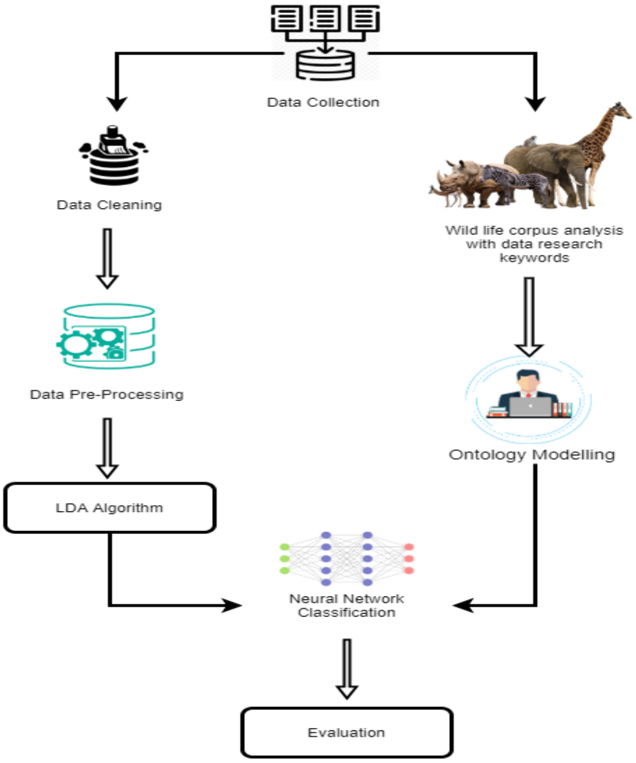


Fig. 1. Methodological Framework

C. Data Preprocessing

Data pre-processing is so important because if our data set contained mistakes, redundancies, missing values, and inconsistencies that all compromised the integrity of the set, we need to fix all those issues for a more accurate outcome [10]. We used GloVe for the preprocessing works of our text data. Glove stands for Global Vector for Word representation which provides a high level preprocessing vocabulary close to the pre-trained embedding [11]. So we can get preprocessing to result in tokens that are mostly covered by word vectors.

D. Topic Modelling-LDA

LDA helped adapt the textual data into a format that could act as an input to the LDA model for training. We began by converting the documents to a simple representation of the vectors as a group of words called Bag of Words (BOW) [12]. First, we translated a list of titles into vector lists, all with vocabulary-capable lengths.

Topic modeling is one of the unsupervised methods in trending. In other words, it is a text mining strategy in which is used to form topics for subjects or themes of documents can be derived from a broader set of documents [13]. LDA, one of the most famous & efficient modeling techniques, is a similarity model of a corpus-based on Bayesian models. This is often considered a probabilistic extension of Latent Semantic Analysis (LSA). The LDA's basic idea is that each document has a word distribution that can be defined as.

E. Ontology Modelling

Ontologies contain features such as general vocabulary, reusability, machine-readable content, as well as ordering and structuring information for the Semantic Web application, enabling agent interaction, and semantic searching [14]. Automated learning is the problem in ontology engineering, such as the lack of a fully automated approach to shape ontology using machine learning techniques from a text corpus or dataset of various topics.

The ontology model was finalized using protégé tools, which are the most popular tool of ontology visualization [15]. The Protégé 5.5.0 tool is being applied for further development in various disciplines for a better understanding of knowledge with the aid of domain professionals in the wildlife.

F. Neural Network Classification

We used the Recurrent Neural Network (RNN) classification to train and test the model of our automatic labeling process. Here to train our model, we used Long Short Term Memory (LSTM). Three sections or layers, which are the input layer, the secret / intermediate layer, and the output layer, may characterize the neural network. The input layer is used to receive the outer field's input signals. It is made up of neurons that go to the secret layer. The supervised based method is used in Neural Network, so there is a response or output to the input provided to the neural network. The neural network processed input values and weights which took as input from the input layer and then goes to the hidden layer where the weights are summarized by the algorithm and the results are mapped to the proper output layer units.

For the training, we used 70% of our data set and testing 30% of the data. We input the abstract of each paper and train through ontology classes. We selectively trained for 7 major classes.

G. Evaluation

In our research, we used the output analysis method which is used to assess the outcomes of the research concerning its objectives. A new approach to automatic ontology learning has been established, and the LDA model has been applied to generate topics by this process, and the advancement of learned ontology does not need the seed of ontology, but only the text corpus[16].

IV. Results

The results of this study were represented using abstract past research which serves as an input in Sri Lanka. We used python language for LDA implementation. The text used as input is interpreted and tokenized with the result that input nouns, adjectives, and verbs are compiled. Also, it removes all the stop words in the research papers.

The tokenized and pruned text is then subjected to the algorithm of LDA modeling. That created word sets that could collect words that are connected as word sets. These word sets are listed as separate subjects. To organize, synthesize a large corpus, and to retrieve subjects and words, the LDA model approach is used.

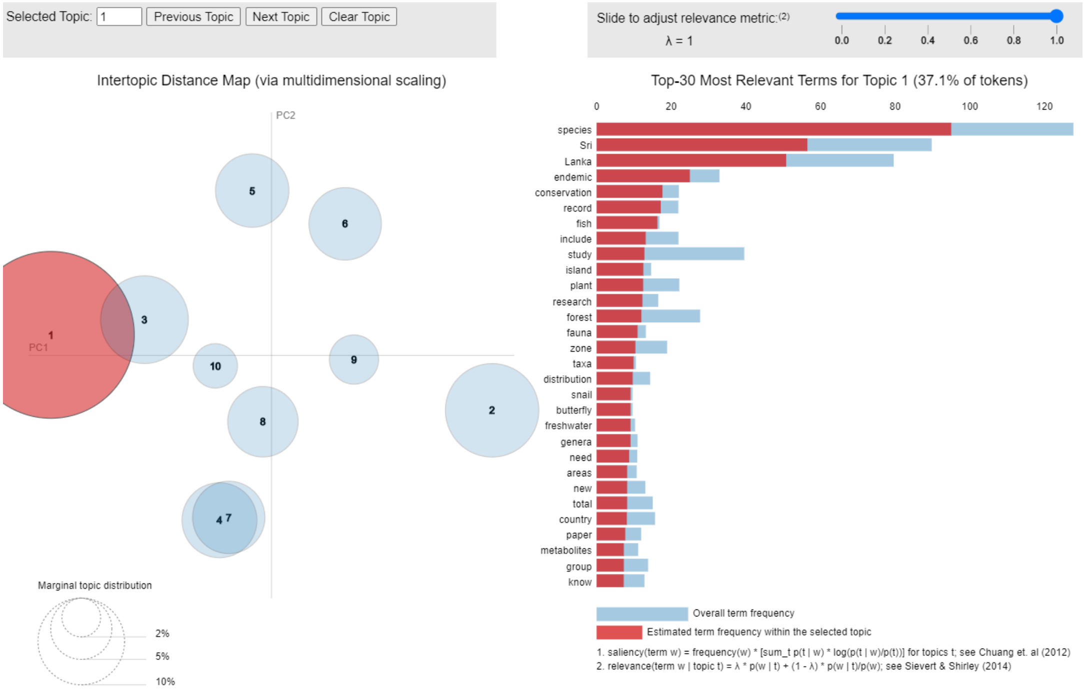
Fig. 3 is the final visualizations of the LDA model which shows the overall keyword for each research paper and the essential keyword using the pyLDAvis library in python. This output allowed the detection of hidden keywords from every abstract. To get the output of the pyLDAvis method we used the equation of saliency and relevance to accommodate the keyword distributions.

The intertropical distance map is indicated via multidimensional scaling by our LDA output. In CE literature and inter-topic distance, the top 30 salient keywords.

Saliency is used to describe the overall term frequency according to your data set. In all research papers, we found the top 30 overall terms to form a similarity cluster to form each topic. The saliency [17] is defined by the blue color bar in the given graph.

$$Saliency = frequency \times \left[sum\ p(t|w) \times \log\left(p((t|w)/p(t))\right)\right] \quad (1)$$

Where (1), t- Topic, Frequency (w) –frequency of word w, p (t|w) - conditional probability: the probabilistic which identified word w was generated by latent topic t, p (t) - the probability of topic t, sum p (t|w) - the sum of the probability of identified word w was generated by latent topic t.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 2. LDA Topic Model Output

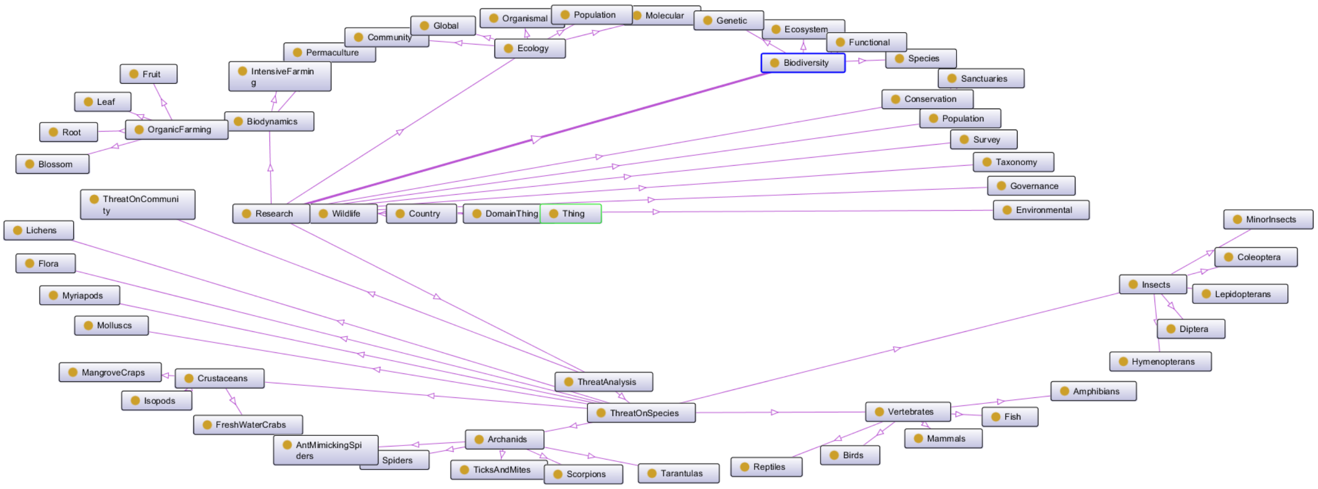


Fig 3 Ontograph Final Output

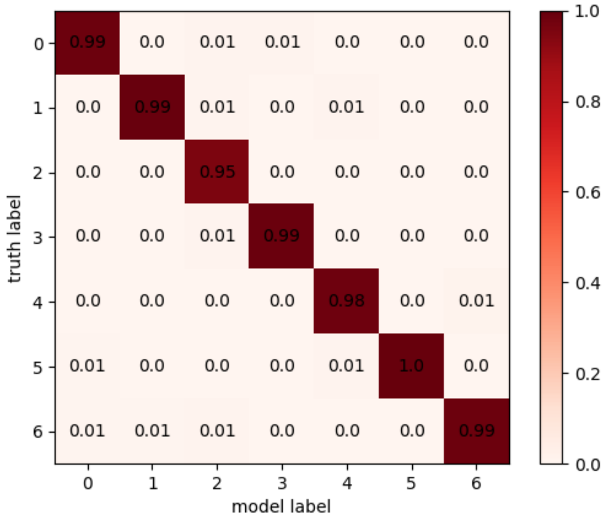
The formulation (1) defines (in a theoretical context of information Sense) the meaningfulness of specific term w, versus a randomly selected word, is for obtaining the generating subject. In case, if a word w appears in all topics, observing the word tells us nothing about the topical mixture of the document; thus the word will obtain a score of low distinctiveness.

Relevance is used to express the estimated term frequency compared to overall term frequency which means saliency in the given corpus. Relevance is used to measure the similarity between each keyword to find the final topics.

$$Relevance = \lambda * p(w|t) + (1 - \lambda) * p(w|t)/p(w) \qquad (2)$$

Where (2), λ –slide to adjust relevant metric, p (w|t) - conditional probability: the likelihood that identified word w was generated by latent topic t, p (w) –the probability of word w [18].

Using this output from LDA we compared the ontology output. Analyzed the estimated keywords and their ontology domain formation. The protégé tool used the Sri Lankan wildlife research domain ontology to be developed. The partial view of the final ontology production is shown in Fig 3.



	0	1	2	3	4	5	6
0	0.99	0.0	0.01	0.01	0.0	0.0	0.0
1	0.0	0.99	0.01	0.0	0.01	0.0	0.0
2	0.0	0.0	0.95	0.0	0.0	0.0	0.0
3	0.0	0.0	0.01	0.99	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.98	0.0	0.01
5	0.01	0.0	0.0	0.0	0.01	1.0	0.0
6	0.01	0.01	0.01	0.0	0.0	0.0	0.99

The tokenized and pruned text is then inputted into the algorithm for LDA modeling. This created word sets that could collect words that are linked to each other as word sets. Such word sets are known as topics. To organize, synthesize a large corpus, and to retrieve subjects and words, the LDA model strategy is used. The intertropical map of distances is indicated by our LDA performance through multidimensional scaling. The top 30 salient keywords in CE[19] literature and inter-topic distance.

After the LDA keyword extraction and Ontology model output, we performed the classification manually with the help of domain expert advice.

For the automatic classification part, we need the labeled data set as the input of ANN. The model trained through 3 major layers such as Embedding, LSTM, and dense layers [20]. LSTM is a kind of Recurrent ANN. It allows for the iteration of the model for efficient output. Also, the ontology model provided the best-labeled dataset to the ANN. Because learning a text like an abstract of a research paper is a little bit hard to implement. The dataset after cleaning used to vectorize because ANN only understands Float instead of text. Fig 4 has shown the vectored classification for each class. From ontology, we selected the 7 major classes. Fig 4 describes the accuracy of the ANN model through LSTM.

Fig. 3. ANN Vectorization model for classes

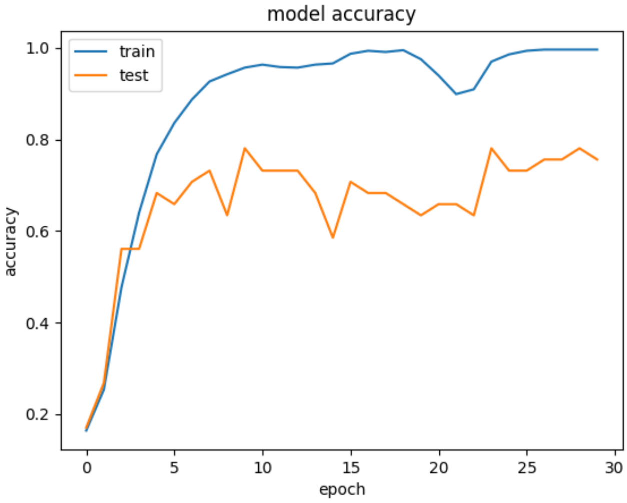


Fig. 4. ANN model accuracy graph

After that, we calculated the loss and accuracy to measure the deviations between predictions and real outputs. We used the Sri Lankan research papers for the testing of the model. In our research, the neural network model's accuracy was 80% with the less and sensitive data we achieved this number.

V. Conclusion And Future Works

In this paper, we have suggested an automatic labeling model for the research papers on the wildlife of Sri Lanka. We used a methodology which first LDA to extract the hidden keyword of each paper and after that Ontology to model the domain to label the data set with the help of domain experts. Finally, the dataset was trained and test through RNN. After the training of the test data, the model tests the testing data and it shows the accuracy of every algorithm with different features. We manually compared the results from both LDA and terminology ontology. Our study's output evaluation was 80% accurate to the overall conclusion.

This work reduced the complexity to label the research papers without any domain pre-knowledge. Using this method the hidden keyword and the relations between the keywords also identify to help future research ideas.

In this topic labeling method, there is some inefficient while ontology classification. Because there are several cross path hierarchy moves of keywords identified from LDA. So when we used ontology it collapsed the different paths into a graph. So we will use other classification methods to fully automated our methods.

Acknowledgment

We acknowledge the Department of Wildlife Conservation of Sri Lanka for the research permit (WL/3/2/60/15) granted to E. P Kudavidanage.

References

[1] L.P.Jayatissa, *Present Status of Mangroves in Sri Lanka*. 2012.

[2] X. Zhu, D. Klabjan, and P. N. Bless, "Unsupervised

terminological ontology learning based on hierarchical topic modeling," *Proc. - 2017 IEEE Int. Conf. Inf. Reuse Integr. IRI 2017*, vol. 2017-Janua, pp. 32–41, 2017, doi: 10.1109/IRI.2017.18.

[3] S. Chowdhury and J. Zhu, "Towards the ontology development for smart transportation infrastructure planning via topic modeling," *Proc. 36th Int. Symp. Autom. Robot. Constr. ISARC 2019*, no. Isarc, pp. 507–514, 2019, doi: 10.22260/isarc2019/0068.

[4] I. Namatevs, K. Sudars, and I. Polaka, "Automatic data labeling by neural networks for the counting of objects in videos," *Procedia Comput. Sci.*, vol. 149, pp. 151–158, 2019, doi: 10.1016/j.procs.2019.01.118.

[5] L. Qing, W. Linhong, and D. Xuehai, "A novel neural network-based method for medical text classification," *Futur. Internet*, vol. 11, no. 12, 2019, doi: 10.3390/FI11120255.

[6] W. Wang, P. M. Barnaghi, and A. Bargiela, "Probabilistic topic models for learning terminological ontologies," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 1028–1040, 2010, doi: 10.1109/TKDE.2009.122.

[7] R. Adhitama, R. Kusumaningrum, and R. Gernowo, "Topic labeling towards news document collection based on Latent Dirichlet Allocation and ontology," *Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017*, vol. 2018-Janua, pp. 247–251, 2018, doi: 10.1109/ICICOS.2017.8276370.

[8] M. Govindarajan and R. M. Chandrasekaran, "Classifier Based Text Mining for Neural Network," *Proc. World Acad. Sci. Eng. Technol. Vol 21*, vol. 21, pp. 200–205, 2007.

[9] S. Bhatia, J. H. Lau, and T. Baldwin, "Automatic labeling of topics with neural embeddings," *COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap.*, no. 1, pp. 953–963, 2016.

[10] N. T. R. Editors, *Technologies in Data Science and Communication*. 2019.

[11] P. M. Brennan, J. J. M. Loan, N. Watson, P. M. Bhatt, and P. A. Bodkin, "Pre-operative obesity does not predict poorer symptom control and quality of life after lumbar disc surgery," *Br. J. Neurosurg.*, vol. 31, no. 6, pp. 682–687, 2017, doi: 10.1080/02688697.2017.1354122.

[12] M. Rani, A. K. Dhar, and O. P. Vyas, "Semi-automatic terminology ontology learning based on topic modeling," *Eng. Appl. Artif. Intell.*, vol. 63, no. August, pp. 108–125, 2017, doi: 10.1016/j.engappai.2017.05.006.

[13] J. Lee, J. H. Kang, S. Jun, H. Lim, D. Jang, and S. Park, "Ensemble modeling for sustainable technology transfer," *Sustain.*, vol. 10, no. 7, 2018, doi: 10.3390/su10072278.

[14] D. Movshovitz-Attias and W. W. Cohen, "KB-LDA: Jointly learning a knowledge base of hierarchy, relations, and facts," *ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf.*, vol. 1, pp. 1449–1459, 2015, doi: 10.3115/v1/p15-1140.

[15] G. Hussein, A. L. I. Ahmed, L. Kovács, G. Hussein, and A. Ahmed, "ONTOLOGY DOMAIN MODEL FOR E-TUTORING SYSTEM," vol. 5, no. 1, pp. 37–44, 2020.

[16] Z. Lin, "Terminological ontology learning based on LDA," *2017 4th Int. Conf. Syst. Informatics, ICSAI 2017*, vol. 2018-Janua, no. Icsai, pp. 1598–1603, 2017, doi: 10.1109/ICSAI.2017.8248539.

[17] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," *Proc. Work. Adv. Vis. Interfaces AVI*, no. June, pp. 74–77, 2012, doi: 10.1145/2254556.2254572.

[18] T. R. Frasier, S. D. Petersen, L. Postma, L. Johnson, M. P. Heide-Jørgensen, and S. H. Ferguson, "Bayesian abundance estimation from genetic mark-recapture data when not all sites are sampled: an example with the bowhead whale," *bioRxiv*, vol. 22, p. 549394, 2019, doi: 10.1101/549394.

[19] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," no. June, pp. 63–70, 2015, doi: 10.3115/v1/w14-3110.

[20] B. Shishov, "Mental workload estimation on facial video using LSTM network MENTAL WORKLOAD ESTIMATION ON FACIAL VIDEO USING LSTM," no. June 2017.

A Review on Mining Software Engineering Data for Software Defect Prediction

J.P.D.Wijesekara
Faculty of Information Technology
University of Moratuwa
Sri Lanka
dulanjalijpw@gmail.com

P.G.T.P. Gunawardhana
Faculty of Information Technology
University of Moratuwa
Sri Lanka
tharin.gunawaradhana@gmail.com

Abstract— **Maintaining quality and reliability is a foremost challenge faced by software Developing professionals amid the software development process. A defective module can lead to software failures, changes in development time, and costs as well as leads to customer dissatisfaction. Usually, a Software repository is used in maintaining software for a long period with its upgrades and bug fixing. This paper forwards a literature review that describes software failures along with the consequences. So identifying these defects as earliest as possible is quite important in the software development life cycle. In identifying these defects researchers and Professionals are using different data mining techniques along with defect tracking systems to find out these defects accurately. Since these defect tracking systems and software repositories have a great deal in locating issues in software, it is very much important to improve the defects tracking systems. Defect tracking systems are the places where we can review the history of the software developments in failures as well as in successors. This paper details how to make a conversion between software repositories into active software repositories. Further, this paper details how to use data mining approaches to identify fault-prone modules in the software. This also consists of details on time prediction in defect fixing cases; post or pre-release and various metrics that we can use to predict software defects.**

Keywords—Software Defect Prediction, Data Mining Approaches, Software Metrics, Defect Tracking System

I. INTRODUCTION

Software engineering data analytics has been a major research area among software researchers and practitioners over the last decade [43], [44], [45]. Software Engineering research information and collection of historical data from previous projects have an enormous capacity to develop and improve the management of software projects. Different Research [1] has focused on how far a review of the history of the software development process or in evolution progress of software can be useful for enterprise level software development. Such studies have mainly focused on areas such as software bug prediction[5], software visualization[42], and software security[41]. In order to maintain a project among the allowed budget and the timeline which corresponds to the quality of the product, we can use the knowledge that has been discovered using different data mining approaches using the past tracks of software repositories during different software projects. In order to extract statistically important data points from a selected data set, data mining uses different approaches such as classification, clustering, and rules like association rules. Data mining is not only important in marketing, designing, identifying patterns. Etc. but also this can be used in order to enhance the software development process. Different approaches to data mining can be used along with bug tracking systems and software repositories in order to identify defects, analyse them, predict them, within testing processes as well as to enhance the understanding related to software bugs. There are already developed tools and techniques for defect prediction, identifying, correction, testing, and prevention as well. There are various tools available to carry out data mining in very large datasets and to work on analyzing them[2]. Among such research conducted, Data mining in software repositories has come into the stage with various challenges. Without obstructing the enterprise Application development activities at the development stage necessary information required for predictions can be obtained using the software intelligence which has been created using the same set of approaches and techniques used in decision making in decision support systems.

Researchers like Hassan [3] have detailed different challenges faced in these circumstances main two of them are; challenges in mining only selected important information from these software repositories and challenges in automating and extracting information in them. There are lots of challenges in mining software repositories such as defining future research dimensions, featuring success in prediction as described by the various researches[1]. Recent research conducted [5] have identified that structured mining can identify defective and defect-prone modules up to an extent. Various patterns, important contents, and useful details were able to be discovered using the historical repositories, repositories of the codes, and repositories of the run time projects which have the enamour potential in making decisions related to software projects[6]. In reviewing most of the research conclusions [7][8],it shows that information that was obtained from historical repositories or past software development projects and information taken from previous open-source projects can aid in very large software development systems.

II. MAJOR RESEARCHES

There are different research works conducted in mining research repositories to facilitate the software defect prediction in order to enhance the software development and maintenance field. This section mainly focuses about such research studies that have already been conducted in this context

A. *Mining of Software Repositories for Defect Prediction and Challenges*

Research conducted by Zimmerman et.al [5] has concluded that modules or areas that are mostly fault-prone and highly defective also can be identified with close examination using structured data mining. Defects in the software are the main concerns to reduce the quality of the product which leads to customer dissatisfaction. Defect repositories keep a continuous track record of software in its development phase where we can track the pitfalls of the software in different stages. So this is a rich information database to track the pitfalls of software in their successors and failures. Not only have the technical failures led a software project to be a failure. There are some other social concerns that can affect project failures such as negligence of estimations, impossible deadlines. Both of these technical and social factors that can lead to software project failures are broadly discussed in many types of research [4]. Further technical problems like lack of chances for growth in project requirements, least estimation approaches during development also can cause project failures. Especially in non-technical fields and also in cases where software is not the major drive like in Banks, insurance companies...etc., there is a very lack of dedicated tools for estimations, separate project managers for software projects and inadequate estimation methods will be used along with informal methods. Most commonly the major reason for schedule changes, overruns of cost and resources, and immediate withdrawal of large-scale projects is a load of bugs and defects which hold the operations. Identifying defect prone modules is a crucial but difficult task for the management. It is very hard to forecast the time and the resources needed to correct once a defect or a fault is diagnosed at the deployment. It is said that it is very much important to get the guidance through patterns identified and analysis taken from open source projects taken from historical repositories where they can guide the development in large scale systems [7][8]. There are some other researches [8] conducted in order to reveal whether there is a chance in converting a passive repository to an active software repository which itself can be used to forecast, planning process, and testing different artifacts of software projects. Through these means, developers can reuse software codes which will eventually lead to fewer faults, less time required for developing, no redevelopment problems, lesser error correction whereas finally all of these leads to better quality software at last. Historical or traditional repositories mostly contain unstructured and unlabelled texts which steer the difficulties in analysing. But Thomas [17] has pointed out the use of Latent Dirichlet Allocation for automatic identification of structures as a solution to this. This study has discussed the challenges faced in the application of different models for software repositories.

B. *Mining with Defect Tracking Systems for Defect Prediction and Challenges*

Software bugs result in pausing the project schedules as well as an effect on software quality. There are two main types of software defects as pre-release and post-release defects. The defects that have been identified before the deployment among the development stages and testing duration are called pre-release defects and defects that have been identified after the deployment of the software or system during clients' work on the system are called post-release defects. It has been discovered by Schroter in the paper [36] these post-release defects can be identified and forecasted beforehand after carefully investigating the relationships and patterns in the information present in historical repositories. In supporting this idea Nagappan and Ball in the papers [16] have discussed that these post-release defects can be revealed by the dependencies having across different components of a project. In general, the quality of the software is taken into account considering the post-release bugs and bugs that have not been closed down or isolated for years. There is some sort of dormant bugs in software that has not been disclosed or emerged after years throughout a few releases. There are researches conducted on these factors as well. Chen Hasun is such a researcher who has conducted research [28] on dormant and non-dormant bugs in software. Particular research has discovered that the dormant defects are quickly fixable since they are the results of being isolated and as a result of spontaneous control flows.

In order to reduce the cost overruns in software development developers need to develop new prediction models for software defect prediction and choose reliable software metrics for the predictions. So that these metrics can be used as inputs to mining methods in order to draw out useful patterns that can be used to categorize and classify these software defects [3]. Currently, it is a really tough task to forecast the resources required in defect correction. But specifically, Cathrin [14], explained a method leads to stable scheduled releases where the project managers can exactly predict the time required for bug fixing, guide in assigning defect fixers, effort estimations. In that paper, they experiment with the proposed approach using the JIRA bug tracking system with the use of Project JBoss. There they have observed that more or less seven hours are required to issue reporting in the particular project when compared with the actual time required to work on reporting defects. Simultaneously most of the researchers have the idea of [16] rather than traditional prediction models, just in time prediction models are practical in predictions.

Currently, software defect prediction is a highly concerned area in software engineering (SE). There are researches [9] conducted with the aim to identify how a particular module is defected prone and some other module is not. This paper has studied how previous repository data can be used to identify different properties in future software. why some modules distinguished defect prone levels from others. There they identified that bug tracking systems are vital to grabbing previous failure information than any other source. But not always these bug tracking systems contain all the information like what, how, when, where, and who assigned to fix these defects. There are many techniques [30][31][32] that were developed to correct these tracked defects.

Deeper into the analysis of defect tracking systems Zimmerman in one of the research [10] conducted an analysis using software developers and users of software like MOZILLA, ECLIPSE, APACHE. According to the analysis, they have collected bug reports from both developers and the users where they found that there is a big difference in the bug reporting with differences in reporting length, the format they used, the description they provided, attachments provided along with the report. With the observations, they could conclude that bug reporting makes the defect fixing life much easier except for a few serious bugs that arose and most of the defects reported were duplicated. It costs only lesser time than usual since most of the developers' bug reports contained enough details to the fixer.[26]

Further in deeper research [27] some of the researchers also have analysed the reasons to reopen the issues and evaluated the previous bug fixing process. In this research, a reconstruction of statistical models has been taken statistically in order to show the effect from the metrics of reopened defects with respect to how the bug was identified as well as concerning the reputation of the founder. So all of these hints at the major role played by defect tracking systems in software engineering predictions. This is a major tool and a medium which holds and interacts with the bond between developers and the users of a system. But a weaker defect tracking system ends in maintaining a weaker relationship like submission of duplicated defects which cost more time in managers' life [26]. Researchers have suggested that in order to minimize the time spent on a defect that was tracked using a bug tracking should be updated in order to get the very current status of the particular issue to the developers[24]. Moreover, this research points out more information through a study on managing an interacting defect tracking system whereas users get some defined questions that are relevant to the reporting defect. This would get the developers that complete information on the defect that they are assigned to fix as well as this will reduce the time in understanding the defect itself. Silvia in the research [23] has shown that different options to enhance this bug tracking system. The study has highlighted the importance of users' interactive participation in the defect fixing process.

Ostrand [20] has presented a study on the automation of this defect prediction process. In this study, they have discussed the development of automated tools that can forecast issues in a system without the involvement of testers and expert judgments. In doing this they also have proposed an updated bug tracking system that can collect more information on defects which are vital for developers in fixing it. There are seven suggestions forward to an updated bug tracking system in research conducted by Sascha [13]. Usually, a bug is fixed whenever the system malfunction or not functioning as required. There is no observation on how to design the debugging process. Emerson [22] in the research has discussed the ways that a defect can be fixed. At the end of their surveys, they have developed a view on multidimensional design space for defect fixing.

There is a study carried out on a time forecasting model which is for defect prediction using Naïve Bayes. This Defect Fix Time Prediction model can forecast the time required to fix a particular defect using a very huge dataset to analyse and forecast the time required to fix for new bugs. Using these results of time a typical developer or a project manager can prioritize the work as needed [21]. Among them, some defects only need a few minutes of work as well as there are rare cases like years of time required for the fixing process. For such outlining cases in his study, he has proposed a filtering technique to keep the model quality. Except for outliers after filtering they have reached acceptable and accurate experimental results of time for bug fixing. Philip [25] has studied the reassignment of the bug fixing process. And they found that it is not always harmful. There are five main reasons for reassignment as per the study such as in cases where a proper fix is difficult to decide, problems within root cause, problems in workload assigning, and balance. In his study, he has mentioned that there is only a small space for effective reassignment in defect tracking systems.

C. Software Metrics for Defect Prediction and Associated Challenges.

In order to improve the quality of defect prediction models, different metrics for software were introduced to be used with the defect prediction models. There are investigations carried out on the usage of an anti-pattern for this [19]. In this investigation, they have revealed that different metrics built on anti-patterns can be used to increase the accuracy of a prediction model. A research paper published by Fukushima [16] has revealed that a JUST IN TIME prediction model is much more effective than previous prediction techniques used. But this Just in time model needs a huge amount of training data. Further they have revealed that rather than projects having less historical data, projects that have used data even from cross-projected to learn to have a better prediction result. But in the case of the Just In Time model it requires a huge amount of historical data to train the model to get a reliable accurate prediction result. Moreover, they found that this model output prediction results that are robust, very stable to noisy data, and very accurate data when they used the Random Forest algorithm with Just in Time cross-project data. Different researchers have studied various software metrics. Some of such metrics are Metrics of complexity, Object-Oriented Metrics, Process metrics, and traditional metrics. Papers published by Zimmerman [29] have discussed the traditional and complexity metrics for software defect prediction. As a summary, most of the research finalized that than most of the popular traditional metrics, complexity metrics have better results in prediction [15]. Hassan in research concluded that Complexity metrics can predict future faults in a much better way for large software systems in contrast to using a prior modification of prior faults.

In considering Object-oriented metrics, it is quite important for defect density prediction [33]. Researchers like Subramanyam [34] have shown that there is a very important relationship between defects and Object-Oriented metrics. Using Relative code churn there are studies that predicted defect densities perfectly [17]. Further in Zimmerman's [19] research papers have concluded that complexity metrics are a better fit for prediction. At the same time there are many other choices for defect prediction than the complexity metrics. In contrast to Zimmerman's, Danijel in his studies showed that rather than complexity metrics, Object-Oriented metrics performed better prediction results, and the model which used the process metrics performed worse in the evaluation with post-release software. Further, it mentions that object-oriented metrics are used twice in comparison to the process and traditional metrics. Among the object metrics, the CK metric is the widely used kind and Not all of the CK metrics are performing equally. The best metrics out of the CK metrics are CBO, WMC, and RFC. LCOM from them can be used in the least prelease cases but not too successful in detecting defects.

Always Prevention is much better than cure. Sakthi [11] also has published in his study that defect prevention is kind of a good investment to maintain the software quality. They discovered that most of the defects are found during the development stage. Using the identified and classified patterns from the set of defined projects the developers can improve the quality of the software which are yet under development. Moreover, in their study, they emphasize that previous defect prevention was about the prediction of defects, size of the project team, and required debugging time. A "Bug Tracking system" tool for track defects was introduced by Chenbin [12]. This was able to track the defects with the least cost and also with high accuracy. The defect classifying approach which was proposed by IBM was called to be the best method to identify defects which are called "Orthogonal Defect classification". There are two main sections in the Orthogonal Defect Classification as opener and closer sections. The opener section is the time a defect is detected, and the closer section is the time of bug fixed. The main aim of the defect prevention is to find out the reason for the defect and stop it being recurred, in order to reduce the time and other resources spent on the project, which on the other hand increase customer satisfaction, to reduce the reassignment of the work thereby moderately increase the product quality.

III. COMPARISON OF RELEVANT RESEARCHES

Following table is a comparison among a set of similar research articles, journal articles and the conference papers that have conducted the research on mining software repositories and the use of this in the context of software defect prediction and related fields in software engineering.

TABLE I: Comparison of Relevant Researches

Paper Title	Area of study	Remarks
Mining software engineering data [1]	Examine How appropriately use the SE data for project management within allocated resources	Latest practices of mining acquired by different practitioners and researchers and discuss the associated challenges.
The Future of Mining Software Engineering Data [3]	Investigate methods and challenges the software repositories in order to evolve software intelligence to facilitate future researchers.	Different practices in software intelligence (SI) and identifies future directions in research with mining data in SE. Demonstrate the capabilities of SI in automating development and management of projects
Social and Technical Reasons for Software Project Failures [4]	Reasons for project failure; assume project failures can be reduced with the pre-known reasons which risk the project a failure.	Emphasize and categorise reasons for project failures with root cause. How to mitigate risks in failures with BRIDGE processes leads to the project success rate in software development projects.
Predicting Bugs from History [5]	Identify the properties (Problem Domain, Complexity of codes, historical data) correlations with bugs in past records with the defect rates, to predict future defect rates and costs	structured Mining can identify defective/not modules. Fine-grained methods and more advanced metrics results in better results in predicting
The Road Ahead for Mining Software Repositories [6]	To identify the cross-links, exist with data present in software repositories to discover the actionable information in the decision-making process	Highlights specific supports (identifying codes risk in failures and future costs,) results in complex changes in mining techniques used with historical changes and bugs existed.
CVSSearch: Searching through source code using CVS comments [7]	To present an algorithm to source code fragments using version control systems	Develop an algorithm named CVSSearch; featured in search engine for open source communities
Guest Editors' Introduction : Special Issue on Mining Software Repositories [8]	Potential uses of repositories to improve the design, reusability, and maintenance of the software and to empirically validate these findings.	Tools like version controlling systems, bug tracking systems were identified to have issues in their records. Validate and built theories, approaches newly founded with current records.
Predicting Bugs from History [9]	Studied how previous repository data can be used to identify different properties in future software	Identify why some modules can be distinguish as defect prone as others and what makes them identify.
What Makes a Good Bug Report ? [10]	Identify what makes a quality defect report to use in future research works with the identified anomalies present in the current bug reports present in selected tools.	Prototype called CUEZILLA to enhance the prevailing defect tracking tools, which predict quality of defect reports based on the details present in these bug repositories
Defect Analysis and Prevention for Software Process Quality Improvement [11]	Examining the information in bugs identified in previous projects and to know how to avoid some issues by improving the existing approaches or using new approaches.	Elaborates on how to initiate the measures to eliminate bugs in newer projects to reduce similar bugs. Initiating level first ODC for classification of bugs with better bug elimination measures.
Defect Tracking System Based on Orthogonal Defect Classification [12]	To provide another step for bug measuring with a newer method in analysing software bugs	New approach with Orthogonal classification for defect analysis. Has better accuracy compared to the traditional method as well as higher popularity, and a lower cost.
Towards the Next Generation of Bug Tracking Systems [13]	Survey to examine information requirement and common issues with reporting defects in developers and other users	Suggestions for new defect tracking systems are provided. There are seven suggestions forward to an updated bug tracking system. They state that there is no empirical on how to design the debugging process
Predicting Effort to Fix Software Bugs [14]	How to forecast the resources (time and effort)required in defect correction or identifying beforehand	Methodology in predicting bug fixing efforts for a defect. Automated model estimate effort to stable release schedules..
Predicting Faults Using the Complexity of Code Changes [15]	To more precisely examine how a change in a complex code in a system will affect the product.	Models quantify complexity using the historical changes in code instead of the attributes of the source code. (change complexity model).
An Empirical Study of Just-in-Time Defect Prediction using Cross-Project Models [16]	Different methodologies in creating JIT bug prediction models, detect changes in source code that have a higher risk of being defective.	The model has better performances when the model is trained with Correlated training data. Higher training is required for the model.
Mining Software	Examine the challenges in the	Automatic tool for detecting textual

Repositories Using Topic Models [17]	application of the topic model in repositories	repositories by using static topic model. capability of Structures in defect prediction and recovery traceability
Software defect prediction models for quality improvement: a literature study [18]	The effort to minimize and control defects in software engineering in using different models to maintain their accuracy.	Any model in defect prediction relies on the volume, size, nature of defect data, and predictors'/ classifiers' accuracy.
Predicting Bugs Using Antipatterns [19]	Examine the abilities of anti-patterns for bug prediction	Metrics built on anti-patterns can use to increase the accuracy of a prediction model.
A Tool for Mining Defect-Tracking Systems to Predict Fault-Prone Files. [20]	Automate defect prediction process to work without statistical knowledge for any user	Automated framework to a model of defect prediction. How to detect modification requests that contain the defect is yet a problem.
Improving Bug Fix-Time Prediction Model by Filtering out Outliers [21]	Study carried out on a time forecasting model which is for defect prediction using Naïve Bayes	Defect Fix Time Prediction using a huge dataset to analyse and forecast on new bugs. Used a filtering technique of mild outlier to remove outliers.
The Design of Bug Fixes [22]	Ways a defect can be fixed alternatively and perform an investigation on how engineers do selection about how to fix a bug.	Multidimensional design for defect fixing based on the identified factors that affect the selection of defect fixing method with few implications for localization.
Information Needs in Bug Reports: Improving Cooperation Between Developers and Users [23]	How far a bug report affect for quality maintenance of a system and how important the different users' engagement with bug reporting in relevant tools	showed different options to enhance bug tracking system. this release that users engagement with up-to-date information in bug reporting is viable in defect fixing
Improving Bug Tracking Systems bug tracking [24]	Address the concerns in defect tracking systems to elicit all the required information.	This initial work presents a system including context-sensitive questions to take out the required information to bug fixing process
" Not My Bug !" and Other Reason For Software Bug Report. Reassignments [25]	Reassignment of bug fixing based on the optimal capabilities of the developers in handling the defect report.	Reassignment of bug fixing process. There are five main reasons for reassignment as per the study.
Duplicate Bug Reports Considered Harmful . . . Really? [26]	This study work on managing the duplicate bug reports which have a different collection of related information on the same defect	Weaker defect tracking system ends in maintaining a weaker a relationship like submission of duplicated defects. Seven suggestions to mitigate this issue with a prototype
Characterizing and Predicting Which Bugs Get Reopened [27]	Study on the effectiveness of defect fixes by considering the stability of the defect tracking system concerning the rate of reopened defects.	Reconstruction of statistical models to show the effect from the metrics of reopened defects concerning how the bug was identified
An Empirical Study of Dormant Bugs Categories and Subject Descriptors [28]	Measure the effect of software quality related to the post-release defects	Use of bugs reported is misleading in judging the quality.
Predicting Defects for Eclipse [29]	Investigate how bugs originate by connecting failures to components; traditional and complexity metrics for software defect prediction	Results are not satisfactory.Propose followings for the future world; to improve the complexity metrics to, Presented a Bug Database for future researchers.
When do changes induce fixes? On Friday [32]	Study on fix-inducing changes	Prototype automatic- -ally locate changes in fix-inducing by connecting to version controlling system into a defect database, Bugzilla.
A validation of object-oriented defective design metrics as quality indicators [33]	Observe how much internal and external OO classes' design characteristics are affected by defectiveness.	OO metrics are useful for predicting defect density. Chidamber and Kemerer's metrics in defect prediction in earlier phases than traditional code metrics, which can only be collected at a final stage of development.
An empirical analysis of CK metrics for object-oriented design complexity [34]	Based on previous analysis, validate the performances of CK metrics in defect predicating at the stage of user acceptance testing	Identified that there is a significant relationship between high-quality software products, their language, and the OO approach (CBP, WMC, and DIT).
Predicting failure-prone components at design time [36]	Investigation on how decisions of design affect the quality of the software	Post-release defects can be identified and forecasted, from patterns in historical repositories. Propose an approach require only relationship of components (design phase)
Predicting Software Defectiveness by Mining Software Repositories [38]	Aims to progress the software development process as to localize and remove all the defects exists in software as fast as possible	Use Prime Machine Learning algorithm and regression decision trees to forecast residual defects in a program. Correlated Algorithm was used to remove unrelated data. Possible to work on 1000 file - version of a software
Fine-grained just-in-time defect prediction [39]	Investigate the importance in partially buggy commits and must these commits be a specific amount, conceive a bug prediction model to detect the buggy file in a commit	This work presents a new fine-grained JIT bug prediction model to predict significant files in a commit, are defective or non-defective files, this lets practitioners 'to manage efforts accordingly
Mining software defects: Should we consider affected releases? [40]	Nature of datasets collected with heuristic and realistic approaches that influence the earliest impacted version which estimated realistically for a particular bug.	Heuristic approach got a larger effect from bug count and binary defect datasets.

IV. DISCUSSION

In this study we have analysed the current topic that are vital to examine and challengers in the field of mining repositories in using them in project decision making. We have identified and analysed how mining software repositories have been done with different datasets, different algorithms used and current validation techniques of them. In this study we further found that repositories hold a huge amount of data but there is no defined or better extractor to extract important data available from them. Most of the time datasets used in research are from open source projects.

As per the observations from the reviewed researches we found that most of the time Data mining always confront with the datasets from conferences on software mining repositories. As reference [37] mentioned most of the time research datasets form open source projects that are taken from mining software conference. Only very few datasets were collected from enterprise level projects. In order to evaluate software as a whole and its quality descriptive statistics are used by the companies to mining and to initiate some software metrics to measure relevant qualities. In identifying different ways to use data in repositories the community of Mining software Repository plays a vital role. Anyway, there is very little work being conducted on the reliability and realistic of the practicality in the application of these findings to respond to the input data quality. There is always a question on the adequateness of the reliability.

Moreover, there is an increasing need in future discussions on data selection methods and techniques of processing these details. The reason for this is there is a big set of variations, and a high capacity of information on defects and commits. Here we significantly figured that mining techniques like topic models are emerging strategies on this work. Further sentiment Analysis also could be used to identify what kind of emotions or feelings are there to a developer on fixing particular defects by analysing pieces of texts. And this can be used by a project manager in assigning particular tasks to one developer from another but still there is no much usage of these sentimental analysis approaches in this field.

As mentioned in the associated challenges under mining with software repositories section, there must be a standard format in defect reporting. This is very important, though it seems fundamental to future credibility and practicality in application of research finding. These standard formats then direct the future tool and new techniques to emerge as well as to improve the practical usage of them around the mining software repositories.

V. CONCLUSION

A defect database is a good source of information for improvements in software engineering. In order to reduce the cost over projects new software defect prediction approaches are required to introduce at the same time there must be new software approaches in order to minimize cost overruns and new software metrics or blends of software metrics for inputs into different data mining techniques for the purpose of obtaining better prediction results. Further, this study revealed that Just In Time defect prediction is a better approach than traditional approaches. In this study, we found that anti-pattern codes have much more defect density than others. A defect report is vital to content for the developers and defect fixing process. It showed that the anti-pattern can support defects prediction and files used in anti-patterns have the highest density in its defects. This paper contains better and effective methods in finding out defect reports.

REFERENCES

[1] Ahmed E. Hassan and Tao Xie. 2010. Mining software engineering data. In Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 2 (ICSE '10), Vol. 2. ACM, New York, NY, USA, 503-504. DOI=10.1145/1810295.1810451

[2] http://en. wikipedia.org/wiki/ Data_mining #Data_ mining

[3] A. E. Hassan, "Software Intelligence : The Future of Mining Software Engineering Data," Proc. FSE/SDP Work. Futur. Softw. Eng. Res., pp. 161–165, 2010.K. Elissa, "Title of paper if known," unpublished.

[4] J. Caper, "Social and Technical Reasons for Software Project Failures", The Journal of Defence Software Engineering, vol.19, no. 6, 2006

[5] T. Zimmermann, N. Nagappan, and A. Zeller, "Predicting Bugs from History," Software Evolution. Springer Berlin Heidelberg, pp 69-88, 2008.

[6] A. E. Hassan, "The Road Ahead for Mining Software Repositories," Front. Softw. Maint., pp. 48–57, 2008.

[7] A. Chen, E. Chou, J. Wong, A. Y. Yao, Q. Zhang, S. Zhang, and A. Michail. CVSSearch: Searching through source code using CVS comments. In Proceedings of the 17th International Conference on Software Maintenance, pages 364–374, Florence, Italy, 2001.

[8] A. E. Hassan, A. Mockus, R. C. Holt, and P. M. Johnson, "Guest Editors ' Introduction : Special Issue on Mining Software Repositories," IEEE Trans. Softw. Eng., vol. 31, no. 6, pp. 426–428, 2005.

[9] T. Zimmermann, N. Nagappan, and A. Zeller, "Predicting Bugs from History," 2008.

[10] T. Zimmermann, R. Premraj, N. Bettenburg, C. Weiss, S. Just, and A. Schro, "What Makes a Good Bug Report ?," vol. 36, no. 5, pp. 618–643, 2010.

[11] S. Kumaresh and R. Baskaran, "Defect Analysis and Prevention for Software Process Quality Improvement," Int. J. Comput. Appl., vol. 8, no. 7, pp. 42–47, 2010.

[12] T. Pann, L. Zheng , C. Fang , "Defect Tracing System Based on Orthogonal Defect Classification" Computer Engineering and Applications, vol. 43, pp 9-10, 2008.

[13] S. Just and T. Zimmermann, "Towards the Next Generation of Bug Tracking Systems," Proc. - 2008 IEEE Symp. Vis. Lang. HumanCentric Comput. VL/HCC 2008, pp. 82–85, 2008.

[14] C. Weiß, T. Zimmermann, and A. Zeller, "Predicting Effort to Fix Software Bugs," Comput. Inf. Sci., pp. 2–3, 2006.

[15] A. E. Hassan, "Predicting Faults Using the Complexity of Code Changes," Proc. 31st Int. Conf. Softw. Eng., pp. 78–88, 2009.

[16] T. Fukushima, Y. Kamei, S. Mcintosh, K. Yamashita, and N. Ubayashi, "An Empirical Study of Just-in-Time Defect Prediction using CrossProject Models," pp. 172–181, 2014.

[17] S. W. Thomas, "Mining Software Repositories Using Topic Models," Mach. Learn., 2011.

[18] Rawat Mrinal Singh, and Sanjay Kumar Dubey. "Software defect prediction models for quality improvement: a literature study." IJCSI International Journal of Computer Science Issues 9.5 (2012): 16940814.

[19] S. Ehsan, S. Taba, F. Khomh, Y. Zou, A. E. Hassan, and M. Nagappan, "Predicting Bugs Using Antipatterns," IEEE Int. Conf. Softw. Maintenance, ICSM, pp. 270–279, 2013.

[20] T. J. Ostrand, P. Avenue, F. Park, E. J. Weyuker, P. Avenue, and F. Park, "A Tool for Mining Defect-Tracking Systems to Predict FaultProne Files."

[21] W. Abdelmoez, M. Kholief, and F. M. Elsalmy, "Improving Bug FixTime Prediction Model by Filtering out Outliers," Int. Conf. Technol. Adv. Electr. Electron. Comput. Eng., pp. 359–364, 2013.

[22] N. Carolina, E. Murphy-hill, T. Zimmermann, and C. Bird, “The Design of Bug Fixes,” 2013.

[23] S. Breu, R. Premraj, J. Sillito, and T. Zimmermann, “Information Needs in Bug Reports : Improving Cooperation Between Developers and Users,” Proc. 2010 Comput. Support. Coop. Work Conf., pp. 301– 310, 2010.

[24] T. Zimmermann and S. Breu, “Improving Bug Tracking Systems bug tracking,” 31st Int. Conf. Softw. Eng. - Companion Vol., pp. 247–250, 2009.

[25] P. J. Guo, ” Not My Bug !” and Other Reason For Software Bug Report Reassignments”. Comput. Inf. Sci., 2011.

[26] N. Bettenburg, T. Zimmermann, and S. Kim, “Duplicate Bug Reports Considered Harmful . . . Really ?,” IEEE Int. Conf. Softw. Maintenance, ICSM, no. Section 2, pp. 337–345, 2008.

[27] [27] T. Zimmermann and P. J. Guo, “Characterizing and Predicting Which Bugs Get Reopened,” Proc. 34th Int. Conf. Softw. Eng., pp. 1074– 1083, 2012.

[28] T. Chen, M. Nagappan, E. Shihab, and A. E. Hassan, “An Empirical Study of Dormant Bugs Categories and Subject Descriptors,” ACM, vol. 14, no. 05, 2014.

[29] T. Zimmermann, R. Premraj, and A. Zeller, “Predicting Defects for Eclipse,” Third Int. Work. Predict. Model. Softw. Eng. (PROMISE’07 ICSE Work. 2007), pp. 9–9, 2007.

[30] D. Cubranic and G. C. Murphy, "Hipikat: Recommending pertinent software development artifacts." in 25th International Conference on Software Engineering (ICSE), Portland, Oregon, 2003, pp. 408-418.

[31] M. Fischer, M. Pinzger, and H. Gall, "Populating a release history database from version control and bug tracking systems." in Proc. International Conference on Software Maintenance (ICSM 2003), Amsterdam, Netherlands, 2003.

[32] J. Śliwerski, T. Zimmermann, and A. Zeller, "When do changes induce fixes? On fridays." in Proc. InternationalWorkshop on Mining Software Repositories (MSR), St. Louis, Missouri, U.S., 2005.

[33] V. R. Basili, L. C. Briand, and W. L. Melo, “A validation of objectoriented design metrics as quality their paper they have suggested the simple and easy technique to search bug reports. Bug reports are crucial information to developer.

[34] R. Subramanyam and M. S. Krishnan, "Empirical analysis of ck metrics for object-oriented design complexity:Implications for software defects." IEEE Trans.Software Eng., vol. 29, pp. 297-310, 2003.

[35] A. B. Binkley and S. R. Schach, "Validation of the coupling dependency metric as a predictor of run-time failures and maintenance measures." in Proceedings of the International Conference on Software Engineering, 1998, pp. 452-455.

[36] A. Schröter, T. Zimmermann, and A. Zeller, "Predicting failure-prone components at design time." in Proceedings of the 5th International Symposium on Empirical Software Engineering (ISESE 2006), Rio de Janeiro, Brazil, 2006.

[37] Prechelt, L., Pepper, A.: Why software repositories are not used for defect-insertion circumstance analysis more often: A case study.

[38] S. Kasianenko, “Predicting Software Defectiveness by Mining Software Repositories,” Linnaeus University, Sweden, 2017.

[39] L. Pascarella, F. Palomba, and A. Bacchelli, “Fine-grained just-in-time defect prediction,” J. Syst. Softw., vol. 150, pp. 22–36, 2019.

[40] S. Yatish, J. Jiarpakdee, P. Thongtanunam, and C. Tantithamthavorn, “Mining software defects: Should we consider affected releases?,” in 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), 2019.

[41] Abeyratne, A., Samarage, C., Dahanayake, B., Wijesiriwardana, C., & Wimalaratne, P. (2020). A security specific knowledge modelling approach for secure software engineering. *Journal of the National Science Foundation of Sri Lanka*, *48*(1).

[42] Chen, J., Hu, K., Yu, Y., Chen, Z., Xuan, Q., Liu, Y., & Filkov, V. (2020, June). Software visualization and deep transfer learning for effective software defect prediction. In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (pp. 578-589).

[43] Wijesiriwardana, C., & Wimalaratne, P. (2018). Fostering Real-Time Software Analysis by Leveraging Heterogeneous and Autonomous Software Repositories. IEICE TRANSACTIONS on Information and Systems, 101(11), 2730-2743.

[44] Wijesiriwardana, C., & Wimalaratne, P. (2019). Software Engineering Data Analytics: A Framework Based on a Multi-Layered Abstraction Mechanism. IEICE Transactions on Information and Systems, 102(3), 637-639.

[45] Gurcan, F., & Cagiltay, N. E. (2019). Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. IEEE Access, 7, 82541-82552.

An Integrated Solution to Enhance the Flood Disaster Management Process

WWMNSB Wijekoon
Department of Information Technology
General Sir John Kotelawala Defence University
Ratmalana, Sri Lanka
wijekoon997@gmail.com

Maj RMD Pradeep
Department of Information Technology
General Sir John Kotelawala Defence University
Ratmalana, Sri Lanka
pradeep@kdu.ac.lk

Abstract - **Floods are the most catastrophic and cataclysmic events of all-natural disasters. According to the world-wide researches, most of the human lives are lost due to these flood disaster situations. Not only the human lives but also it will negatively influence the stability of a country in many. Therefore, every country must have a proper mechanism to effectively manage disaster situations. Flood management process can be categorized into three stages. 1. Flood detection and identification stage. 2. Flood alerting processes and early warning stage. 3. Refugee relocating and rescue processes stage. In order to acquire the fully utilization of this process, these three stages should be properly communicated and co-operated. The main problem of currently available disaster management systems is those are not properly co-operated and linked. Therefore, this research paper suggests a proper integrated software solution for address the abovementioned problem. This system performs flood monitoring, flood detection, early warning sending, refugee relocating and locating processes in an integrated manner with the help of android, cloud, windows application and IOT technologies.**

Keywords— Cloud, IOT, Disaster management, android.

I. Introduction

Floods are commonly occurring natural catastrophes that intervene in human livelihoods every now and then. The causes leading to such devastating events are multiple, yet the aftermath is intense and destructive, from a perspective of both, livelihood as well as country's overall economy. Therefore, it is mandatory to establish and maintain a proper flood disaster management mechanism. When comes to flood disaster management, there are three main visible stages to address. 1.Flood detection and identification stage. 2. Flood alerting processes and early warning stage. 3. Refugee relocating and rescue processes stage. The main goals in first stage is to detect and predict the flood situations early. The effectiveness of all other stages and activities directly depend on the efficiency and accuracy of this stage. The second stage, flood alerting, and early warning stage is responsible for pass the flood warnings among civilians, disaster management department, emergency response teams and other required parties as soon as possible. Third stage is the most important and critical stage. All the relocating processes, refugee searching missions, manage refugees in safe locations, etc. are conducted under this stage. All the processes and actions that consisting with in this stage will directly affect to the safety of human life and safety. The existing mechanisms are not abdicated enough to address such situations properly. Therefore, above mentioned three stages should upgrade with a more advanced and novel mechanism, in order to face for this increased flood disaster scenarios. And the main aspect is, the entire effectiveness of this flood disaster management directly relies on the better integration and communication among those three stages. The existing systems consist of poor integration and communication among these three stages since the lots of human interventions and manual inefficient activities are capped with the flood management mechanism. The Globe, currently, in living in a world dramatically ruled by technology, can opt for no better route for flood disaster management rather than making full utilization of information and communication technology. This is because communication is one of the most integral and crucial parts when tackling such intricate issues. Quite recently, the world has begun to incorporate the technology paradigms, such as the 'Internet of Things', mobile technologies, cloud technologies, and Geolocational systems and services in providing and integrating solutions to such calamities. These technological concepts, when used together, have the capacity to renovate and ameliorate the existing technologies of communication to machine-to-machine basis (M2M) [2]. This is primarily done by developing multiple pathways and prospects for even better and appropriate applications for disaster management authorities. This research paper aims to address the basic requirements for the ones in-charge of flood disaster management operations in a country through an integrated system. It further reiterates how to use IoT, Mobile, Cloud, and geographical information processing technologies as an integration solution for timely handling of the forthcoming disaster. . The research paper has been carefully sought to provide profound findings and analysis for the following aspects The analysis extends and validates the task-technology fit approach and discusses why and how IoT and GIS can give us the strategic value required. The research paper has been carefully sought to provide profound answers and analysis for the following questions: 1. Critical review about flood disaster management domain and existing flood management systems. 2. Index study of appropriate technologies, practices and theories. 3. Designing, development and implementation of the integrated flood disaster management system

II. Literature Review.

This chapter will discuss about the existing flood disaster management strategies, the new technologies which can in the solution and the identify the gap among the existing, and desire state of the flood disaster management processes. According to the "Shubhendu S Shukla and others" disaster management is the discipline of dealing with and avoiding both natural and manmade disasters [1]. According to the global senses, a world bank report states that there are approximately 3.8 million Km2 and 790 million individuals are exposed to at least two natural hazards [2]. In every disaster scenario the disaster situation divided into main three categories. 1. Preflood disaster stage, 2. During flood disaster stage. and 3. After flood stage. This research is about the flood disaster management and following three stages related to flood disaster situation are addressed from this research. 1. Flood detecting and identification processes. 2. Flood alerting and early warning stages. 3. Relocating and rescue activities.

978-0-7381-4403-0/20/$31.00 ©2020 IEEE

The effectiveness of each stage will do a dramatical influence on other stages. According to the "WL Delft Hydraulics" in pre-flood preparedness stage, the level of flood awareness is most valuable thing to proceed on further doings[3]. Evacuation/ Relocating is a response to the immediate or forecast threat of flooding that is expected to pose a risk to life, health or well-being. It involves people moving from their houses or places of business to safe locations which are locating out of the flood plains [3]. According to the "Sunarin chanta" shelter site is a public place which named by the government for gather the people who got affected from flood after the considerable tests, or experiments that confirmed as safe from flood [4]. But whenever the disaster happened people used to move for the nearest public place (temple, church or school) without considering the safety measures of those places. Sometimes those places also will get affected by the flood situations. This happens because of the poor communication among disaster management department and civilians and there is no proper guidance to the people whenever the disaster has happened. According to the "UNHCR (United Nations High Commissioner for Refugees") searching missions mainly focuses on where the victims are likely to be located and area of entrapment [4]. Currently these missions are conducted by the armed forces of Sri Lanka and assigned people form the government and the disaster management department. When considering the nature of these missions, normally large area was searched during thus mission to find if there is any person who trapped to the flood. Sometimes, there are people can be found but sometimes, there is not. Due to this reason it costs, lots of cost and time..

A. IoT Technologies in Disaster Scenarios.

The main components of a disaster situation management are environmental monitoring and pre-identification of disaster. In order to gain the success of this processes, the major and critical component is to identify the parameter which directly related with the disasters [5]. According to the "Sharma and others" IOT is a sort of "universal global neural network" in the cloud which connects various things. These IOT intelligently connect devices and systems which comprised of smart machines interacting and communicating with other machines [6]. Using these IOT, we can provide the ability to sense/ the external environment for the system (sensors, etc.). These systems can automatically interact with the external entities accordingly. Therefor it is possible to make automated systems that can take inputs, process them & provide outputs. Through this field, we can gain a secured & ultimate remote management control. Mainly this IOT can be used for disaster preparedness rather than the preventing from disasters or stopping them. Prediction processes, early warning systems & etc. Ex: - observance of forest fires. Sensors on trees will take measurements that indicates any robust risk. (temperature, moisture, green-house emission, and CO2 levels) [7]. According to the "Akash Sinha and others" IOT is proven to be fundamentally capable enough to provide more significant, scalable, portable, and energy efficient solutions to various problems in the disaster management. Specially, for the monitoring purposes and alerting purposes[8]. According to the "Stefan Poslad" an early warning system (EWS) is a core type of data driven Internet of Things (IoTs) system used for environment disaster risk and effect management[9]. According to the "Akash Sinha and others" one of the main key problem faced in disaster management is the communication gets hampered at the disaster region [8]. Sensor and satellite data collected at the site must be communicated to the DMU so that necessary actions can be taken as quickly as possible.

B. Google Map Related Geo-Information Processing and Mobile Technologies in Disaster Scenarios.

The awareness of the routes, geographical details about the area, exact locations of safe zones will enhance the effectiveness of this project. This will increase the efficiency of relocating purposes, rescue operations and addition to that logistics, emergency services etc. According to the "Varsha S. Sonwane", use of geographical information is very helpful when the disaster situation. There is an application which uses the Google map services for assisting relocating processes during a disaster situation. When the user in a disaster zone, it will send a message notification along with path to the member who can help.in this message user sends text message and path to reach for help. The path contains the distance from helper to the user. The distance can be given with the help of Google Map to show the distance by walk, by railway or by bus [10]. According to the "Janeetta Silvester and other" use of the google maps with mobiles give the maximum performances and utilizations for disaster scenarios. Because google map is easy to utilizes by any person, at any time. And with the use of mobile computing increases the accessibility and the portability of the system[11]. According to the "Teresa onarati, Igancio aedo" modern mobile devices embed several sensors such as GPS receivers, Wi Fi, accelerometers or cameras that can transform users into well-equipped human sensors[12]. For these reasons, emergency organizations and small and medium enterprises have demonstrated a growing interest in developing smart applications for reporting any exceptional circumstances. Considering the spread of such technologies, data collected from these sensors are useful for figuring out what is happening in a specific context. One of the greater examples for human sensors' participation is given by the earthquake and subsequent tsunami that happened in 2011 in Japan. The authors in this paper were discussed how people used Twitter or Skype to share positions, texts and photos and to stay in contact with their families[12]. According to the "Jovilyn therese B.fajardo, carlos m.oppus", different media channels allows people to communicate during crisis situations through the years. And, that authors pointed out, the social technologies are already merged with disaster relief [11]. And, those are assisted disaster management in four ways. 1. communication – quickly communicating with citizens for individual needs, 2 Self-help communities – cooperation through emergent groups. 3. integration of citizengenerated content – integration of information from various social software sources. 4. inter-organizational crisis management – cooperation among professional organization communities. According to the "M.L Tan and others" mobile social application which are not solely designed for the purpose of the disaster management such as Facebook, twitter can be used for gathering information and details during disaster. And some applications were designed only for the disaster management purposes. The applications which determines the optimum route along different geographical locations that the volunteers and rescuers need to take in order to serve the greatest number of people and provide maximum coverage of the area in the shortest possible time. Genetic algorithm was applied for optimization and different parameters were varied to determine the most optimum route. [13]

III. Research Methodolhy.

This study has been carried out with the primary aim of testing and launching a mechanism that provides adequate communication and transmission of key information during flood detection, generation of alert and sending it, evacuation of areas and handling refugees. For this purpose, a system has been developed that can provide multiple services to three kinds of stakeholders for dealing with flood disaster management using technology. Civilians, Emergency response teams and flood management department.

A. Problem Identification.

After developing a profound review of existing literature regarding flood disaster management, several loopholes were identified. Such as lots of human interventions/ manual processes are complies within the existing flood disaster management activities. These were related to flood detection, flood alerting and refugee relocation processes. Manual methods lead to ineffective communication and inadequate integration of above mentioned main three activities to reach the desired outcomes.

B. Analysis

An analysis of the current situation of methods of flood disaster management reveals that the system in-use has the following issues:

- Manual water level monitoring mechanism.
- Poor communication among flood detection, flood alerting and refuge relocating and locating processes.
- Isolated system in each level.
- Improper awareness about safe locations.

As a solution for the above-mentioned problems this research suggests and carried out an automated and integrated system for flood disaster detection, alert sending, refuge locating and relocating assisting. This is carried out through the development of an integrated system which based on the following aspects of technology: Android development, IOT (Internet of things), Windows application development. (ASP.net C# language), Cloud technology (Firebase), Google Map API, database (SQL server).

After the proper gap analysis following requirements were identified to be present in the proposed system in order to patch the pre-identified loopholes and fill-up the gap. Respective requirements are listed as follows under three main categories.

Functional Requirements.

- Flood Detection and identification. – View real time water level, identify floods, store water level details.
- Flood alerting and broadcasting. – generating flood alerts, send alerts to admins, admins should confirm alerts, civilians and emergency response teams should receive and view alerts.
- Refugee relocating and rescue operations. – View safe locations, navigate to safe area, send SOS messages, view SOS requests and view the refugee's locations, emergency response teams should navigate to the refugee's location.

Non-Functional Requirements.

a. User friendly. b. Accuracy c. Reliability d. Availability e. Portability and accessibility d. Security.

Technical Requirements.

a. Platform which system will be running, b. Databases. (Local and Cloud databases) c. Geographical information processing d. Programming languages and other development equipment e. Hardware requirements – Android smart phones, Arduino micro controller board and ultrasonic sensors, windows computers.

All key processes were developed, keeping in mind the three stages of flood disaster management. The software was used to develop a number of modules to substitute each process of the flood disaster management and cloud database was used for the integrating purpose of those modules.

C. Design and Development of Solution System.

The following section of this research paper gives a thorough insight of the developed application as a proposed solution to the above identified problem and how the entire requirements are embedded into this project. The entire system is designed and developed, keeping in mind that integrity and the integration of the main activities which should perform under three main stages of flood disaster management.

1) Overall Architecture of the System.

This portion of the research methodology gives profound analysis of the system and the three-tier architecture that was integrated towards finding optimum solutions, categorized into the following three layers: client layer, application layer, and data layer. Figure 2 exhibit the layered architecture of the proposed system.

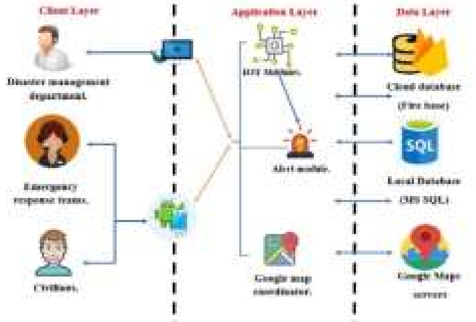


Figure – 2 Three layers of the proposed system

a) Client Layer - The client layer for the system had been designed with the primary aim of administering human and system interactions. It was mainly utilized by three stakeholder groups, as mentioned above: civilians, disaster management department, and the emergency response teams. It was rather intricate to implement user interaction using the 'client layer' on such a widespread scale. Therefore, system was designed two ways to enable for users to consume the services of this system. through the Android application civilians and emergency response teams interact with the system and through a windows application disaster management department is interacted with the system.

b) Application Layer - The application layer handles the core logic of the entire application and it is comprised of the following:

- IOT component.
- Alert generating and sending component.
- Geo-map/ coordinate component.

These three components handle the entire application logic and the communication among all main three types of stake holders as well as the three stages of the flood disaster situation throughout the system independently to each other.

c) Data Layer - This tier contains the managing the connection with data sources and maintain the data related processes. (CRUD – create, read, update and delete). This structured data layer, for this project, handled three integral data sources. The first one was the local SQL database, where all information regarding the users of disaster management department such as alert details, coordinate location of safe zones, sensor readings, sensor locations were retrieved. This database was only accessible by the top personnel of the flood disaster management department and could not directly meet external parties. The second data source was the Cloud database. This particular source kept records of the emergency response teams and the local population, and contained valuable details like sensor values, details of safe ones, coordinate location of various relevant locations, etc. The third source involved the use of Google Maps data servers by incorporating the Google Map API. It is not straightway run or administered by the system.

2) Module Architecture.

As mentioned above of this paper, this entire system consisting with three main modules. Those respective main modules were further dividing into the sub-module in order to achieve the separation of concerns. All the modules and sub-modules are discussed as follows.

a) Flood detection and identification module

Water level measuring sub-module. - measure the water level in real-time basis and send them to the windows application.

Flood identification sub-module. – compare the received water level values with the knowledge base for identify the flood situations.

Water level view module – enables to view the water levels for civilians, emergency response teams and flood disaster management department. Figure 3 exhibit the logical design of flood detection module.

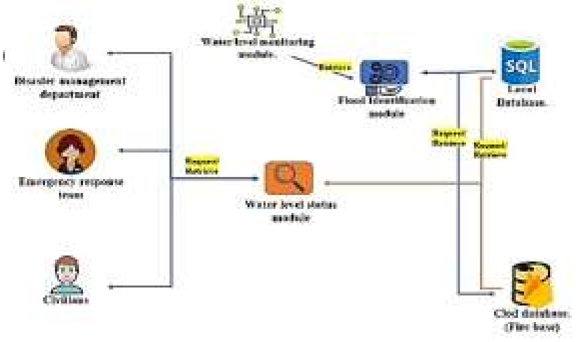

Disaster management department
Water level monitoring module.
Retrieve
Flood identification module
Local Database.
Emergency response team
Request Retrieve
Request Retrieve
Request Retrieve
Water level status module
Civilians
Clod database. (Fire base)

Figure – 3 Design of the flood detection and identification module

b) Alert generating and broadcasting module.

Alerts generating sub-modules. – automatically generated an alert and send it to admins to conform.

Alerts confirmation sub-module. – admins confirm the system generated alerts prior to the broadcasting.

Alert sending sub-module. – broadcast the alerts to the other stakeholders of the system. Figure 4 exhibit the logical design of alert generating and broadcasting module

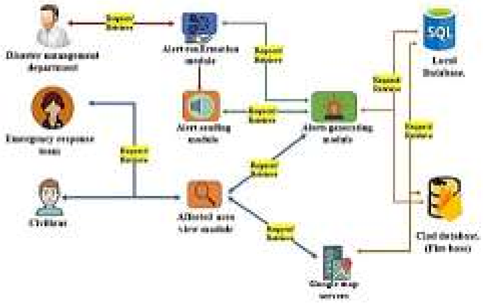


Figure – 4 Design of the alert generating and broadcasting module

c) Refugee relocating and locating module.

Relocating assisting module – navigates civilians, SOS teams to safe areas.

SOS module – enables for civilians to initiate SOS messages and view SOS requests for SOS teams. Figure 5 exhibit the logical design of refugee relocating and locating module.

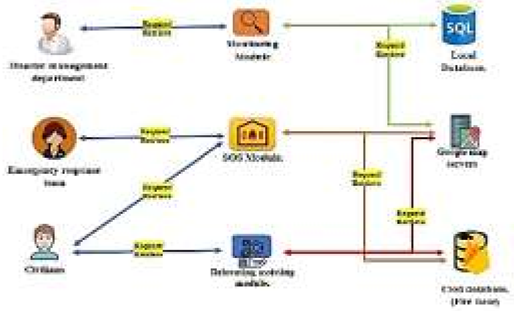
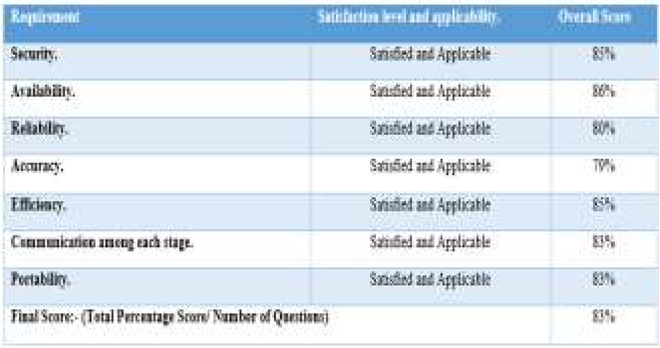


Figure – 5 Design of the refugee relocating and locating module.

IV. System Evaluation.

The summative evaluation was utilized to verifying the results of the proposed system to meet the functional and non-functional requirements of end-users. Simply, this is an assessment of the utilization of prerequisites and the exactness of the arrangement.

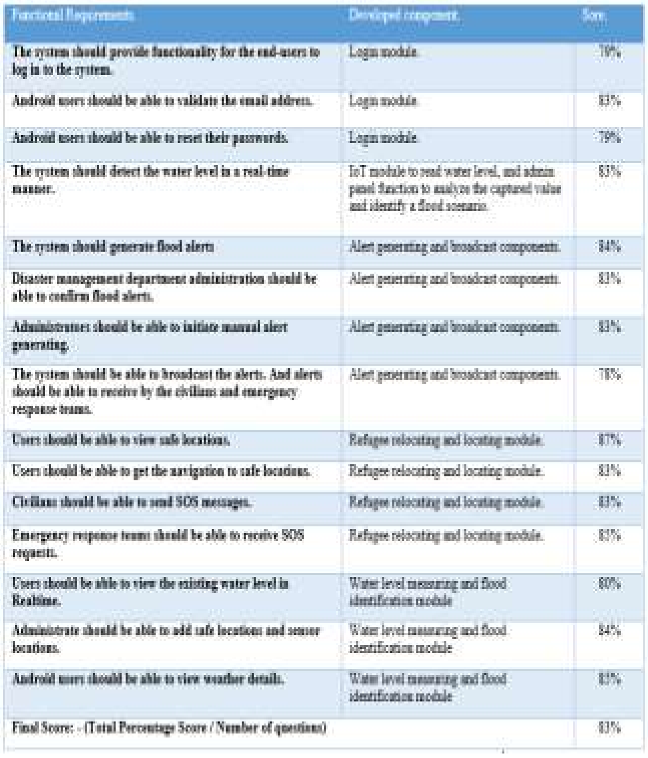
A. Evaluation Results of Non-Functional Requirements.



Requirement	Satisfaction level and applicability.	Overall Score
Security.	Satisfied and Applicable	85%
Availability.	Satisfied and Applicable	86%
Reliability.	Satisfied and Applicable	80%
Accuracy.	Satisfied and Applicable	79%
Efficiency.	Satisfied and Applicable	85%
Communication among each stage.	Satisfied and Applicable	83%
Portability.	Satisfied and Applicable	83%
Final Score:- (Total Percentage Score/ Number of Questions)		83%

Table – 1 Evaluation results of Non-functional requirements.

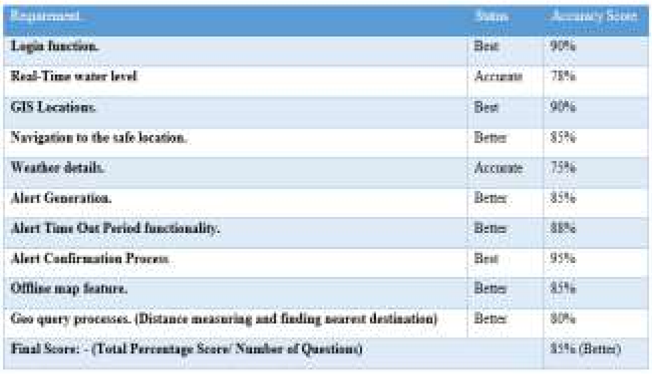
B. Evaluation Results of Functional Requirements.



Functional Requirements	Developed component.	Score.
The system should provide functionality for the end-users to log in to the system.	Login module.	79%
Android users should be able to validate the email address.	Login module.	83%
Android users should be able to reset their passwords.	Login module.	79%
The system should detect the water level in a real-time manner.	IoT module to read water level, and admin panel function to analyze the captured value and identify a flood scenario.	83%
The system should generate flood alerts	Alert generating and broadcast components.	84%
Disaster management department administration should be able to confirm flood alerts.	Alert generating and broadcast components.	83%
Administrators should be able to initiate manual alert generating.	Alert generating and broadcast components.	83%
The system should be able to broadcast the alerts. And alerts should be able to receive by the civilians and emergency response teams.	Alert generating and broadcast components.	78%
Users should be able to view safe locations.	Refugee relocating and locating module.	87%
Users should be able to get the navigation to safe locations.	Refugee relocating and locating module.	83%
Civilians should be able to send SOS messages.	Refugee relocating and locating module.	83%
Emergency response teams should be able to receive SOS requests.	Refugee relocating and locating module.	83%
Users should be able to view the existing water level in Realtime.	Water level measuring and flood identification module	80%
Administrate should be able to add safe locations and sensor locations.	Water level measuring and flood identification module	84%
Android users should be able to view weather details.	Water level measuring and flood identification module	83%
Final Score: - (Total Percentage Score / Number of questions)		83%

Table – 2 Evaluation results of Functional requirements.

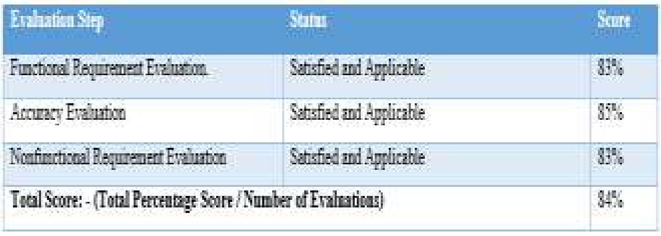
C. Evaluation of The Accuracy of Critical Components.



Requirement.	Status.	Accuracy Score.
Login function.	Best	90%
Real-Time water level	Accurate	78%
GIS Locations.	Best	90%
Navigation to the safe location.	Better	85%
Weather details.	Accurate	75%
Alert Generation.	Better	85%
Alert Time Out Period functionality.	Better	88%
Alert Confirmation Process.	Best	95%
Offline map feature.	Better	85%
Geo query processes. (Distance measuring and finding nearest destination)	Better	80%
Final Score: - (Total Percentage Score/ Number of Questions)		85% (Better)

Table – 3 Evaluation results of Accuracy of critical components.

D. Overall Evaluation Results.



Evaluation Step	Status	Score
Functional Requirement Evaluation.	Satisfied and Applicable	83%
Accuracy Evaluation	Satisfied and Applicable	85%
Nonfunctional Requirement Evaluation	Satisfied and Applicable	83%
Total Score: - (Total Percentage Score / Number of Evaluations)		84%

Table – 4 Overall Evaluation Results

According to the interpreted results (which are showing in above table 1, 2, 3, and 4) from the evaluation function, the applicability of the system, this developed version can be delivered as a minimum viable product to the actual environment to perform its actions with parallel to the existing system.

Limitations –
Connectivity with the internet during a flood situation. – this limitation was addressed to a considerable extent within this project by adding offline features, a separate thread approach, and manual alert generating processing through the system.
Civilians need android smartphones – currently, the mobile application is developed only for android phones. But as a further enhancement researcher has mentioned the spread of this application to the IOS platforms.
Technological literacy – users (especially the civilians, should have to have technical literacy to some extent. This problem is also addressed by the researcher a considerable amount by adding more user-friendly interfaces, and automatic configurations options into the android applications.
Complete test activities. - since a real flood situation is needed to test this application to find out threshold values, breaking points, etc. Therefore, parallel implantation should be taken place until the full validation of the system. Other than that, all the system functionalities are tested and validated through several test activities.

V. Conclusion and Recommendations

Based on a rigorous examination of existing literature on the subject matter, it is clearly seen how the incorporation of newer technologies in flood disaster management systems can simplify the processes and increase their efficiency and productiveness. In addition to this, IoT, cloud, mobile, and geo-information processing technologies can significantly alleviate losses of human lives. Floods are prevailing throughout the world and need prompt attention. Countries like Sri Lanka should, as soon as possible, adapt and evolve by integrating newer solutions in their departments and systems of disaster management. The research paper identifies all benefits of incorporating the above-mentioned technology aspects into systems. On the other hand, it provides a detailed insight into the drawbacks of the existing manual, inefficient and inaccurate mechanisms. IoT gives us a consistent interconnection among heterogeneous gadgets. By applying information investigation and coordination of different IoT gadgets utilized for early alerts about flood situations. If there should arise an occurrence of a flood circumstance part of relocatable cycles are conveyed. By utilizing Geo-information processing and mobile technology innovation to address this issue (for safe area choice, identifying refugee locations, navigating, etc.) gives a compelling and fruitful result.

Furthermore, the following recommendations were made by the researcher after the serious analysis of the identified limitations and the planned further improvement on this research project.
a. Since the system needed to be tested on an actual flood scenario and it was unable to be carried out within this research work researcher recommends going with parallel implementation along with the existing mechanisms, b. Assign some dedicated users from the flood departments' administration side to operate this system and monitor the

progress, issues, limitation during a flood situation as well as in the normal way, c. And the developed flood identification module is better to be upgraded with flood prediction capabilities using machine learning concepts, d. Use many training camps and workshops for all these stakeholders to make them more familiar with the developed system. (especially for the civilians.), e. Place the water level measuring IoT components in a place where anyone can easily reach for the maintainable and replacing purposes. But with more physical security procedures.

ACKNOWLEDGEMENT

This entire research was made possible through the greater support and guidance of my supervisor and my parents. Therefore, this acknowledgment will address the gratitude towards them.

REFERENCES

[1] Shubhendu S Shukla, “Disaster Management: ‘Managing the Risk of Environmental Calamity,’” vol. 1, no. 1, p. 7, 2013.

[2] N. C. Bronfman, P. C. Cisternas, P. B. Repetto, and J. V. Castañeda, “Natural disaster preparedness in a multi-hazard environment: Characterizing the sociodemographic profile of those better (worse) prepared,” *PLOS ONE*, vol. 14, no. 4, p. e0214249, Apr. 2019, doi: 10.1371/journal.pone.0214249.

[3] D. Lumbroso and B. Woods-Ballard, “Review of evacuation rescue methods and models,” 2006, doi: 10.13140/RG.2.1.3347.3368.

[4] “Search and rescue response and coordination (natural disasters) - UNHCR|Emergency Handbook.” https://emergency.unhcr.org/entry/222599/search-and-rescue-response-and-coordination-natural-disasters (accessed Apr. 06, 2020).

[5] D. Purkovic, L. Coates, M. Hönsch, D. Lumbeck, and F. Schmidt, “Smart river monitoring and early flood detection system in Japan developed with the EnOcean long range sensor technology,” in *2019 2nd International Colloquium on Smart Grid Metrology (SMAGRIMET)*, Apr. 2019, pp. 1–6, doi: 10.23919/SMAGRIMET.2019.8720390.

[6] V. Sharma and R. Tiwari, “A review paper on ‘IOT’ & It‟s Smart Applications,” vol. 5, no. 2, p. 5, 2016.

[7] H. N. Saha *et al.*, “Disaster management using Internet of Things,” in *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, Bangkok, Thailand, Aug. 2017, pp. 81–85, doi: 10.1109/IEMECON.2017.8079566.

[8] Akash Sinha, P. Kumar, N. P. Rana, R. Islam, and Y. K. Dwivedi, “Impact of internet of things (IoT) in disaster management: a task-technology fit perspective,” *Ann. Oper. Res.*, vol. 283, no. 1–2, pp. 759–794, Dec. 2019, doi: 10.1007/s10479-017-2658-1.

[9] S. Poslad, S. E. Middleton, F. Chaves, R. Tao, O. Necmioglu, and U. Bugel, “A Semantic IoT Early Warning System for Natural Environment Crisis Management,” *IEEE Trans. Emerg. Top. Comput.*, vol. 3, no. 2, pp. 246–257, Jun. 2015, doi: 10.1109/TETC.2015.2432742.

[10] V. S. Sonwane, “Disaster Management System on Mobile Phones Using Google Map,” vol. 5, p. 4, 2014.

[11] J. T. B. Fajardo and C. M. Oppus, “A Mobile Disaster Management System Using the Android Technology,” vol. 9, no. 6, p. 12, 2010.

[12] “(PDF) Improving emergency response: citizens performing actions,” *ResearchGate*. https://www.researchgate.net/publication/262412269_Improving_emergency_response_citizens_performing_actions (accessed Sep. 27, 2020).

[13] M. L. Tan, R. Prasanna, K. Stock, E. Hudson-Doyle, G. Leonard, and D. Johnston, “Mobile applications in crisis informatics literature: A systematic review,” *Int. J. Disaster Risk Reduct.*, vol. 24, pp. 297–311, Sep. 2017, doi: 10.1016/j.ijdrr.2017.06.009.

Deep Neural Network and Image Processing Based Approach for Identifying Road Signs

V.Diluxshan*
Department of Computing and Information Systems,
Sabaragamuwa University of
Sri Lanka.
vdiluxshan@std.appsc.sab.ac.lk

BTGS Kumara*
Department of Computing and Information Systems,
Sabaragamuwa University of
Sri Lanka.
kumara@appsc.sab.ac.lk

Abstract - **Road sign identification in images is an important issue, in particular for vehicle safety applications. Normally it will notify into some stages, those are detection, recognition, tracking and identify with the help of the dataset. Here Convolutional Neural Network [CNN], Robot Operating System [ROS] is using to undertake this project. In this study, as an Objective, To obtain an accurate image representation, this paper would discuss the key methods of separating an image into regions by using data levels. We mainly compare the Counter-based, Region-based, Colour-based, segmentation using edge detection with the boundary estimate. For many image processing and computer vision algorithms, Image segmentation is an important step, while an edge can be informally defined as the edge between adjacent parts of an image. According to the recent year progress of vehicle manufacturing, Bulk of people like to buy this kind of vehicles to use. Because of the safety and stress release of the driving part. Trending map technology navigates to needed places is easier to access. The proposed work aims to detect traffic sign which has illumination variation, Support Vector Machine [SVM]. For the whole system, we need to use the Raspberry pi 3 processor and camera which automatically captures the video data. Research says, that people will get safe while using the automotive vehicles to have all the facilities to safe the driver/ passenger in the proper way.**

Keywords— Autonomous Self-driving vehicle, Road Signal identification, Neural Network image processing, Feature extraction in Image processing, Deep Learning, SVM classifier

I. INTRODUCTION

The Road Signs we were around us date long back in history. The earliest road signs were milestones, giving distance or direction. In the middle ages, multidirectional signs at intersections became common, giving directions to cities and towns. With the advent of motorized traffic and its increasing pressure on the road, many have adopted pictorial signs and standardized their signs to facilitate international travel, where language differences would create barriers. Also, Here going to mention about the impact of Automation technology in the world. The Riders, Actually the new license holders don't know the Rules and Regulations of the traffic system. That's why plenty of accidents will happening on roads [1]. Thus, we need to verify the sensors were working properly or not. While Autonomous technologies have improved substantially, they still ultimately view the drivers around to get to know the drivers' mind in a proper manner (Every system need to be improved by the users' review and experience). Road and traffic signs must be properly installed in the necessary locations and an inventory of them is ideally needed to help ensure adequate updating and maintenance. The main purpose of the Driving Assistance System (DAS) is to collect information for drivers to reduce their effort in safe driving. CNN improved a lot by the use of emergence of GPUs, ReLU activation neurons, max-pooling instead of average pooling, and dropout regularization. CNN is being extended into many other areas such as face recognition, image retrieval, and object detection, not only recognizing hand-crafted features. Dataset is taken part from GTSRB site to introduces high performance. According to the sudden development, the need for software and hardware impacts on the processing power which is much enough to identify perfectly. Gathered inappropriate background images, such as images partially including a traffic sign [2]. The system introduces a method of sign identification with the help of Raspberry pi. When a vehicle is moving high speed, then it is difficult to process the better image classify.

II. BACKGROUND & MOTIVATION

A. Purpose of the Study

Travelling is one of the main parts of every person. Because it will full fill the person to healthy. According to travel, we need to ensure the quality of the driver. So our research will help the driver to drive the proper way. The main reason for conducting this research is, improving the CNN model for getting the best output from the dataset, Which is used to install in the vehicle. The driver will be notified while nowadays Accidents which was happened by the mistake of Driver. The main purpose of the Driving Assistance System (DAS) is to collect significant information for drivers to reduce their effort in safe driving. From the exciting research, it different from comparing the features to getting high accuracy [1].

B. Significance of the Project

- To know which features of feature extraction is suitable for a traffic sign identification.
- The main objective is to compare the features by feature extraction such as Counter based (Area, Perimeter, Aspect ratio, Entropy), Region-based (Circularity, Rectangularity), Colour based(Contrast, Red channel). Combining all three main features to generate more accuracy than the previous study.
- Analyzing the entire model by the help of benchmark dataset, change the preprocessing rate (epochs-number of counts preprocess the model)

and the Batch size of the benchmark dataset → Getting the well-trained model to use into my vehicle sign identification system [3].

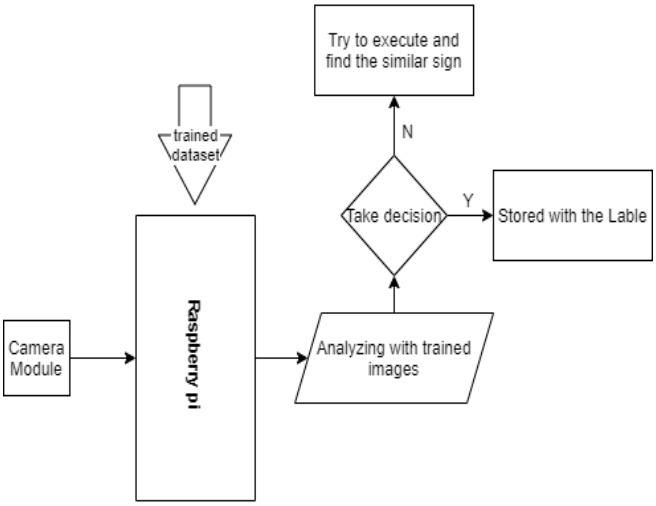


Figure 1. Proposed Block Diagram

Above fig1 shows the detail description of the flow of system working. According to the flowchart, can understand the main purpose of researching this topic. The system should detect the traffic signs in real-time and alert the user of the information. For this detection, we must create a data set of all needed traffic signboards and must train the system so it can compare and detect the real-time images. The system will be trained by collected datasets, the data sets will be created by capturing the signboards and identify the signs and name the information and store it. The classification was undertaken using an SVM classifier

III. Proposed System

The paper used to reveal about image processing concepts to detect various signs in the transport system. CNN helps to identify an exact image to extend the dataset. Self-driving cars need traffic sign recognition to properly parse and understand the roadway. Similarly, driver alert systems inside cars need to understand the roadway around them to help aid and perfect drivers.

Belove Fig.1 shows the block path of road sign detection and processing of the image flow. The images are processed with the CNN algorithm for sign detection and classification. The output of the image will be stored into the folder, with the help of the label name.

A. Abbreviations and Acronyms and Units

CNN- convolutional neural network, SVM- support vector machine, ROS- robot operating system, SLR- systematic literature review, AI- artificial intelligence, DAS- driving assistance system, IoT- the internet of things, RPI- raspberry pi. *Units:* These following units were used to train the model successfully. Raspberry pi 3B, Camera Module for RPI, Tensor flow, Keras [6].

Each country has its shape of road signs. So it's difficult to use one model to another country. So there were delimitations include. More than 39000 images were included in the research to train the model [1]. The entire images were labelled as 43 main classes.

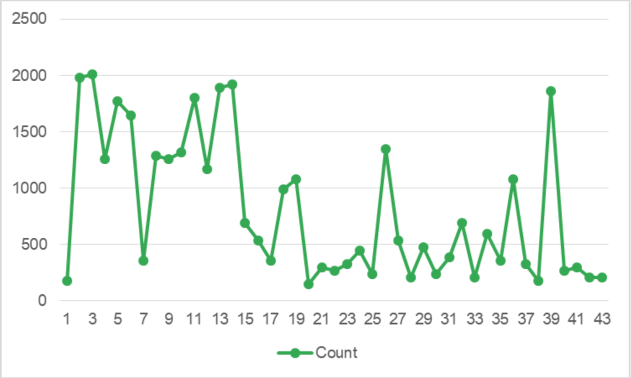


Figure 2. Classes and train data count for the collected dataset

Fig.2 is showup the entire dataset count and class details. The used dataset is the benchmark analyzing by the existing researchers [1]. The loaded image was loaded and get ready for making changes. The changes respectively rotated, zoom, shift, flip and shear to classify.

IV. Methodology

Road traffic signs are difficult to be detected easily in a complex environment. According to CNN, the model is used to train by the two categories. Detection stage and Classification stage. On the image acquisition stage, all the real-time images have to go through this feature extraction. And then those images resized for the import size. The images in each class were randomly identified for the classification [3] [4].

Data were labelled as 43 classes and Divided to train the model multiple time. Below graph will show up the entire dataset history for mapping the images. The excess images have to resize for modelling the dataset. Test dataset has implemented as 12000. On the preprocessing part, image contrast filtering is mainly focused to manipulate. Also, SVM is used to analysing the feature extraction to the collected dataset.

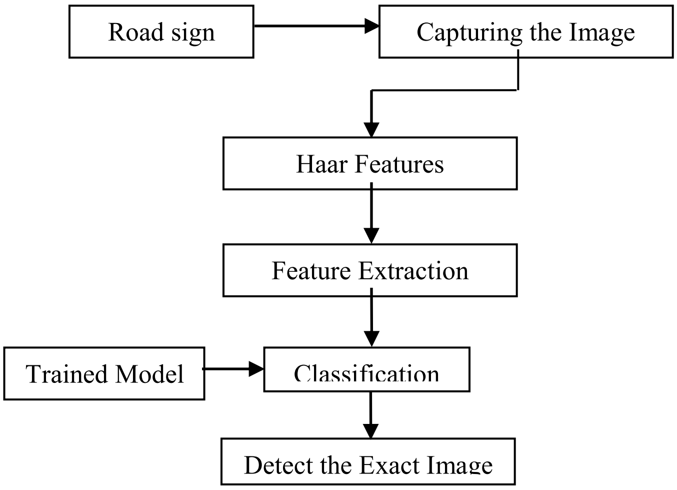


Figure 3. Block Diagram of the Algorithm

Fig.3 detailly show the process pathway. Here the captured image is sent to the detection stage. Localization was detected wherefrom the input. And the Recognition part was classified the traffic sign. According to the multiple models, I have chosen the best model for testing the image with high accuracy [3].

And according to the image feature extraction, research was taken part to the Suppor vector machine(SVM) method to specific features to identify how deeply involved with the Accuracy. Below shapes are standard shapes to segment the traffic signs according to the attributes [8].

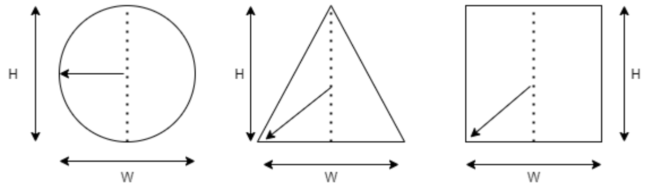


Figure 4. Schematic diagrams of the basic Shape

$$C = \frac{4\pi S}{L^2} \qquad R = \frac{S}{W \times H}$$

Here, Roundness(C), Circumference of the Mark(L), Area of the mark(S), Rectangular (R), Width(W), Height(H) of the particular road sign [8]. According to SVM traffic sign need to be segmented from the background, then the information contained in the segmented region is assigned to the matrix. These images go through Figure 3 data flow to pass the images through the steps. Otherwise, it will be more complex to identify by the trained model [1]. In figure 4 mentioned shapes defaulted sign outlet to entire signboards [2].

V. Results

From the collected dataset, models were trained and the average accuracy is calculated from the learning rate.

TABLE I. CNN based accuracy from the selected model.

Num	Parameters			Accuracuy			Average
	Epoch	Learning Rate	Batch Size	precision	recall	f1-score	
1	1	3	64	23	20	21	21.33
2	5	3	64	89	88	87	88.00
3	10	3	64	88	85	85	86.00
4	10	3	80	87	85	85	85.67
5	12	3	80	90	89	89	89.33
6	11	3	64	89	87	87	87.67
7	18	3	70	83	80	79	80.67
8	22	3	65	94	94	94	94.00
9	25	3	70	83	70	69	74.00
10	27	3	70	93	92	92	92.33
12	30	3	50	79	67	65	70.33

According to the CNN method, there were hard to restrict the feature extraction. SVM conducted by below features to analyzing the exact specification, which is mostly involving inaccuracy. Those are Area, Perimeter, Aspect_ratio, Rectangularity, Circularity, Contrast, Correlation, Entropy.

TABLE II. High accuracy from the selected model of CNN.

	precision	recall	fi-score	support
Accuracy			0.94	12630
Macro avg	0.93	0.92	0.92	12630
Weighted avg	0.94	0.94	0.94	12630

Table.1 shows the obtained model accuracy by average from 43 classes. The trained model reflects 94% of accuracy than other predefined models.

Above bar chart shows the detailed description of the accuracy of the image by the supported dataset [5]. It will show the particular image counts with the trained model count. Accuracy is narrowly increased and constant, while approaches going to increase continuously,

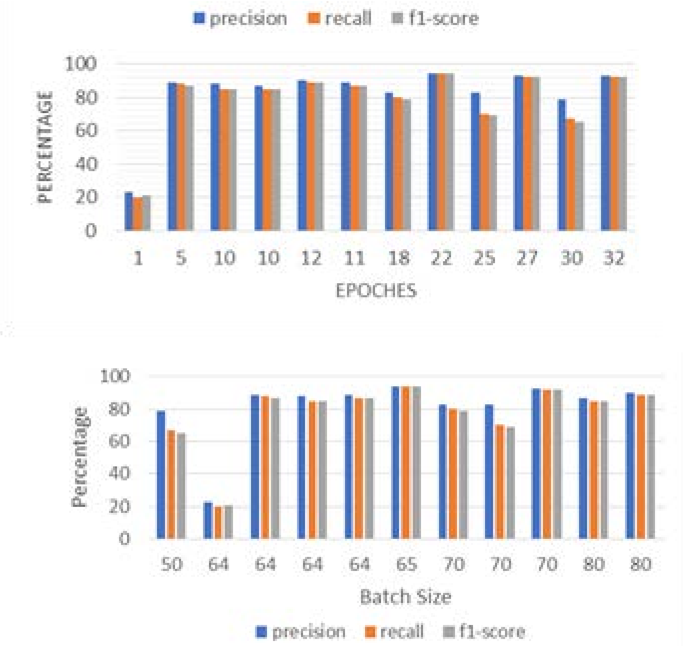


Figure 5. Accuracy deviation with Epoaches & Batch size of image

We can come up with the solution, which is made to analyzing the Loss and Accuracy values from the below graphs. The graph is generated from model training time. Matplotlib is helping to generate the above graph. Figure 3.8 is generated from the classified part. There you can see the flow of the graph with the increment of Epoaches(count of Epochs) [6].

- Train Loss (Redline)
  - The loss of train model will slowly go down through the increment of Epochs. So we can get a decision,

- Validation Loss (Blue line)
  - Bothe Train and validation go to lose. But validation loss decreases, not in a proper way.
- Tain Accuracy (Purple line)
  - Train accuracy slowly goes up to near to 1 (100%). It slowly increases to 90% till 10 epochs and then there were no changes up to 30th
- Validation Accuracy (Black-line)
  - Its slowly go up with train accuracy and suddenly have a drop in 25th epoch.

These above main points can describe below graphs well. Loss and Accuracy values can narrowly change their path to efficient values [10].

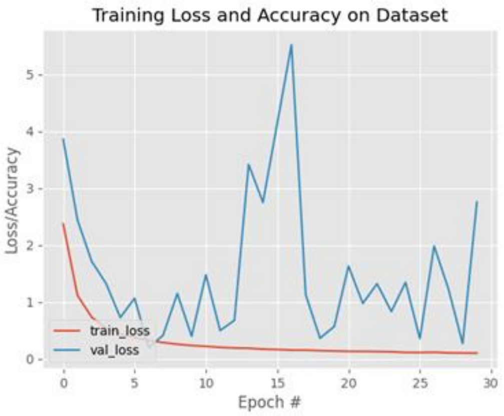


Figure 6. The loss value change with the epoch

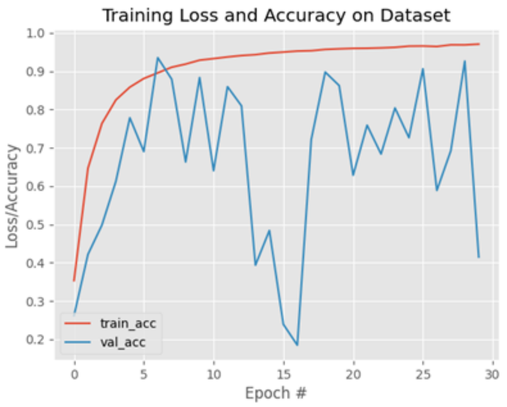


Figure 7. The Accuracy value change with the epoch

The saved model can use by prediction code to identify well. The model successfully trained as 94% percentage. Here detailedly shows each label and their preprocessing times, recall, support data by suitable labelled folder. Finally, the weighted average was calculated by the system. For almost near to 12000 testing images. According to the model training period [7].

TABLE III. SVM model specification comparison for future extraction.

No	Area	Peremeter	Aspect_ratio	Rectangularity	Circularity	Contrast	Corerelation	Entropy	Percentage
1	▨	▨	▨	▨					85.35
2	▨	▨	▨	▨	▨				85.59
3	▨	▨	▨						85.95
4	▨	▨	▨	▨	▨	▨			88.57
5	▨	▨	▨	▨	▨	▨	▨		90.52
6	▨	▨			▨				91.42
7	▨	▨	▨	▨	▨	▨	▨	▨	92.23

According to table 3, image classification with future extraction is more important to image processing methodology [9].

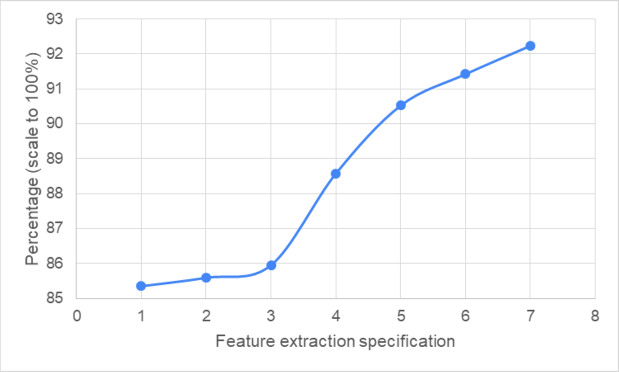


Figure 8. The feature extraction differences between each specific categories.

The flow of the graph is slightly increased with the trained features. According to this, we can get the decision to increase the image-specific features. To the image processing, feature extraction takes part to increase accuracy. And shows the pattern of graph for features vs accuracy of the model. The trendline values can be involved to mapped to the sample. Likewise, we can add more features to analyzing the accuracy high. But always the important features can be involving the changes inaccuracy. Other features might be a support to lose accuracy [10].

VI. Discussion and Conclusion

Area, Perimeter, circularity, contrast, Aspect ratio, Rectangularity, Correlation, Entropy was mainly influenced by the data preprocessing. So we have to be sure about the above factors to apply before classify. So traffic sign recognition has to pass the preprocessing part to better performance. The videos are real-time captured by the Pi Camera Module which is connected to the Raspberry Pi installed in a moving vehicle. Hence, the brightness, contrast, clarity and noises of scenes may have a large difference when the weather or other conditions are changed by time and locations. However, these variables could increase recognition difficulty and affect recognition results. 1. To increase the robustness of the proposed scheme, some pre-processes are used to reduce the influence of variable conditions. 2. Include the user profile to provide the point system for every driver, and attached to the License documents. 3. Attach the location-based road sign identification, to the feedback to RDA department (about the repairs and errors).

ACKNOWLEDGEMENT

I would like to express my deepest gratitude towards my thanks to Mr Tharmakulasingam Jeyjenthan, CEO at SenzMate (Pvt) Ltd, Mr K.Suthagar, Software Engineer at Sysco LABS Sri Lanka, Mr Easanka Diddenya, Demonstrator, Sabaragamuwa University of Sri Lanka for shared their valuable time which is given to me freely and for the discussions and criticisms. And advice is given by Mr K.Banujan the Lecturer Sabaragamuwa University of Sri Lanka.

VII. References

[1] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German Traffic Sign Recognition Benchmark: A multi-class classification competition," *Proc. Int. Jt. Conf. Neural Networks*, pp. 1453–1460, 2011, doi: 10.1109/IJCNN.2011.6033395.

[2] R. Belaroussi, P. Foucher, J. P. Tarel, B. Soheilian, P. Charbonnier, and N. Paparoditis, "Road sign detection in images: A case study," *Proc. - Int. Conf. Pattern Recognit.*, pp. 484–488, 2010, doi: 10.1109/ICPR.2010.1125.

[3] S. O. Ali Chishti, S. Riaz, M. Bilal Zaib, and M. Nauman, "Self-Driving Cars Using CNN and Q-Learning," *Proc. 21st Int. Multi Top. Conf. INMIC 2018*, 2018, doi: 10.1109/INMIC.2018.8595684.

[4] N. P. Botekar and M. N. Mahalakshmi, "Development of road sign recognition for ADAS using OpenCV," *Proc. 2017 Int. Conf. Intell. Comput. Control. I2C2 2017*, vol. 2018-January, pp. 1–4, 2018, doi: 10.1109/I2C2.2017.8321941.

[5] R. Hmida, A. Ben Abdelali, and A. Mtibaa, "Hardware implementation and validation of a traffic road sign detection and identification system," *J. Real-Time Image Process.*, vol. 15, no. 1, pp. 13–30, 2018, doi: 10.1007/s11554-016-0579-x.

[6] N. Kryvinska, A. Poniszewska-Maranda, and M. Gregus, "An approach towards service system building for road traffic signs detection and recognition," *Procedia Comput. Sci.*, vol. 141, pp. 64–71, 2018, doi: 10.1016/j.procs.2018.10.150.

[7] M. Bénallal and J. Meunier, "Real-time color segmentation of road signs," *Can. Conf. Electr. Comput. Eng.*, vol. 3, pp. 1823–1826, 2003, doi: 10.1109/ccece.2003.1226265.

[8] C. Y. Fang, C. S. Fuh, S. W. Chen, and P. S. Yen, "A road sign recognition system based on dynamic visual model," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2003, doi: 10.1109/cvpr.2003.1211428.

[9] S. P. S. Subramanian and S. Ganesh Vaidyanathan, "Automatic traffic sign identification system for real time operation," *Proc. Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), I-SMAC 2018*, pp. 423–427, 2019, doi: 10.1109/I-SMAC.2018.8653797.

[10] Y. Y. Nguwi and A. Z. Kouzani, "Detection and classification of road signs in natural environments," *Neural Comput. Appl.*, vol. 17, no. 3, pp. 265–289, 2008, doi: 10.1007/s00521-007-0120-z.

Ontology-Based Decision Support System for Subfertility – A Case Study on Female Subfertility

Thenuka
Computing and Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Ratnapura
thenuka94@gmail.com

Vasanthapriyan
Computing and Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Ratnapura
priyan@appsc.sab.ac.lk

Banujan
Computing and Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Ratnapura
bhakuha@appsc.sab.ac.lk

Abstract— **Decision-making in the treatment method of subfertility of a female is a vast area than the male subfertility in the gynecological section. Information technology-based decision-making helps doctors, nurses, and medical students to increase the quality of service for their subfertility patients in their routine work, studies, and further researches. It assists to find the causes and actual treatment method as a recommendation by using the current medical status of the patients. Ontology is selected because it suits to encapsulate the term of concepts and their relationship and specify modeling primitives. Domain knowledge for the subfertility of a female is gained from doctors and medical students. Protégé Ontology Editor 5.5 is used to implement the terms and concepts in the subfertility of a female. Evaluation of developed ontology evaluated by ontology experts, OOPS, DL Query, and SPARQL Query tools. Decision Support System (DSS) is developed using the owl file. The system is validated by doctors and medical students.**

Keywords— ontology, female, subfertility, decision support system

I. Introduction

Failure to conceive after 12 months of unprotected and regular sexual intercourse is called as subfertility [1]. Childlessness is a major life problem. Human beings' subfertility is an average monthly rate of only 20%. An average of 10-15% of couples visits at least once for fertility care specialists who have difficulties in conceiving as primary or secondary [2]. In the gynecological section, at first, males are checked for childlessness causes. Then, females need to check for their childlessness cause. Subfertility of a female is dependent on two factors that are the age of the woman and the period of childlessness. Couples need to wait 3-5 years for the investigation until the treatment. Diagnostic tests are done as part of three stages. Specialists use screening tools for patients who need further treatment [3]. As the first stage-specific assessment and after that make a diagnosis to predict results with or without treatment. Ovulatory Disorders, Tubal Disorders, Uterine Problems, and unexplained factors are the subsets of diagnostic causes [4].

Ontology is a basic form of knowledge representation of a real-world thing. Ontology is a formal which is represented in a specific formal language, explicit which describes proper assumptions written here and specification which covers the specific domain's artifacts of a shared conceptualization of a specific domain [5]. Ontology's main components are concepts, attributes, and relationships. Concepts usually refer to classes of objects. Attributes represent the features of the objects. The relationship between the concepts is represented by properties. In advance, properties of relation, the cardinality of relations, rules, axioms, and restrictions are depicted in the ontology. The ontology domain defines all the guidelines of the decision-making process [6]. Ontology is selected because ontology has common terminology and formal semantics, it provides consistency checking, integrating heterogeneous data from various sources, reuse of the ontology, interoperability, and taking benefits from existing ontology [7].

Sometimes, this concept is expressed by using various terminologies because of the incomplete, unstructured, general nature, and different formats of the information, and the knowledge are not reaching everybody. Further, computers need to understand the meaning or semantics of the information clearly. The semantic web enables this understanding of computers. Ontologies are a powerful mechanism for representing knowledge presented in the semantic web. Therefore, ontology can be used to find a response to queries within a specified context in the domain of subfertility of females [8].

Decision Support System (DSS) helps to decision making in related knowledge for human. These DSS used widely in most of the decision making processes. All these DSS are focused on some critical parts of the domain and customized for that. DSS uses various methodologies and ontology modeling is rarely used in developing DSS [9].

Evaluation of ontology and DSS also needed for error-free and actual decision making. Evaluation and validation of ontology are usually done by OOPS! Evaluation Editor to find pitfalls, DL queries and SPARQL Queries used to find correctness of the ontology modeling and ontology experts check the developed ontology is correct or not. DSS is checked by domain experts [10].

This paper focuses to provide a DSS for decision-making in female childlessness treatment by a specialist. By developing a DSS that helps to find causes to treatment methods for childlessness for females. Ontology helps to decision making without no ambiguity in the related domain that describes and organizes clearly [11].

The base to the top-level details was collected from medical students and doctors in the decision-making process. This would help the doctors and their assistants in the gynecological department in hospitals to decision making to find the causes and treatment methods and also to gain knowledge for medical students.

The objective of this paper is the presentation of a DSS for decision making using ontology modeling in the subfertility domain. This paper is structured as follows; Section 2 describes related works of literature review, section 3 clearly explained the methodology and experimental design, section 4 provides the results of the ontology domain, section 5 presents the evaluation and methods of the proposed system and section 6 contains the conclusion and future work.

II. Related Work

Ontology-based medical domain research papers were mainly used to get knowledge about the research area. Other than that, fertility and subfertility papers, Ontology-based

DSS papers were used for literature study. Some related works:

Clinical pregnancy failure after 12 months of regular sexual intercourse with unprotected is called as subfertility. Subfertility is a disease that general term describes the failure to get clinical pregnancy after twelve months of unprotected & regular sexual intercourse. Ovulatory Problem, Tubal Disorders, Uterine Abnormalities, Endometriosis, and in advance female's age are general causes of subfertility in females [12].

S Vasanthapriyan and K Banujan presented Ontology modeling of dental extraction forceps and Knowledge Management (KM) Portal. Here, grounded theory was used for data collection. An OWL ontology language is used for ontology modeling. This paper gave clear concepts of properties and their relationships. This KM portal assists hospitals, dental students in knowledge sharing, and learning practices [13].

P. C. Sherimon and team developed Clinical DSS (CDSS) for Diabetic patients. In that paper, ontology design and modeling, implementation, and the reasoning process are depicted in the OntoDiabetic CDSS. An OWL ontology language is used for the implementation of domain concepts. OntoDiabetic DSS calculates and predicts the score of diabetics which becomes a risk for the patients by alcohol, smoking, cardiovascular disease, sexual and physical activity. This provides care methods suggestions and precautions to the patients [6].

This paper is about developed DSS to find causes and treatment methods for cancers. Existing disease ontology is reused for the system to interlinking data and strengthen the knowledge. It enhances the reasoning ability of the cancer treatment of DSS. Here, the patient's condition is analyzed to find the stage of cancer. Based on previous stages indexed in the Case-Based Reasoning (CBR) database and the search results will be returned to the doctors to utilize as references in cancer diagnosis [14].

DSS for Subfertility using ontology is not available. By using these concepts and the domain knowledge, So, We intend to develop the Ontology-Based DSS.

III. Design And Methodology

Modeling ontology has various methodologies and it requires appropriate methods and tools. Moreover, constructing a scientific domain ontology from the beginning is a complicated task. In the literature review, all the methodologies of ontology development and evaluation were reviewed. After that, Grüninger and Fox's methodology [15] was selected and used for the ontology modeling for female subfertility. The methodology of the female subfertility DSS using ontology is shown in Fig. 1.

This methodology describes in three parts of our research. Those are modeling ontology for subfertility of female, development of DSS for decision making in the treatment method of female subfertility and evaluation, and validation of developed ontology and DSS of female subfertility.

A. Data Collection

Data was collected using grounded theory. An extensive literature survey in fertility & subfertility and field expert collaboration was used to get relevant data. Those medical experts are interviewed and brainstormed for our purpose. One Doctor and three medical students with extensive knowledge in this field and two experts of ontology engineering helped to design the ontology [16].

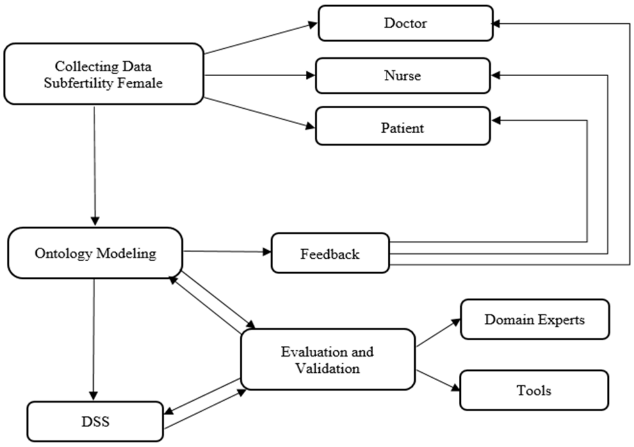
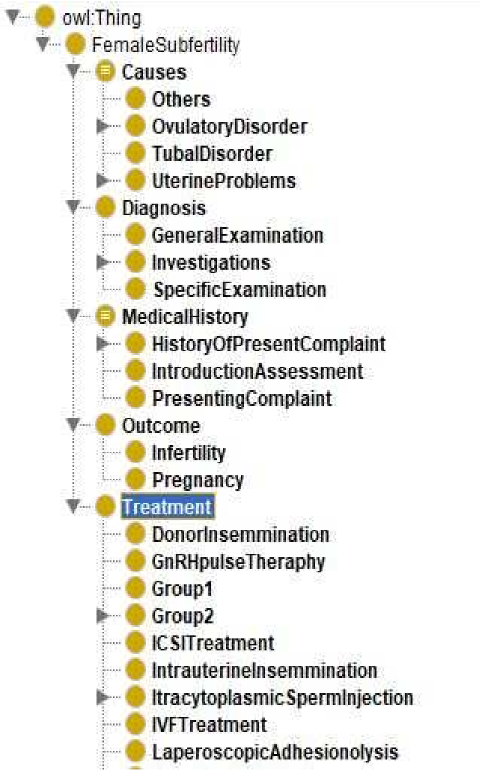


Fig. 1. Methodology

B. Ontology Modelling

After collection of the information from the above-mentioned interviewees, The Competency Questions (CQs) were developed. Ontology modeling using some axioms to answer the set of competency questions. Here, CQs are used as requirement specification of the domain. Ontology tries to answer the competency questions using Description Logics [17]. Table 1 shows some of the competency questions. After the creation of CQ and answers, and Ontology hierarchy was developed. The ontology hierarchy was developed by the knowledge of finding causes and treatment methods of the subfertility of a female. Fig. 2 describes the high-level hierarchy of ontology for subfertility by using the competency questions which is developed to help the decision making in the treatment method process of the subfertility of the female.



```
owl:Thing
  FemaleSubfertility
    Causes
      Others
      OvulatoryDisorder
      TubalDisorder
      UterineProblems
    Diagnosis
      GeneralExamination
      Investigations
      SpecificExamination
    MedicalHistory
      HistoryOfPresentComplaint
      IntroductionAssessment
      PresentingComplaint
    Outcome
      Infertility
      Pregnancy
    Treatment
      DonorInsemmination
      GnRHpulseTheraphy
      Group1
      Group2
      ICSITreatment
      IntrauterineInsemmination
      ItracytoplasmicSpermInjection
      IVFTreatment
      LaperoscopicAdhesionolysis
```

Fig. 2. High-level hierarchy of subfertility ontology

C. DSS

Ontologies are the best example of depict decision-making methods in the semantic web. Best reasoning capabilities are presented using ontology and semantic web technologies. We present the ontology-based DSS to the treatment method of the subfertility of females in this part. This system was built by using the Java and JENA API distributed component environment [18].

TABLE I. SOME OF THE COMPETENCY QUESTIONS

Information needs for Subfertility of Female
What are the causes of subfertility?
What is medical history need to gather?
Which tests used to find premature ovarian failure?
What is the treatment method for tubal disorder?
What are the types of tubal disorder?
List out all the tests for causes?
What is the basic information need to get from the patient?
What are the reasons for uterine problems?
What are the general examinations done for patients?

Our DSS design consists of Experience Sharing, Ontology, Knowledge Retrieval, Storage, and Reasoning Layer and is shown in Fig. 3. Decision support system developed using ontology file (RDF/XML). We fetched data from ontology file using JENA API also Springboot used to develop the backend and ReactJS used to develop the front-end using JAVA. Whole system developed using these tools and technologies.

We have implemented as a starting point of specifying the DSS. Importantly, our presented concepts, properties, and relationships here are found according to the characteristics of the hospital's decision-making environment [17].

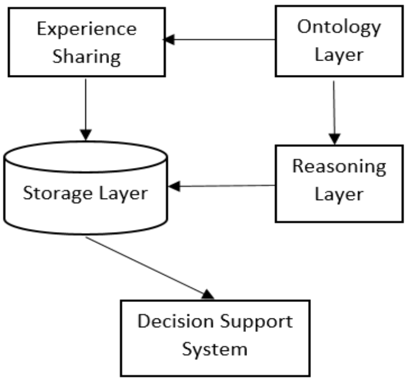


Fig. 3. The architecture of the DSS

- Ontology Layer – Female subfertility is described using the ontology layer which contains, classes, individuals, relationships, domain rules, axioms, etc. Ontology Layer is develop Using the Protégé Ontology Editor 5.5, these individuals and their relationships.
- Experience Sharing Layer –Doctors share their knowledge about female subfertility to diagnose the causes and treatments by the Experience Sharing layer. The domain knowledge is used for the ontology layer.
- Storage Layer – Here, Triple-store storage is used for the storage layer which stores RDF triples and is queried by SPARQL query. Jena API was selected to utilize in this research because Jena API is a part of Jena for RDF storage and query. It is used to get high performance of the RDF store and it supports Jena API ranges.
- Reasoning Layer – Combination of OWL Language and Rule Markup Language is called Semantic Web Rule Language (SWRL). By using the terms of OWL classes and relationships of instances is depends on rules in SWRL reasoning. Moreover, those rules define more complex relationships and constraints between relationships & concepts. Decision-making rules were created with Protégé SWRL Editor. This is a plugin in Protégé and it is used for the above-mentioned purpose.
- DSS Layer – We show how our ontology can be used to decision making of treatment method in the DSS by using the semantic web technologies. Jena API ontology models are used for all the models read by the API and system developed for decision making.

IV. RESULTS AND FINDINGS

A. Ontology Modelling

Ontology model developed using the software Protégé Ontology Editor 5.5. It is developed by the collected knowledge of classification under the main concepts of the decision-making method. Following Fig. 4 describes the part of Ontograf of developed ontology which shows how to make decisions using ontology.

Ontology is based on the relationship of their concepts. Table 2 describes the associative relationship and their inverse relationship with some concepts.

TABLE II. ASSOCIATIVE RELATIONSHIP AND THEIR INVERSE RELATIONSHIP

Concept	Relationship	Concept
Causes	findby, finds	Diagnosis
MedicalHistory	hasSelect, isSelect	Causes
Causes	hasTreat, isTreatFor	Treatment
SpecificExamination	examineBefore, examineAfter	Investigation
Laperoscopy	hasObstructions, isObstructions	TubalDisorder

B. Development of DSS

DSS developed using the above-mentioned DSS architecture using Java Language and Jena API Rule. This system shows how the system is used for decision making in subfertility [17].

V. EVALUATION AND VALIDATION

The evaluation and validation of the DSS done after ontology modeling evaluation and validation. The ontology model's evaluation and validation are done separately. The evaluation and validation have been done separately. OOPS! Evaluation editor which is a web-based tool used to evaluate the ontology model. All the critical, important, and minor errors were found by using this editor and all the critical errors were rectified. Fig.5. shows the critical, important, and minor issues of developed ontology. After that, we discussed some competency questions which were used to collect the details of the domain, and by using that competency questions that developed ontology's correctness is evaluated by the DL query.

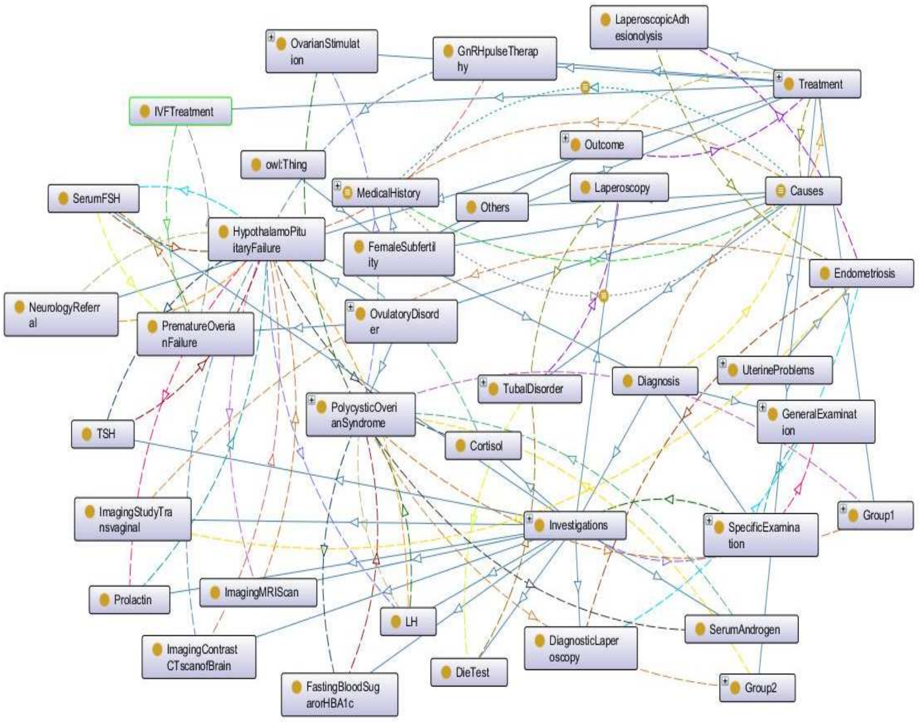
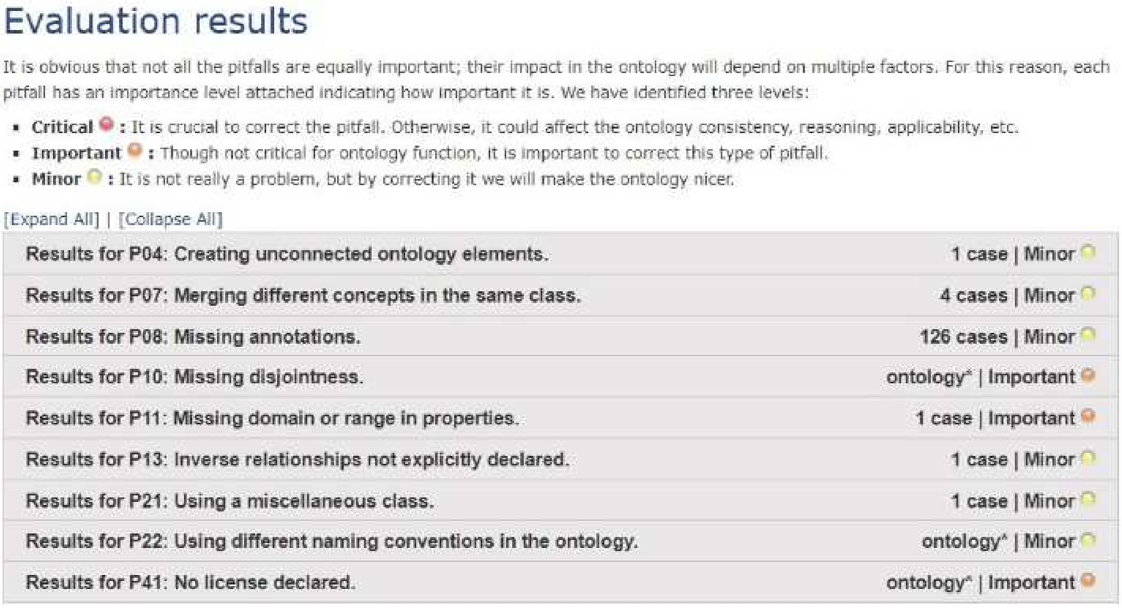


Fig. 4. Part of Ontograf of developed ontology



Evaluation results

It is obvious that not all the pitfalls are equally important; their impact in the ontology will depend on multiple factors. For this reason, each pitfall has an importance level attached indicating how important it is. We have identified three levels:

- **Critical** ● : It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.
- **Important** ● : Though not critical for ontology function, it is important to correct this type of pitfall.
- **Minor** ○ : It is not really a problem, but by correcting it we will make the ontology nicer.

[Expand All] | [Collapse All]

Results for P04: Creating unconnected ontology elements.	1 case | Minor ○
Results for P07: Merging different concepts in the same class.	4 cases | Minor ○
Results for P08: Missing annotations.	126 cases | Minor ○
Results for P10: Missing disjointness.	ontology* | Important ●
Results for P11: Missing domain or range in properties.	1 case | Important ●
Results for P13: Inverse relationships not explicitly declared.	1 case | Minor ○
Results for P21: Using a miscellaneous class.	1 case | Minor ○
Results for P22: Using different naming conventions in the ontology.	ontology* | Minor ○
Results for P41: No license declared.	ontology* | Important ●

Fig. 5. OOPS! Evaluation Results

Table 3 describes DL Queries and answers. FaCT++ reasoner is used to validating using the DL query. FaCT++ reasoned is an inbuilt tool of Protégé Ontology Editor 5.5. SPARQL query is also used to evaluate the developed ontology model. False feeding terms of ontology modeling and relationships were corrected by ontology experts. DSS are evaluated by some doctors and medical students. They gave some comments and suggestions if any tweaks need to do in the DSS and all the tweaks mentioned by them were tweaked.

Ontology experts need to check the syntax, vocabulary, structure, semantics, context and representation of the ontology model. Ontology model evaluated by two ontology experts. By their comments and suggestions, ontology modeling was redeveloped and corrected.

TABLE III. DL QUERIES AND ANSWERS

Competency Questions	DL Query	Answers
What are the treatment method for Tubal Disorder Cause?	TubalSurgery and hasSurgeries only TubalSurgery	Labaroscopic_Tubal_Surgery Surgery_for_hydrosalpinges Tubal_catheterisation_or_cannulation Tubal_microsurgery
What are the medical history that helps to find Tubal Disorder?	TubalDisorder and is select some MedicalHistory	Scar_Tissue Pelvic_Inflammatory_Disease History_of_Gonorhea_or_Chlaymydia

The quality and level of satisfaction of domain experts such as doctor and medical students who evaluate the DSS for their decision making in treatment method. The DSS hosted in a private hospital of Jaffna district and also provided the system to medical students.

This system usage method training session provided to doctors and medical students. We went to the hospital and clearly explained how to use the system to the doctors. We arranged zoom meeting because of the COVID19 pandemic period to medical students and described how to use the system for their studies. Their user satisfaction, correctness, user friendly and etc details were gathered from both doctors and medical students using questionnaire. Five point Likert – type questionnaire is used to check the satisfaction and quality level. Following shows the level of agreements of the questionnaire.

i. Strongly Disagree

ii. Disagree

iii. Moderately Agree

iv. Agree

v. Strongly Agree

Score assigned by following: Strongly Disagree- 1, Disagree -2, Moderately Agree – 3, Agree – 4, Strongly Agree – 5. Participant's profile details, work details and job experiences also gathered when collecting the user satisfaction level of questionnaire.

Ontology modelling and DSS was able to give the following user advantages:

i. Express the full knowledge about female subfertility

ii. Support decision making in treatment method

iii. Support to medical studies in gynecological area

iv. User satisfaction

The time period of data collection is limited for twenty days. This collection was performed by 5 doctors, 10 medical students and two lecturers were participated in the evaluation. Most of the domain expert's Likert scale came between 4 and 5. Table IV shows some of the questions of questionnaire outcome.

TABLE IV. SOME QUESTIONS AND OUTCOME

Question	Mode	Conclusion
Is all the information of female subfertility provided in diagnosis and treatment method?	4	Agree
Is the decision making process correct?	4	Agree
Can the system helps to the doctors and medical students?	5	Strongly Agree
Is the female subfertility norm is maintained or not?	5	Strongly Agree

VI. CONCLUSION

Female subfertility is a vast area than male subfertility in the subfertility area. So Normally, Doctors and medical students have complexity in decision making. This DSS helps to decision making of finding causes and treatment. Designing an ontology model and DSS is not a simple task. Difficulties faced by the researcher is the need to gain clear knowledge about subfertility. These terms and all the connections are very new and tedious. It includes female subfertility treatment method concepts, their properties such as object property and data property, and their relationships. We confidently believe that our female subfertility DSS can help the gynecological area, doctors and medical students, and other active researchers in this field to improve not only decision-making and also the knowledge sharing and experiences. In the future, we planned to develop a DSS with some critical treatment methods of subfertility and infertility.

ACKNOWLEDGMENT

We pay our heartfelt gratitude to the Doctor Shinthuja Ranjit of Nation Hospital of Sri Lanka. Further, we acknowledge medical students, Miss. Karthiga Rajasingam, Mr. Sithambarabanan Kathirgamanathan, Mr.Thanistan Raj, Mr. Ajith Jerom Mathuranayagam, and the other domain experts who have directly and indirectly tangled to finish the research work effectively and successfully.

REFERENCES

[1] U. Larsen and D. Ph, "Research on infertility : which definition should we use ?," vol. 83, no. 4, pp. 0-6, 2005.

[2] A. G. Obstet, A.-t. Anti-adhesions and G. Expert, "Adhesions and endometriosis : challenges in subfertility management," pp. 16-18, 2016.

[3] C. Gnoth, E. Godehardt, P. Frank-Herrmann, K. Friol, J. Tigges and G. Freundl, "Definition and prevalence of subfertility and infertility," Human Reproduction, vol. 20, no. 5, pp. 1144-1147, 2005.

[4] G. D. Adamson, "Subfertility : causes , treatment and outcome," pp. 169-185, 2003.

[5] A. I. Walisadeera, G. N. Wikramanayake and A. Ginige, "An Ontological Approach to Meet Information Needs of Farmers in Sri Lanka," pp. 228-240, 2013.

[6] P. C. Sherimon and R. Krishnan, "OntoDiabetic: An Ontology-Based Clinical Decision Support System for Diabetic Patients," Arabian Journal for Science and Engineering, pp. 1145-1160, 2016.

[7] P. Delir Haghighi, F. Burstein, A. Zaslavsky and P. Arbon, "Development and evaluation of ontology for intelligent decision support in medical emergency management for mass gatherings," Decision Support Systems, vol. 54, no. 2, pp. 1192-1204, 2013.

[8] E. Kontopoulos, G. Martinopoulos and D. Lazarou, "An ontology-based decision support tool for optimizing domestic solar hot water system selection," Journal of Cleaner Production, vol. 112, no. 2016, pp. 4636-4646, 2020.

[9] M. Choraś, R. Kozik, A. Flizikowski and M. W. HołuboChoraś, "Ontology applied in decision support system for critical infrastructures protection," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 671-680, 2010.

[10] H. Satria, R. S. Priya, L. H. Ismail and E. Supriyanto, "Building and Reusing Medical Ontology for Tropical Diseases Management," vol. 6, no. 1, 2012.

[11] A. Kokossis, L. Jim, A. Moreno and D. Ria, "An Ontology-Based Knowledge Management Platform," no. May 2014, 2003.

[12] J. B. Stanford, G. L. White and H. Hatasaka, "Timing Intercourse to Achieve Pregnancy : Current Evidence," vol. 100, no. 6, pp. 1333-1341, 2002.

[13] S. Vasanthapriyan and K. Banujan, "An ontological approach for dental extraction decision making and knowledge dissemination," Journal of Computer Science, pp. 832-843, 2019.

[14] Y. Shen, J. Colloc, A. Jacquet-Andrieu, Z. Guo and Y. Liu, "Constructing ontology-based cancer treatment decision support system with case-based reasoning," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 278-288, 2018.

[15] M. Gruninger and M. Fox, "Methodology for the Design and Evaluation of Ontologies," Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal, 1995.

[16] M. Gawich, M. Alfonse, M. Aref and A.-B. M. Salem, "Developing a System for Medical Ontology Evolution," Egyptian Computer Science Journal, vol. 41, no. 2, pp. 53-62, 2017.

[17] S. Vasanthapriyan, J. Tian, D. Zhao, S. Xiong and J. Xiang, "An ontology-based knowledge management system for software testing," Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE, pp. 230-235, 2017.

[18] I. T. Afolabi, G. E. Olujinmi and R. O. Nwokoye, "Ontology Based Decision Support System for Youth Counseling," vol. 0958, 2019.

LSTM and FFNN based Exchange Rate Prediction Model

Mauran Kanagarathnam
Department of Computing &
Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
kmauran@appsc.sab.ac.lk

Vasanthapriyan S.
Department of Computing &
Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
priyan@appsc.sab.ac.lk

Banujan Kuhaneswaran
Department of Computing &
Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
bhakuha@appsc.sab.ac.lk

Prasanth S.
Department of Physical Sciences and
Technology
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
senthanprasanth007@gmail.com

Abstract— **The exchange rate is the value of one country's currency concerning another country's currency. Exchange rates are determined by the foreign exchange market. In the economic state, the currency exchange rate of every country tends to vary from time to time. So, it can be said that the exchange rate affects the countries economical system. As we all know, the USD is used by most countries and it is taken as a reference for the exchange rate. So, it is an important indicator for everyone to predict the exchange rate in advance corresponding to their country. During this research, one of the variants of Artificial Neural Network (ANN) has been employed to fulfill the above purpose. ANN can be learned and adapted by the environment quickly which ability makes the ANN could be categorized into various types and each type consists of many models. In this research study, Long Short-Term Memory (LSTM) and the Feed Forward Neural Network (FFNN) have been used to predict the exchange rate of three currencies namely Sri Lankan rupees, Japanese yen, and Chinese yuan concerning the USD. The aforementioned neural network algorithms were compared against each other for finding the best algorithm to develop the final prediction model. There are several performance evaluation methods available among those, accuracy and Mean Absolute Error (MAE) were used to evaluate the prediction performance. Ten years of past exchange data have been incorporated in this regard. LSTM has shown the best performance out of the two algorithms. Finally, the LSTM model has been used for the development of the prediction model.**

Keywords— Exchange Rate, Economic System, ANN, LSTM, FFNN

I. Introduction

Every country in the world usually carries out trading activities with other countries. Countries buy goods and services from other countries. While making a trade, it is important to consider how currencies of one country can be exchanged for currencies of other countries. Here, the Exchanging of currencies making the trade easier. An exchange rate is the value of a country's currency concerning another country's currency or economic zone. The government of any country has the right to change the exchange rate when needed. Exchange rates are used in international markets, finance, trading, and investment. Appreciation and depreciation are the words used in a change in the price of a currency. Appreciation occurs when a currency becomes more valuable or more expensive. Depreciation is the opposite of appreciation. Depreciation occurs when a currency becomes less valuable or less expensive. Exchange rates may change from time to time. Exchange rates are commonly governed by forces called supply and demand. Purchasing goods and services require the conversion of the local currency of a country concerning the currency of other countries which involves trade.

USD/EUR = 1.25 = It means 1 euro buys 1.25 US dollars

It is very important to track the exchange rates for doing business with other countries. Because only trading at the right time with the right approach can help to gain expected profit. In the absence of required knowledge in trading, investors and businessmen may incur heavy losses. Prediction of the exchange rate is a critical factor to achieve success in many businesses. Future movement of exchange rates can be predicted using the records of the exchange rate. These predictions are useful for decision making in the financial sector. Changes in exchange rates can have a significant impact on most economies.

In this study, ANN has been used to predict the exchange rate. The idea of ANN is based on the belief that working as a human brain. ANN is composed of multiple nodes which imitate biological neuron of the human brain. The nodes take input data and perform simple operations on the data. The results of the operations are would be passed on to the other neurons. The output gained from each node is called its activation or node value. There are two types of topologies on artificial neural networks. Those are feedforward and feedback.

Past data has been utilized with the ANN model to train the network. Initially, the data are pre-processed and analyzed. Then the data are divided into three sets namely training, testing, and validation. The neural network model has been trained using the training portion of the data. For adjusting the hyperparameters validation part of the data has been utilized. For evaluating the model testing part of the data has been incorporated.

II. Literature Review

M. Bonilla et al. have done research using Artificial Neural Network (ANN) to predict the US dollar against Sri Lankan rupee (USD/LKR) with a higher accuracy level. In this study, Static and Dynamic neural network architectures were considered. These neural network models are compared with each other and evaluated to build the prediction model. The best prediction model produced a result with 76%

accuracy. The limitation of this study is only the feedforward and the time delay neural network architectures were considered [1].

A. Emam et al. have done research using Artificial Neural Network (ANN) to predict the daily fluctuations of foreign exchange rates. To Overcome the shortcoming of the traditional forecasting techniques ANN was used [2].

M. Ismail et al. have researched the forecast the exchange rate of the US dollar expressed in terms of Malaysian ringgit. This research is conducted by considering two techniques: Artificial Neural Network (ANN) and Autoregressive Integrated Moving Average (ARIMA) time series. The mean squared error has been considered as an evaluation metric[3].

M. Markova et al. have researched forecasting the exchange rate using Nonlinear Autoregressive with Exogenous Input (NARX) Neural Network models to predict Euro against the US dollar currency. Matlab is used to model the neural network and used for the evaluation. In this research, neural network models were implemented using different parameters like the number of neurons in the hidden layer, time delay, and training algorithm. The best validation performance for the network is 1.9771E-5 and the regression R-value is 0.99617. In future work, they have proposed to increase the accuracy of the prediction model by considering more parameters and to decrease the time consumed in the process, they have proposed to reduce the memory consumption[4].

S. Nayakovit et al. have done research on exchange rate forecasting based on U.S. Dollar with Great Britain Pounds by using Artificial Neural Networks (ANN) and Meiosis Genetic Algorithms (MGA). Several models are compared with each other such as ANN, ARIMA, Linear regression (LR), Random Walk (RW), and Regularized Least Squares Time Series (RLS-TS). According to the evaluation, The ANN+MGA model provides better performance. The accuracy of the model is about 25-40% more than other methods [5].

D. Y. Perwej et al. have researched to examine the effects of several important neural network factors on model fitting and forecasting the behaviors. The number of inputs and hidden nodes and the size of the training sample is changed to get better results on predicting the Indian Rupee (INR) against US Dollar (USD) rate using Artificial Neural Network (ANN). In this paper, they have analyzed the Above mentioned factors to find out how they affect the model [6].

S. Ranjit et al. have researched different machine learning techniques namely Artificial Neural Network (ANN), Recurrent Neural Network (RNN) to develop a prediction model between Nepalese Rupees against three major currencies Euro, Sterling Pound, and US dollar. Prediction model was implemented based on different RNN architectures, feed-forward ANN compared against backpropagation algorithm and then checked for evaluating the accuracy of each model. Different trials were performed by changing the number of hidden neurons until the best result was found. Add more parameters on input for more accurate prediction and also minimize time, space complexity for processing ANN architecture as a future improvement [7].

S. Ranjit et al. have researched foreign exchange market prediction using neural network and sentiment analysis. The neural network model has been developed by considering certain parameters such as the number of neurons, the use of bias neurons, the number of hidden layers, activation functions, and training methods. Root Mean Squared Error (RMSE) was found to be 0.0034 with 6 hidden nodes using ANN. Sentiment analysis was applied using a combination of Naïve Bayes and a lexicon-based algorithm to analyze the opinion of different traders and predict the overall sentiment. In this research, Sentiments are taken from tweets and were classified as positive or negative. In sentiment analysis, accuracy was found to be 90.625% [8].

M. Rout et al. have researched the efficient prediction of the long-range exchange rates using Radial Basis Function Neural Network(RBFNN). In this research, models have been employed to predict currency exchange rates among US dollar, Indian Rupees, and Japanese Yen. The results obtained through RBFNN compared against Multilayer Layer Neural Network (MLANN) and Functional Link Artificial Neural Network (FLANN) [9].

M. L. R. Torregoza et al. have researched on forecasting Philippine Peso to US Dollar exchange rate. The Forecasting accuracy of an artificial neural network is highly dependent on the volume of the training data. So in this paper, an alternative algorithm that will increase the accuracy of the conventional artificial neural network with a limited volume of training data is presented and analyzed. Finally, they have proposed, if a large volume of training data has produced will result in higher accuracy[10].

III. METHODOLOGY

According to the study, the prediction model was implemented to predict the exchange rate. The approach (Fig 1) has followed to find the predicted value of the exchange rate. This chapter represents the overall description where different analyses have been done to prepare the dataset. This further describes the implementation and evaluation methods.

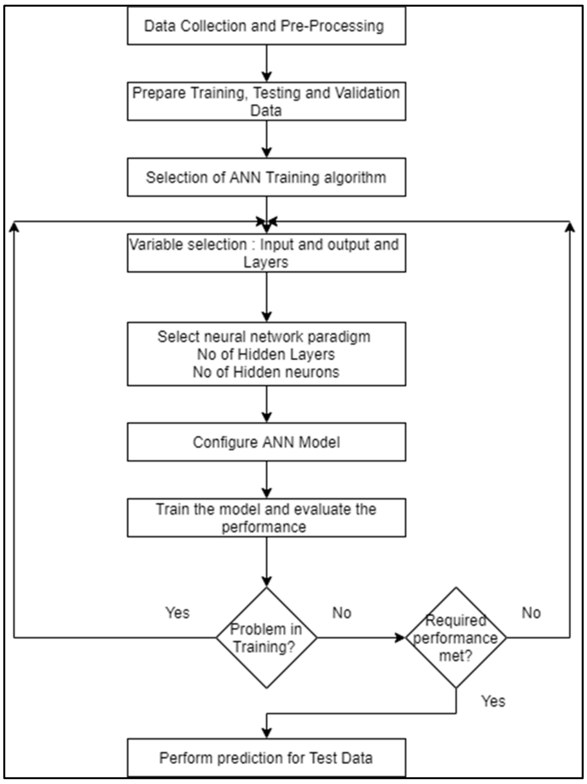

Data Collection and Pre-Processing
Prepare Training, Testing and Validation Data
Selection of ANN Training algorithm
Variable selection : Input and output and Layers
Select neural network paradigm
No of Hidden Layers
No of Hidden neurons
Configure ANN Model
Train the model and evaluate the performance
Yes
Problem in Training?
No
Required performance met?
No
Yes
Perform prediction for Test Data

Fig. 1. Proposed approach

A. *Dataset and Requirement analysis*

Requirement analysis is the first step to analyze the expectation and the flow of any project. Dataset was collected from investing.com. Nearly ten years of exchange rate data were gathered from the website and analyzed during the research.

The closing price is the last level at which it has been traded on any given day. The price of the first trade for any stock will be its opening price daily.

Following are taken as an input for the research,

- Yesterday's closed price
- Yesterday's open price
- Yesterday's low price
- Yesterday's high price

And today's closed price was considered as an output of the research study.

B. *Data pre-processing*

Data pre-processing is the method of extracting missing, noisy, or inaccurate data from the data set. It is a required step to build any prediction model. Normalization is one of the pre-processing methods. Standardization was used to reduce the deviation between the available data and to bring the attributes within a specified range. During the study, pre-processing data was carried out using the "python pandas library".

- Remove unnecessary columns
- Checking for null values
- Normalizing the data set within fixed ranges
- Normalized Value = (Actual Value – Minimum Value) / (Maximum Value – Minimum Value)
- Scaling the data in a standardized way

C. *Training and Testing Model*

The dataset was divided into a training set, testing set and validation set the predicted value was compared with the actual value. To do so, normalization was done to change the whole dataset in the same type of value. MAE was calculated to measure the performance of the model (Equation 1).

$$\text{MAE} = \sum_{i=1}^{n} \frac{|y_i - x_i|}{n} \qquad (1)$$

y_i - The predicted value of the exchange rate

x_i - The actual value of the exchange rate

n – Number of actual or predicted values

IV. Results and Analysis

A. *Results obtained from the LSTM model*

The following Tables I, II, and III are the readings of the MAE and the readings of LSTM for the Sri Lankan rupee, Chinese yuan, and Japanese yen respectively. According to the result of the MAE measures the average magnitude of errors in a set of predictions, without taking into account their direction. It is the average over the test sample of the absolute differences between the prediction and the actual observation where all the differences are of equal weight. MAE readings of LSTM, the best prediction has fifty neurons for the hidden layer 1 and fifty neurons for the hidden layer 2.

TABLE I. Result of Sri Lankan rupee

Number of hidden neurons		**Epochs**	**MAE**
Hidden Layer 1	***Hidden Layer 2***		
40	40	10	1.30646
50	50	10	1.10085
60	60	10	1.88307
40	40	50	2.742
50	50	50	1.754
60	60	50	2.812

TABLE II. Result of Chinese yuan

Number of hidden neurons		**Epochs**	**MAE**
Hidden Layer 1	***Hidden Layer 2***		
40	40	10	0.04296
50	50	10	0.02349
60	60	10	0.03317
40	40	50	0.0397
50	50	50	0.0216
60	60	50	0.0356

TABLE III. Result of the Japanese yen

Number of hidden neurons		**Epochs**	**MAE**
Hidden Layer 1	***Hidden Layer 2***		
40	40	10	0.6408
50	50	10	0.4602
60	60	10	0.6390
40	40	50	0.6179
50	50	50	0.4245
60	60	50	0.598

The following figures are describing the prediction and actual values of the Sri Lankan rupee, Chinese yuan, and Japanese yen (Fig 2, Fig 3, and Fig 4 respectively).

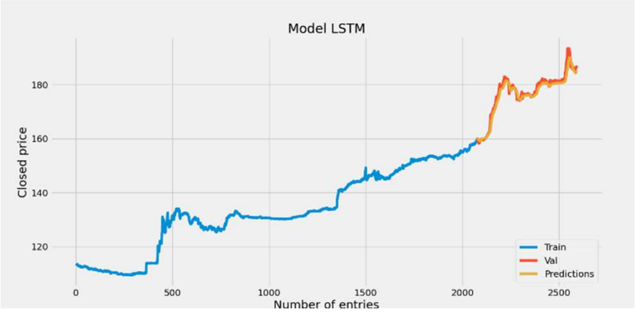


Fig. 2. Best prediction graph of Sri Lankan Rupee for LSTM

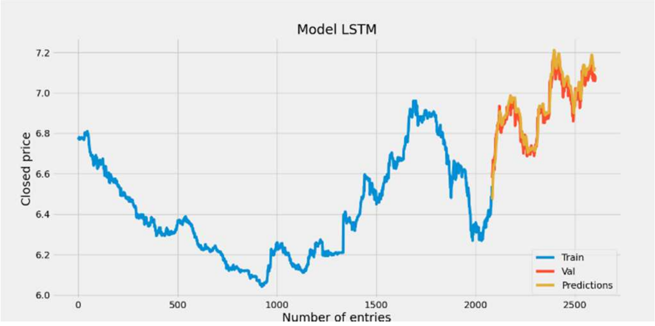


Fig. 3. Best prediction graph of Chinese Yuan for LSTM

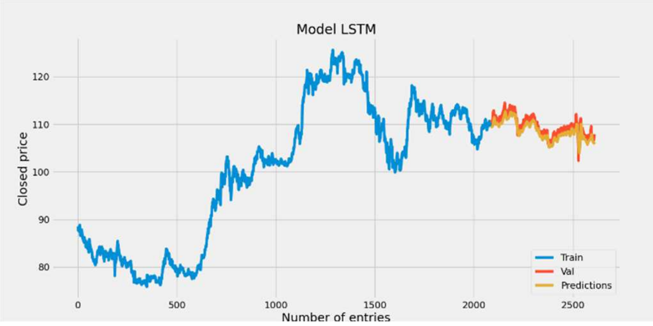


Fig. 4. Best prediction graph of Japanese Yen for LSTM

B. *Results obtained from the feed-forward neural network model*

Following Table IV, Table V, and Table VI are the readings of the MAE readings of FFNN for Sri Lankan rupee, Chinese yuan, and Japanese yen respectively.

TABLE IV. RESULT OF SRI LANKAN RUPEE

Number of hidden neurons		**Epochs**	**MAE**
Hidden Layer 1	***Hidden Layer 2***		
60	60	10	5.787
70	70	10	4.19
80	80	10	6.78
60	60	50	5.94
70	70	50	4.53
80	80	50	6.35

TABLE V. RESULT OF CHINESE YUAN

Number of hidden neurons		**Epochs**	**MAE**
Hidden Layer 1	***Hidden Layer 2***		
60	60	10	1.26
70	70	10	1.12
80	80	10	1.24
60	60	50	3.467
70	70	50	3.124
80	80	50	4.16

TABLE VI. RESULT OF THE JAPANESE YEN

Number of hidden neurons		**Epochs**	**MAE**
Hidden Layer 1	***Hidden Layer 2***		
60	60	10	2.018
70	70	10	1.965
80	80	10	1.994
60	60	50	2.21
70	70	50	2.013
80	80	50	2.38

According to the result of the MAE readings of FFNN, the following figures (Fig 5, Fig 6, and Fig 7) are describing the prediction and actual values of the Sri Lankan rupee, Chinese yuan, and Japanese yen.

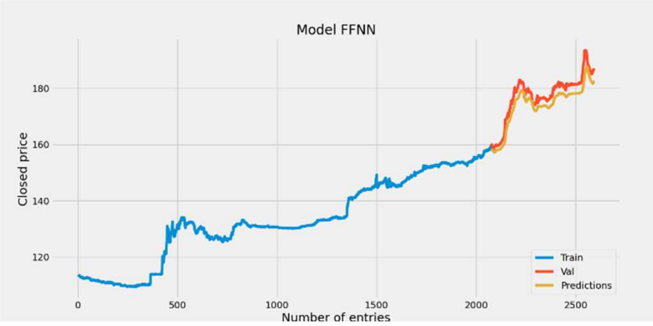


Fig. 5. Best prediction graph of Sri Lankan Rupees for FFNN

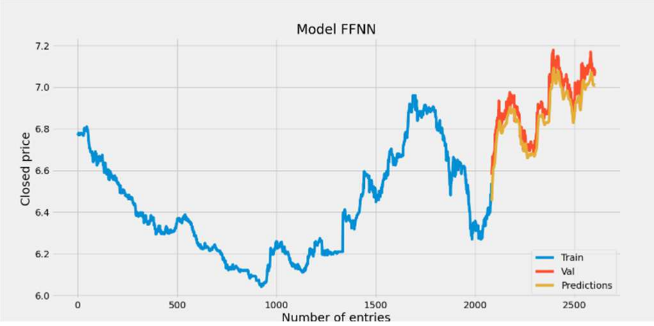


Fig. 6. Best prediction graph of Chinese Yuan for FFNN

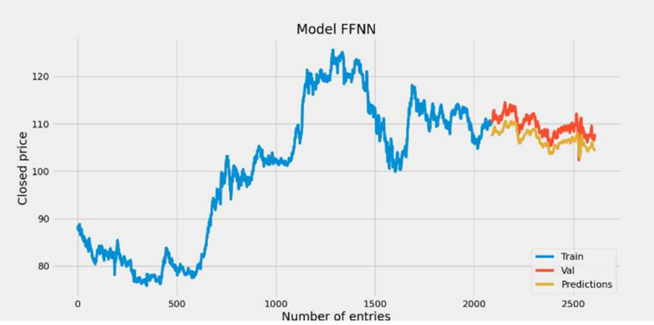


Fig. 7. Best prediction graph of Japanese Yen for FFNN

The graph shown in Fig 2 to Fig 7 were describing the flow of the actual and predicted value of the exchange rate. The red-colored flow shows the actual value of the exchange rate and the yellow-colored flow shows the predicted value of the exchange rate. The results describe the difference between actual and predicted values.

V. CONCLUSION

According to that, the LSTM model performs better than the FFNN model. So, Here the LSTM was shown the best result among the comparison of two different models. In a conclusion, the LSTM model can be used to predict the exchange rate.

The research can be further expanded by comparing other algorithms. And the prediction accuracy will be high with tuning the parameters. Further, develop the prediction model for another neural network algorithm. Collecting more historical data and dividing the Training and testing dataset with different percentages may helpful for further improvement of the research. The exchange rates may vary from time to time. For example, exchange rate values fell during the pandemic period. So, none of the models are perfect at all times. So, the study can be further analyzed more in different situations.

REFERENCES

[1] M. Bonilla, P. Marco, and I. Olmeda, “Forecasting Exchange Rates Volatilities Using Artificial Neural Networks,” pp. 57–68, 2000, doi: 10.1007/978-3-642-57652-2_4.

[2] A. Emam and H. Min, “The artificial neural network for forecasting foreign exchange rates,” *Int. J. Serv. Oper. Manag.*, vol. 5, no. 6, pp. 740–757, 2009, doi: 10.1504/IJSOM.2009.026772.

[3] M. Ismail, N. Z. Jubley, and Z. M. Ali, “Forecasting Malaysian foreign exchange rate using artificial neural network and ARIMA time series,” *AIP Conf. Proc.*, vol. 2013, 2018, doi: 10.1063/1.5054221.

[4] M. Markova, “Foreign exchange rate forecasting by artificial

neural networks," *AIP Conf. Proc.*, vol. 2164, no. October 2019, doi: 10.1063/1.5130812.

[5] S. Nayakovit, U. Jaritngam, and K. Khantanapoka, "Prediction exchange rate of USD/GBP with intelligence cyberspace experimental," *ICEIE 2010 - 2010 Int. Conf. Electron. Inf. Eng. Proc.*, vol. 2, no. ICEIE, pp. 15–19, 2010, doi: 10.1109/ICEIE.2010.5559706.

[6] Y. Perwej, "Forecasting Of Indian Rupee (INR) / US Dollar (USD) Currency Exchange Rate Using Artificial Neural Network," *Int. J. Comput. Sci. Eng. Appl.*, vol. 2, no. 2, pp. 41–52, 2012, doi: 10.5121/ijcsea.2012.2204.

[7] S. Ranjit, S. Shrestha, S. Subedi, and S. Shakya, "Comparison of algorithms in Foreign Exchange Rate Prediction," *Proc. 2018 IEEE 3rd Int. Conf. Comput. Commun. Secur. ICCCS 2018*, pp. 9–13, 2018, doi: 10.1109/CCCS.2018.8586826.

[8] S. Ranjit, S. Shrestha, S. Subedi, and S. Shakya, "Foreign Rate Exchange Prediction Using Neural Network and Sentiment Analysis," *Proc. - IEEE 2018 Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2018*, no. Icacccn, pp. 1173–1177, 2018, doi: 10.1109/ICACCCN.2018.8748819.

[9] M. Rout, B. Majhi, and U. M. Mohapatra, "Efficient long-range prediction of exchange rates using Radial Basis Function Neural Network models," *IEEE-International Conf. Adv. Eng. Sci. Manag. ICAESM-2012*, pp. 530–535, 2012.

[10] M. L. R. Torregoza and E. P. Dadios, "Comparison of neural network and hybrid genetic algorithm-neural network in the forecasting of Philippine Peso-US Dollar exchange rate," *2014 Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag. HNICEM 2014 - 7th HNICEM 2014 Jt. with 6th Int. Symp. Comput. Intell. Intell. Informatics, co-located with 10th ERDT Conf.*, no. November 2014, doi: 10.1109/HNICEM.2014.7016218.

SherLock 1.0: An Extended Version of 'SherLock' Mobile Platform for Fake News Identification on Social Media

MDPP Goonathilake
Department of Computer Science
General Sir John Kotelawala Defence University
Sri Lanka
pathumveyron24@gmail.com

PPNV Kumara
Department of Computer Science
General Sir John Kotelawala Defence University
Sri Lanka
nandana@kdu.ac.lk

***Abstract*— SherLock is a CNN, RNN-LSTM based mobile platform to verify and fact-check information on social media. Today, false news is easily created and distributed across many social media platforms. Due to that, people find it difficult to choose between right or wrong information on those platforms. Therefore, a strong need emerges to develop a fact-checking platform to overcome this problem. Fact-checking means the process of verifying information. An extended version of SherLock mobile platform has presented from this study to verify information on social media including many features. CNN, RNN-LSTM based hybrid model ables to capture the high-level features and long-term dependencies from the input text. Some of the features of the mobile application includes fact-checking, daily news updates, news reporting and social media trends etc. The mobile platform is developed using Flutter as the front-end framework and Firebase as the back-end framework including REST APIs to gather daily news articles. The hybrid model achieved a 92% accuracy when checking the information circulating on social media.**

Keywords—Fake News Detection, Fact-Checking, Deep Learning, Natural Language Processing, Hybrid Approach

I. INTRODUCTION

Today out of 8 billion people worldwide 3.8 billion people are social media users. With the development of new social media platforms, people are moving from traditional news media to those social media platforms. Because they can easily get to know about the things which are happening around the world just surfing through social media news feeds. Because of the freedom and simplicity gives from those social media platforms anyone can express anything at any time and this leads to create and distribute false information comfortably through those platforms. Due to that, people find it difficult to choose between right and wrong information on those social media platforms.

Most of the time due to the lack of verified news sources and fact-checkers on social media platforms in Sri Lanka social media users are failed to identify false information in their news feeds. Because of that, people are sharing those news stories without checking and this leads to spread lot of misinformation through social media platforms in Sri Lanka during the year of 2019. For example, after the Easter Sunday attacks in Sri Lanka government decided to block social media access due to alleged false information circulating on social media occurring lot of misunderstanding between people and religions [1]. Not only that, before and after the presidential election in Sri Lanka in 2019, a lot of false information has created and spread through social media platforms to change the mindset and opinion of people [2].

Many instances have recorded from different countries regarding the widespread impact of false information on social media platforms. During the US presidential election in 2016, "Pizzagate" fake news is widely spread on Twitter by creating more than one million tweets [3] and during the Gubernatorial campaign in Jakarta back in 2016 governor, Ahok sentenced to two years in prison for criticizing a verse of Quaran [4].

These incidents have clearly indicated that a strong need emerges to develop a fact-checking platform to overcome this problem. From this study, we present an extended version for mobile platform called 'SherLock' that can use as a platform for fact-checking on social media using a CNN, RNN-LSTM based hybrid model [5] and including many other features such as daily news updates, news reporting and social media trends etc. 'SherLock 1.0' is the newest and updated version of 'SherLock' mobile application including several new improvements [6].

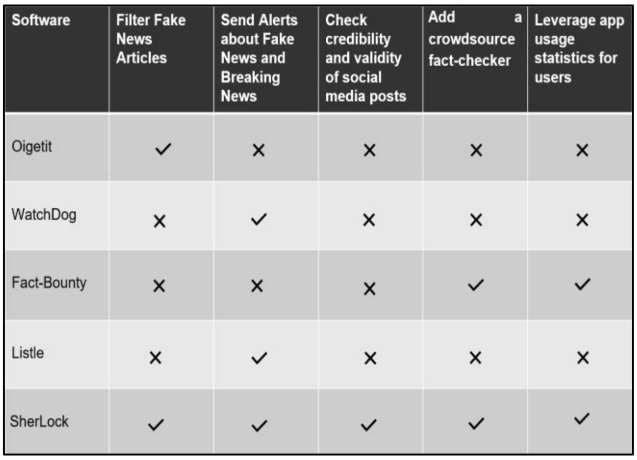
II. LITERATURE REVIEW

Hoaxy [7] is an online tool that can collects and tracks misinformation. Then it can visualize that misinformation by using the technologies like web scraping, web syndication, Twitter API and RSS parser. FakeNewsTracker [8] is another tool to collect and visualize false news on social media using some of the deep learning mechanisms like LSTM, autoencoder etc. Using Hoaxy API, a tool called 'dEFEND' has developed to provide a news propagation network including trending news and top claims. It also can provide some explainable insight into user comments on Twitter. News Verify [9] has developed to detect the credibility of news using the techniques like feature extraction, sentiment analysis and web crawling etc. Authors [10] have developed an extension for both Chrome and Mozilla browsers called 'B.S. Detector' to check unreliable sources against a manually compiled list of domains.

According to [11] they used machine learning techniques to detect false information and proposed a method using word based n-grams. Authors [12] have proposed a method using WEKA machine learning classifiers. Fake News Pattern Detector [13] used a deep learning network to detect the patterns in fake news by applying several techniques like CNN, word2vec word embeddings and feature extraction etc. TRACEMINER [14] also used a LSTM-RNN model to provide high classification accuracy on real-world data sets. Authors [15] have proposed several deep learning networks to detect false information. Rather than using classical models, they proposed a combination of neural networks to use in fake news detection.

978-0-7381-4403-0/20/$31.00 ©2020 IEEE

With that, a CNN, RNN-LSTM based hybrid model has chosen as the method to develop the main feature which is the fact-checking feature of the proposed mobile application. From the next chapters of the paper we present the design and implementation, technology adopted, testing and evaluation and how system works regarding the proposed mobile application. Figure 1 represents some of the related software and comparison of their features.



Software	Filter Fake News Articles	Send Alerts about Fake News and Breaking News	Check credibility and validity of social media posts	Add a crowdsource fact-checker	Leverage app usage statistics for users
Oigetit	✓	✕	✕	✕	✕
WatchDog	✕	✓	✕	✕	✕
Fact-Bounty	✕	✕	✕	✓	✓
Listle	✕	✓	✕	✕	✕
SherLock	✓	✓	✓	✓	✓

Fig. 1. Comparison of features in related software

III. DESIGN AND IMPLEMENTATION

A. *The High-Level Architecture of the System*

The proposed system consists of several parts including MVVM architecture which represents the model-view-view-model pattern of the mobile application. The mobile application has developed including many features namely, daily news updates, fact-checking, news reporting, social media news trends and daily COVID-19 report. Then cloud database has used by including several crud operations for each feature of the mobile app. REST APIs and web scraping methods have used to collect information from different news sources to build the hybrid deep learning model. After that, the same cloud database has used to store the hybrid deep learning model. Figure 2 represents the overall system architecture of the proposed system. As for the front-end of the proposed system developed a mobile application including the above-mentioned features. And for the back-end of the proposed system used a cloud database to store the hybrid deep learning model and included several crud operations according to the features of the mobile application.

B. *Software Process Model of the System*

The incremental software process model is consisted of breaking down requirements into subsystems and modules [16]. Therefore, the proposed system applied a software process model as Incremental.

Figure 3 indicates three subsystems of the main feature of the proposed solution. The first subsystem includes data collection part to develop deep learning model and next subsystem includes checking the social media posts using the hybrid deep learning model. And the final subsystem is about checking the status of the posts using the mobile application. If it is verified news it represents using green colour and if it is fake news it represents in red colour.

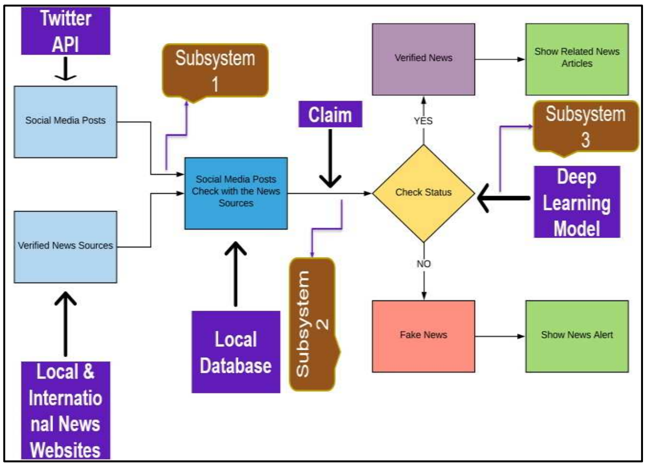


Fig. 3. Software process model for the main feature

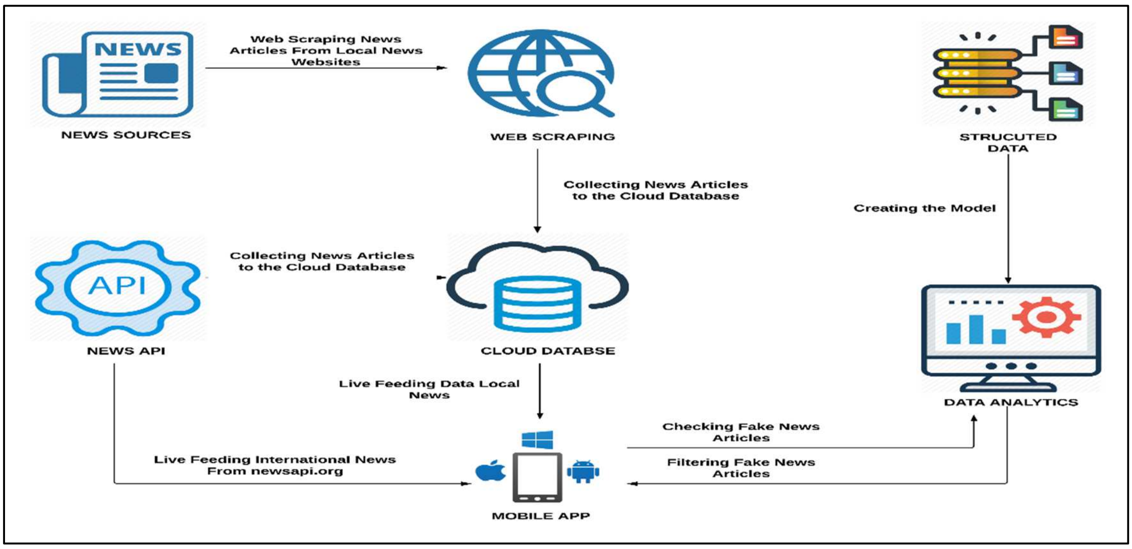
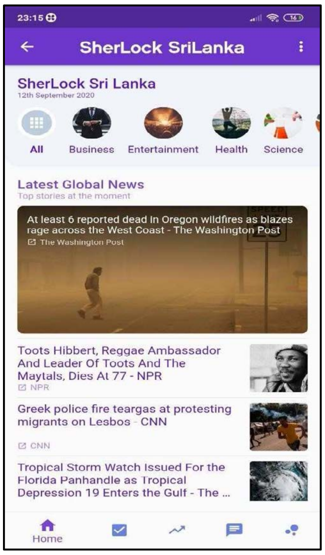

NEWS
NEWS SOURCES
Web Scraping News Articles From Local News Websites
WEB SCRAPING
STRUCUTED DATA
Collecting News Articles to the Cloud Database
Creating the Model
API
Collecting News Articles to the Cloud Database
NEWS API
CLOUD DATABSE
DATA ANALYTICS
Live Feeding Data Local News
Checking Fake News Articles
Live Feeding International News From newsapi.org
Filtering Fake News Articles
MOBILE APP

Figure 2. High-level architecture of the proposed system

C. Design of the Proposed Mobile Application

Design of the proposed mobile application includes several screens of the mobile app such as onboarding screen, registration screen, login screen, home screen and other feature-wise screens etc. But here mentioned only the screens of the main features of the mobile application.

Figure 4 represents the home screen of the mobile application which shows the latest global news around the world by categorizing news for business, entertainment, health, science and sports etc. If the user wants to know more information about news, they can click on the news article and it navigates to a URL where it includes more information about the news article.



At least 6 reported dead in Oregon wildfires as blazes rage across the West Coast - The Washington Post
The Washington Post

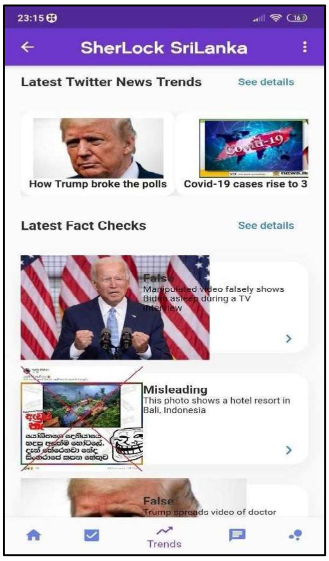
Toots Hibbert, Reggae Ambassador And Leader Of Toots And The Maytals, Dies At 77 - NPR
NPR

Greek police fire teargas at protesting migrants on Lesbos - CNN
CNN

Tropical Storm Watch Issued For the Florida Panhandle as Tropical Depression 19 Enters the Gulf - The ...

Fig. 4. The latest global news screen

Figure 5 indicates the latest news trending on Twitter and the latest fact checks from the websites like AFP news. If the user wants to know more information like the above screen they can click and navigate to find more information on each news.



How Trump broke the polls

Covid-19 cases rise to 3

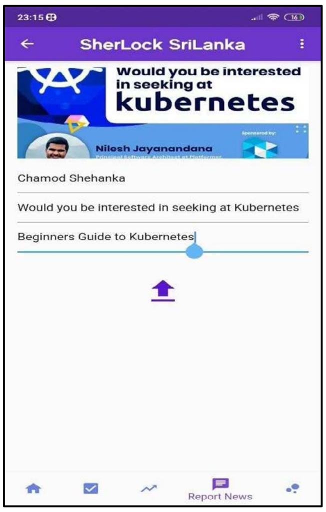
False
Manipulated video falsely shows Biden asleep during a TV interview

Misleading
This photo shows a hotel resort in Bali, Indonesia

False
Trump spreads video of doctor

Fig. 5. Latest twitter trending news screen

Figure 6 represents the news reporting screen where news reporters can report the news to the platform.



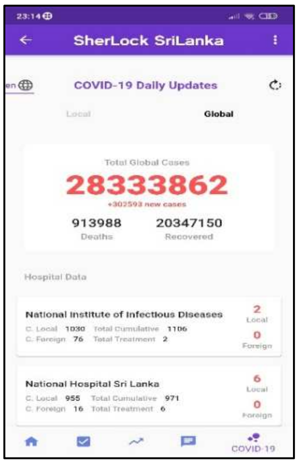
Chamod Shehanka

Would you be interested in seeking at Kubernetes

Beginners Guide to Kubernetes

Fig. 6. News reporting screen

Figure 7 represents the daily COVID-19 report screen where users can see information according to their language preferences. Localization has added to the feature for languages English, Sinhala and Tamil.



Total Global Cases
28333862
+302593 new cases
913988 Deaths
20347150 Recovered

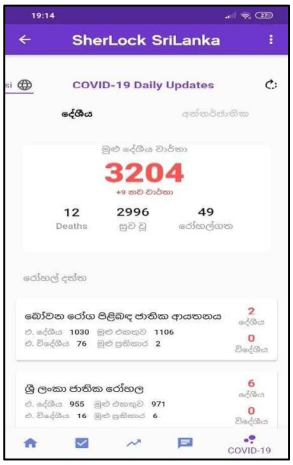
Hospital Data

National Institute of Infectious Diseases
C. Local 1030 Total Cumulative 1106
C. Foreign 76 Total Treatment 2
2 Local
0 Foreign

National Hospital Sri Lanka
C. Local 955 Total Cumulative 971
C. Foreign 16 Total Treatment 6
6 Local
0 Foreign

Fig. 7. COVID-19 updates using English language

Figure 8 represents the daily COVID-19 report screen in the Sinhala language.



මුළු දේශීය වාර්තා
3204
+9 නව වාර්තා
12 Deaths
2996 සුව වූ
49 රෝහල්ගත

රෝහල් දත්ත

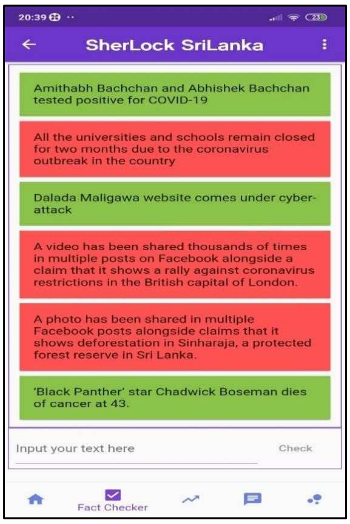
බෝවන රෝග පිළිබඳ ජාතික ආයතනය
එ. දේශීය 1030 මුළු එකතුව 1106
එ. විදේශීය 76 මුළු ප්‍රතිකාර 2
2 දේශීය
0 විදේශීය

ශ්‍රී ලංකා ජාතික රෝහල
එ. දේශීය 955 මුළු එකතුව 971
එ. විදේශීය 16 මුළු ප්‍රතිකාර 6
6 දේශීය
0 විදේශීය

Fig. 8. COVID-19 updates using Sinhala language

Figure 9 represents the main feature of the mobile application which is about the fact-checking of social media posts. User can input the text of the social media post and it classifies as verified news or fake news from the hybrid model.

As an example, if the user inputs the text as 'Dalada Maligawa website comes under cyber-attack' then from the hybrid model it checks and gives a message as verified news by showing the text in green colour. If the user inputs a message like 'All the universities and schools remain closed for two months due to the coronavirus outbreak in the country' then from the hybrid model it labels as a fake one by showing the red colour with the text. This screen is used to fact-check the social media posts using the CNN, RNN-LSTM based hybrid model.



🟩 Amithabh Bachchan and Abhishek Bachchan tested positive for COVID-19

🟥 All the universities and schools remain closed for two months due to the coronavirus outbreak in the country

🟩 Dalada Maligawa website comes under cyber-attack

🟥 A video has been shared thousands of times in multiple posts on Facebook alongside a claim that it shows a rally against coronavirus restrictions in the British capital of London.

🟥 A photo has been shared in multiple Facebook posts alongside claims that it shows deforestation in Sinharaja, a protected forest reserve in Sri Lanka.

🟩 'Black Panther' star Chadwick Boseman dies of cancer at 43.

Fig. 9. Fact-checking screen

IV. TECHNOLOGY ADOPTED

A. *Use Flutter as Front-End Framework*

The reason for using Flutter [17] as the front-end framework of the proposed mobile application is to build beautiful, natively compiled mobile application for both Android and IOS from a single codebase. The best thing about Flutter is free and open source by providing native interfaces. When compare with other mobile development frameworks Flutter has a fast development methodology by providing some easy infrastructure for developers.

B. *Use Firebase as Back-End Framework*

For the proposed system, had to deal with a lot of unstructured data types. Therefore, choose Firebase [18] as the back-end framework of the proposed system. Firebase authentication has used to authenticate users from the login screen. And Firebase Database and Firestore have used to store news reports. Then Firebase Storage has used to store images of the news reports. Finally, used Firebase Machine Learning to store the hybrid deep learning model.

C. *Use Scrapy to collect data*

Before building the hybrid deep learning model collected data from different news sources using the web scraping method. For that, Scrapy [19] which is an open-source web scraper is used.

D. *Use TensorFlow to build the model*

The hybrid deep learning model has used TensorFlow [20] to build the model and Python as the programming language. Because it has rich support to both front-end and back-end frameworks and it has a flexible and comprehensive ecosystem. Figure 10 represents overall technology map of the system.

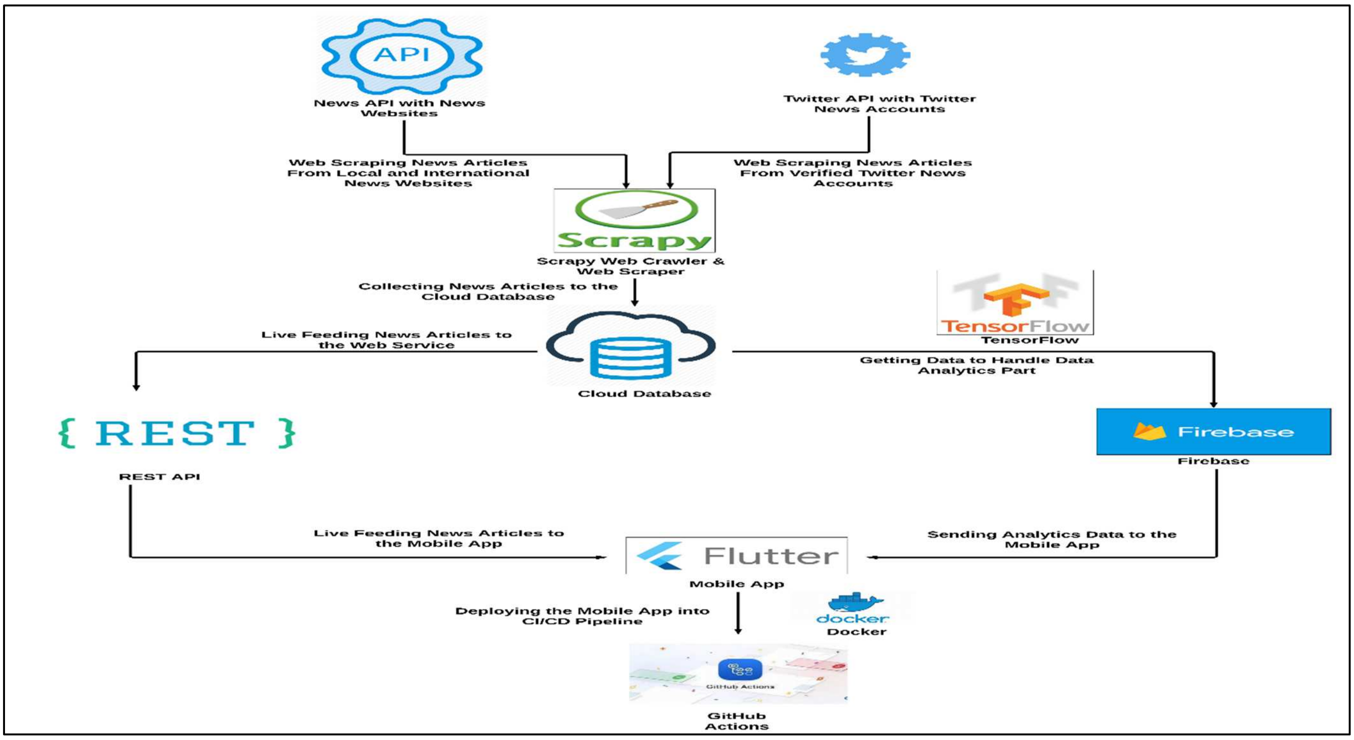

API
News API with News Websites
Twitter API with Twitter News Accounts
Web Scraping News Articles From Local and International News Websites
Web Scraping News Articles From Verified Twitter News Accounts
Scrapy
Scrapy Web Crawler & Web Scraper
Collecting News Articles to the Cloud Database
TensorFlow
TensorFlow
Live Feeding News Articles to the Web Service
Cloud Database
Getting Data to Handle Data Analytics Part
{ REST }
REST API
Firebase
Firebase
Live Feeding News Articles to the Mobile App
Flutter
Sending Analytics Data to the Mobile App
Mobile App
Deploying the Mobile App into CI/CD Pipeline
docker
Docker
GitHub Actions
GitHub Actions

Fig. 10. Technology map for the proposed system

E. Use GitHub to maintain the project

For the proposed system, GitHub [21] used to maintain the project repository. And also GitHub Actions [22] used to build the CI/CD(Continuous Integration/Continuous Delivery) pipeline and deploying the mobile application.

F. Use IntelliJ IDEA as a platform to build the project

IntelliJ IDEA [23] has used as the platform to develop the mobile application because it has rich support to both front-end framework Flutter and back-end framework Firebase as well.

G. Use PyCharm as a platform to build the model

PyCharm [24] has used as the platform to build the data collection method. After that added some of the Natural Language Processing techniques like tokenization, preprocessing and word embeddings etc. Finally, build the hybrid model after doing the above steps.

H. Use Kaggle kernels/Google Colab to run the model

The hybrid deep learning model needed more power to execute with the dataset size and the complexity of the model. Therefore, choose Kaggle kernels [25] and Google Colab [26] cloud environments with built-in TPUs to run the model.

V. TESTING AND EVALUATION

Test driven development methodology has used throughout the project. Different automated tests help to ensure the performance of the mobile app. Unit tests have used to test the functions of the mobile app. And widget tests have applied on UIs to test the widgets of the mobile app. Several API test automation scripts have run using Postman [27] to make sure that all the end points are working fine with different scenarios. After that, integration tests and end-to-end tests have used to test the complete mobile app. GitHub Actions CI/CD pipeline used to run tests automatically when pushing new code changes to the repository.

As for the evaluation of the mobile application, Docker [28] has used to enhance the performance of the app. Then app bundle packages have applied to reduce the size of the app. The proposed mobile application is compatible on both Android and IOS versions when using Flutter as the front-end framework. After giving the mobile app to use as a complete system to some users for testing we found that the user satisfaction and usability rate is high when compared it with other related software in figure 1.

VI. HOW SYSTEM WORKS

The proposed mobile application has several main features such as daily news updates, fact-checking, news reporting, social media news trends and daily COVID-19 report.

First, the user needs to register to the system by providing correct details. Only the admin user has full access to all features of the mobile app. Other users have access to specific defined features only. Which means according to the target stakeholders, the news agency has the full access of the system and other users like daily news reporters, daily news readers and social media users can access specific defined features only. As an example, daily news reporters have the access to news reporting screen but other two users do not have access to that screen. For restricting the access for specific features, role-based authorization has used to manage users. Let's see how the main features of the proposed mobile application work.

A. Latest Global News Feature

After logging to the system by giving correct credentials user navigates to the home page where user able to see the latest global news articles. Users can navigate between different categories of news according to their preferences. This screen shows the latest global news around the world by categorizing news for business, entertainment, health, science and sports etc. If the user wants to know more information about news, they can click on the news article and it navigates to a URL where it includes more information about the news article. This feature is visible to all the users. News API [29] has used here to fetch live news articles to the mobile application using a JSON API.

B. Fact-Checking Feature

After navigating to the home screen user able to see the bottom navigation bar. From that, the user can shift between different features. After clicking the fact-checking feature user navigates to a screen where they can fact-check social media posts using the CNN, RNN-LSTM based hybrid model. User can input the text of the social media post and it classifies as verified news or fake news from the hybrid model.

As an example, if the user inputs the text as ‘Dalada Maligawa website comes under cyber-attack’ then from the hybrid model it checks and gives a message as verified news by showing the text in green colour. If the user inputs a message like ‘All the universities and schools remain closed for two months due to the coronavirus outbreak in the country’ then from the hybrid model it labels as a fake one by showing the red colour with the text.

The hybrid model has an accuracy of 92% and it includes several deep learning mechanisms. CNN, RNN-LSTM based hybrid model ables to capture high-level features and long-term dependencies from the input text. For more information about the CNN, RNN-LSTM hybrid model refers to the following study which is done by the authors [5].

This feature is accessible only for social media users to check the validity and credibility of the social media post. TensorFlow used to develop the hybrid model then Firebase machine learning used to store the model and REST APIs [30] used to interact with the model with front-end framework Flutter.

C. Latest Twitter News Trends and Latest Fact-Checkings Feature

This feature is visible for all the users. From this feature, users can get to know about the latest Twitter trends and latest fact-checkings. If the user wants to know more information about particular news trends, they can click on the news article and it navigates to the post where it includes more information about the trending news articles. Twitter API [31] is used to get the latest Twitter trends and web sites like AFP Fact Check used to get the latest fact checkings using REST APIs.

D. News Reporting Feature

If the logged-in user is a news reporter, they can easily navigate to the news reporting feature from the bottom navigation bar. News reporters can upload news to the system by providing the correct details such as reporters name, the title of the news, description of the news and related photo of the news. This feature is only visible to news reporters only. Firebase Database and Firestore has used to store news reports and Firebase storage has used to store related photos of the news articles.

E. Daily COVID-19 Report

User can navigate to this feature from the bottom navigation bar by clicking the last item of it. This feature represents the daily COVID-19 updates in Sri Lanka and the world. Here, the user can see information according to their language preferences. Information mainly includes total global and local cases, total deaths, total recovered and hospital data. Localization has added to the feature for languages English, Sinhala and Tamil. This feature also visible to all the users. REST APIs have used to fetch live updated data from the cloud hosted database to the mobile application and Google Translate has used to add the localization for Sinhala and Tamil languages.

VII. CONCLUSION AND FUTURE WORK

In conclusion, we have shown that the proposed mobile application is an efficient solution for Sri Lanka. Because it provides many features such as daily news updates, fact-checking, news reporting, social media news trends and daily COVID-19 report. And the main feature of fact-checking is more important for daily social media users to find between real and fake news articles and social media posts as well.

As for future work, the suggested mobile application hope to launch as a complete system to the public with versions for both Android and IOS. With that, hope to publish this mobile application on the Android Play Store and Apple App Store.

ACKNOWLEDGMENT

The authors of this research paper would like to acknowledge the tremendous support given by the supervisors for their assistance and feedback on this research paper.

REFERENCES

[1] "Sri Lankan Government Blocks Social Media Access Over Alleged Fake News," *Reason.com*, Apr. 22, 2019. https://reason.com/2019/04/22/sri-lankan-govt-blocks-social-media-access-over-alleged-fake-news/ (accessed Sep. 12, 2020).

[2] J. C. Wong, "Sri Lankans fear violence over Facebook fake news ahead of election," *The Guardian*, Nov. 12, 2019.

[3] "Pizzagate conspiracy theory," *Wikipedia*. Sep. 23, 2019, Accessed: Oct. 28, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pizzagate_conspiracy_theory&oldid=917379104.

[4] K. Lamb, "Jakarta governor Ahok sentenced to two years in prison for blasphemy," *The Guardian*, May 09, 2017.

[5] MDPP Goonathilake, and PPNV Kumara, "CNN, RNN-LSTM Based Hybrid Approach to Detect State-of-the-Art Stance-Based Fake News on Social Media," in *2020 20th International Conference on ICT for emerging regions (ICTer).*", in press.

[6] MDPP Goonathilake, and PPNV Kumara, "SherLock: A CNN, RNN-LSTM Based Mobile Platform for Fact-Checking on Social Media," in *2020 13th International Research Conference of General Sir John Kotelawala Defence University (KDUIRC).*", in press.

[7] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Hoaxy: A Platform for Tracking Online Misinformation," in *Proceedings of the 25th International Conference Companion on World Wide Web*, Republic and Canton of Geneva, Switzerland, 2016, pp. 745–750, doi: 10.1145/2872518.2890098.

[8] K. Shu, D. Mahudeswaran, and H. Liu, "FakeNewsTracker: a tool for fake news collection, detection, and visualization," *Comput. Math. Organ. Theory*, vol. 25, no. 1, pp. 60–71, Mar. 2019, doi: 10.1007/s10588-018-09280-3.

[9] "Real-Time News Certification System on Sina Weibo." https://dl.acm.org/citation.cfm?id=2742571 (accessed Oct. 28, 2019).

[10] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proc. Assoc. Inf. Sci. Technol.*, vol. 52, no. 1, pp. 1–4, 2015, doi: 10.1002/pra2.2015.145052010082.

[11] H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, 2017, pp. 127–138.

[12] M. Aldwairi and A. Alwahedi, "Detecting Fake News in Social Media Networks," *Procedia Comput. Sci.*, vol. 141, pp. 215–222, Jan. 2018, doi: 10.1016/j.procs.2018.10.171.

[13] "Fake News Patterns Detector." http://fakenews.mit.edu/ (accessed Oct. 28, 2019).

[14] L. Wu and H. Liu, "Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, New York, NY, USA, 2018, pp. 637–645, doi: 10.1145/3159652.3159677.

[15] R. K. Kaliyar, "Fake News Detection Using A Deep Neural Network," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Dec. 2018, pp. 1–7, doi: 10.1109/CCAA.2018.8777343.

[16] "Software Engineering | Incremental process model," *GeeksforGeeks*, May 28, 2018. https://www.geeksforgeeks.org/software-engineering-incremental-process-model/ (accessed Sep. 13, 2020).

[17] "Flutter - Beautiful native apps in record time." https://flutter.dev/ (accessed Sep. 13, 2020).

[18] "Firebase," *Firebase*. https://firebase.google.com/ (accessed Sep. 13, 2020).

[19] "Scrapy | A Fast and Powerful Scraping and Web Crawling Framework." https://scrapy.org/ (accessed May 31, 2020).

[20] "Keras | TensorFlow Core," *TensorFlow*. https://www.tensorflow.org/guide/keras (accessed May 31, 2020).

[21] "Build software better, together," *GitHub*. https://github.com (accessed Sep. 13, 2020).

[22] "Features • GitHub Actions," *GitHub*. https://github.com/features/actions (accessed Sep. 13, 2020).

[23] "IntelliJ IDEA: The Java IDE for Professional Developers by JetBrains." https://www.jetbrains.com/idea/ (accessed Sep. 13, 2020).

[24] "PyCharm: the Python IDE for Professional Developers by JetBrains." https://www.jetbrains.com/pycharm/ (accessed Sep. 13, 2020).

[25] "Kaggle: Your Home for Data Science." https://www.kaggle.com/ (accessed Sep. 13, 2020).

[26] "Google Colaboratory." https://colab.research.google.com/notebooks/ (accessed Sep. 13, 2020).

[27] "Postman | The Collaboration Platform for API Development." https://www.postman.com/ (accessed Nov. 17, 2020).

[28] "Empowering App Development for Developers | Docker." https://www.docker.com/ (accessed Sep. 14, 2020).

[29] "News API - A JSON API for live news and blog articles," *News API*. https://newsapi.org (accessed Sep. 13, 2020).

[30] "REST APIs: An Introduction | IBM." https://www.ibm.com/cloud/learn/rest-apis (accessed Sep. 13, 2020).

[31] "Use Cases, Tutorials, & Documentation." https://developer.twitter.com/en (accessed Sep. 13, 2020).

A Review on Type II Diabetes Prediction using Machine Learning Techniques

P. D. M. Peiris
Department of Information Technology
Faculty of Information Technology
University of Moratuwa
Sri Lanka
dilini.peiris@outlook.com

H. M. S. C. R. Heenkenda
Department of Information Technology
Faculty of Information Technology
University of Moratuwa
Sri Lanka
sankani.ruchirani@gmail.com

***Abstract*— Diabetes is a salient non-commutable disease from which many patients all over the world suffer from regardless of their race, gender and age. Patients with diabetes tend to be affected by many other diseases such as strokes, nerve damage, heart and kidney diseases etc. When considering the mortality rates and life expectancy rates of this disease, it is clear that predicting this disease will contribute to the reduction or even prevention in its early stages. This review paper evaluates various research work that has attempted predicting diabetes mellitus in general and type-2 diabetes using various machine learning techniques such as artificial neural networks, support vector machine, random forest, decision tree etc. The motive of this review paper is to evaluate the selected research work and to suggest the best technique that can be used in clinical facilities to help medical personnel to predict the disease.**

Keywords— diabetes mellitus, type-2 diabetes, machine learning, prediction, forecasting, early detection

I. Introduction

Machine Learning (ML), a branch of Artificial Intelligence, gives machines the ability to perform intelligent tasks. Among them, predicting the future by analysing the past data and identifying patterns in them is a prominent feature. Diabetes is one of the most common diseases among many adults today, not to mention that it is rising among teenagers and children as well. Therefore, prediction of diabetes before actually diagnosing it later can be advantageous in the early prevention of the disease.

This review paper focuses on researches conducted on predicting type-2 diabetes mellitus, using ML techniques during the years 2011-2020. The research papers are gathered from the databases of Science Direct, Google Scholar. The review paper compares and contrasts different ML techniques used in the research works that have been filtered out to be analysed, and discusses the most suitable techniques among them to be used in Clinical Decision Support Systems and other Information Systems to support doctors in predicting type-2 diabetes (T2DM) successfully among their patients.

II. Background

A. Diabetes Mellitus

Diabetes mellitus (hyperglycaemia) is a metabolic disorder of the pancreas where there is a high glucose level in the bloodstream. Five types of diabetes are, type 1 diabetes mellitus (T1DM), type 2 diabetes mellitus (T2DM), latent autoimmune diabetes in adults [1] and gestational diabetes.

Prediabetes is the initial stage of a diabetic patient. If not treated well, prediabetes can be developed into T1DM or T2DM. but if treated well, risk of diabetes can be reduced or even prevented. T1DM is the state where there is a lack of production of insulin by the pancreas. Insulin is the enzyme that controls the glucose level in the blood by binding itself to the target organs' receptor cells to receive glucose. Glucose is the metabolite form of carbohydrates which reacts with oxygen to provide energy to organs. T1DM occurs when the immune system attacks the β cells of the pancreas by mistake that results in a lack of production of insulin [2]. The lack of insulin results in a lack of insulin binding to the receptor cells. This is insufficient to take the needed glucose into the organs which leads to a high glucose level in the bloodstream [2]. T2DM, on the other hand, is the disorder in receptor binding or disorders in the signalling of target cells, that makes the cells non-responsive to insulin [2]. T1DM is mostly diagnosed in younger people while adults are most prone to be diagnosed with T2DM.

Latent autoimmune diabetes (LADA) is diagnosed in adults mostly, usually after the age of 30 [1]. These patients show symptoms of both T1DM and T2DM. Some experts believe that LADA is the slowly developing kind of T1DM because the patients have antibodies that are against the β cells in their pancreas [1]. Gestational diabetes is the form of diabetes that is present in pregnant women. This type of diabetes sometimes fade away after giving birth, but in unfortunate cases develops into other forms of diabetes that last in the patient.

B. Manual Method of Diagnosing diabetes

Diabetes mellitus, regardless of type-1 or type-2, is diagnosed in clinics after carrying out blood tests to measure the glucose level in blood. Fasting blood sugar test is the most common way in clinical procedures while random blood sugar tests can even be done by the patients themselves by the use of glucometers (shortened form for glucose meter). For a fasting blood sugar test, a reading level from 100 to 125mg/dL is considered prediabetes while a reading more than 126mg/dL is considered diabetes [3]. For a random blood sugar test, a blood sugar level of 200mg/dL or higher is considered diabetes regardless of the time the patient ate last [3]. On the other hand, oral glucose tolerance tests are conducted mostly on pregnant women to test gestational diabetes.

However, to diagnose T2DM specifically, a special test called the glycated haemoglobin (HbA1C or simply A1C) is conducted and a result between 5.7-6.4% is considered prediabetes, while an A1C level of 6.5% or higher is considered diabetes [3].

III. Significance of Predicting Diabetes

Most prediabetes patients end up getting diagnosed with T2DM, which is the most common type of diabetes among adults. Even though diabetes can be delayed or even prevented in patients if identified early [4], prediction of the disease is still at the research level. According to the National Diabetes Prevention Program of the United States, by changes in the diet and by from weight loss through exercising, prediabetes can be prevented as much as 58% and up to 71% if the patient is over the age of 60 [4].

In 2016, the Centers for Disease Control and Prevention (CDC) of the United States has estimated that 24.8 in every 100,000 died of diabetes-related causes [5]. Although CDC has not specified the effect it has on the life expectancy of the patients, when at a Canadian study dated 2012, shows that the effects of life expectancy at 55 years of age [5]. They have found that the disease causes an average reduction of 6 years of female and 5 years in male patients [5].

Therefore, it is without a doubt important to predict diabetes rather than trying to control it after diagnosing. This review paper reviews some of the remarkable researches conducted from 2011-2020 to evaluate the best ML technique in predicting T2DM.

IV. Methodology

The research articles were identified following a bibliometric analysis. Google Scholar and Science Direct were used as the databases to search for research work. The keywords 'machine learning', 'diabetes', 'type 2', and 'prediction' were used in Science Direct. The results were refined to the period of 2011-2020 which gave 515 search results. On the other hand, the keywords 'prediction', 'machine learning', 'diabetes', 'type 2', 'diabetes mellitus', 'forecasting' were used in Google Scholar and after restricting time, 598 search results were obtained. Other resources from Google search, PubMed, Publons were also considered.

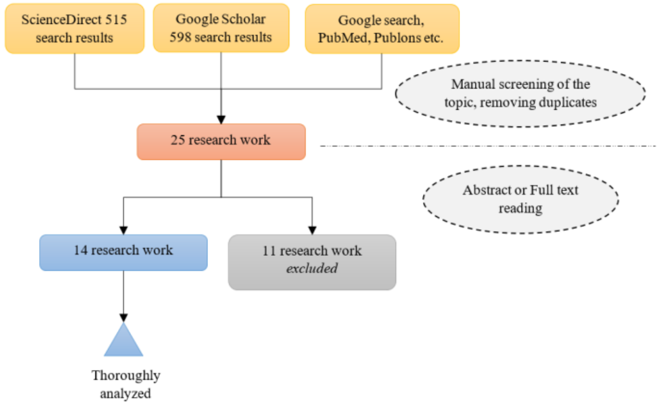

ScienceDirect 515 search results
Google Scholar 598 search results
Google search, PubMed, Publons etc.
Manual screening of the topic, removing duplicates
25 research work
Abstract or Full text reading
14 research work
11 research work excluded
Thoroughly analyzed

Fig. 1: Identification of Suitable Research work

Being inspired from the PRISMA Flow Diagrams, shortened for Preferred Reporting Items for Systematic Reviews and Meta-Analysis [6], the search results obtained from Google Scholar, Science Direct and other resources such as ordinary Google searching, Publons, PubMed were manually screened through for their topic and search description. After removing any duplicates, 25 such research work were identified. Thereafter, the research work was gone through by either an abstract read-through or a full-text read, from which 11 such research work were excluded and left with 14 remarkable pieces of research to analyse thoroughly. *Fig. 1* illustrates the above process of identifying suitable research work to be reviewed in this review paper.

V. Discussion

When taking an overview of the selected research works, the use of different techniques for the prediction purposes over the period 2011-2020 can be identified from the *Fig. 2* which is an overlay visualization of the usage of keywords present in them. The colour variation shows that the artificial intelligence (AI) technique fuzzy logic, has been interconnected with support vector machines (SVM) to do the prediction of diabetes to the latter of the period. It can also be seen that data mining techniques such as k-means clustering and associate rule mining have also been used with parallel to ML techniques in researches done 2018 and later. The thickness of the arcs connecting the nodes in the visualization diagram suggests that the relationship between predicting diabetes mellitus and neural networks (NN), random forests (RF), and decision trees (DT) is having a higher weight than other relationships indicated. On the other hand, when visualizing the density of the keywords (Refer *Fig. 3*), it makes it clear that most researches have used NN, RF, DT and SVMs to predict diabetes in their work. These diagrams (*Fig. 2* and *Fig. 3*) were generated from the scientific visualization tool VOSviewer [7] by Nees Jan van Eck and Ludo Waltman of Leiden University, the Netherlands.

When considering the 14 research work selected, the researchers have used many different techniques on different datasets. Among the datasets they have used, it can be seen that many have used the Pima Indian Diabetes (PID) dataset, which can be used as a common factor to compare and contrast the research results in 10 out of 14 research work cited in this review paper. According to Kaggle Inc. a large repository of published data and code [8], PID is originally from the National Institute of Diabetes and Digestive and Kidney Diseases [9] and was published in the University of California Irvine (UCI) Machine Learning repository. This dataset contains 768 records with 9 columns of data. All the patients in this dataset are females that are at least 21 years of age and Indian heritage [9].

Table I gives an overall view of the 14 research work analysed in this review paper with accordance to the dataset(s), techniques the researchers have used, the published year and the diabetes type they have tried predicting. It can be seen that Gaussian Naïve Bayes (NB), logistic regression (LR), k-nearest neighbour (kNN), classification and regression trees (CART) have also been used in the predictive models.

Subramani Mani et al. in their research have used electronic medical records (EMR) of Vanderbilt University Medical Centre [10] and have separated the data into 3 datasets as D365, D180 and D0 with cut-off days 365, 180 and 0 respectively. D180 means that the patient's data was available in the EMR 180 before him/her to be diagnosed with T2DM. Subramani Mani et al. have tested the datasets with the ML techniques such as Gaussian NB, LR, kNN, CART, RF and SVMs. The results show that RF has given the overall best performance among the other techniques.

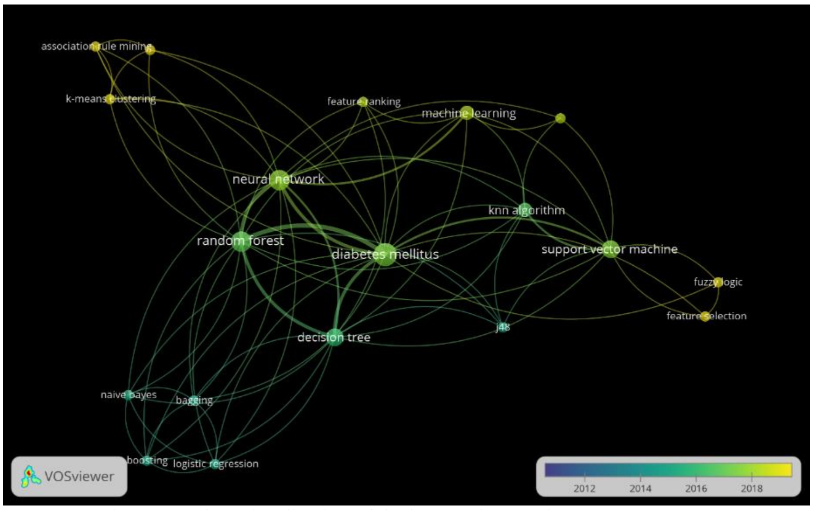


Fig. 2: Overlay Visualization of the keywords over the years 2011-2020

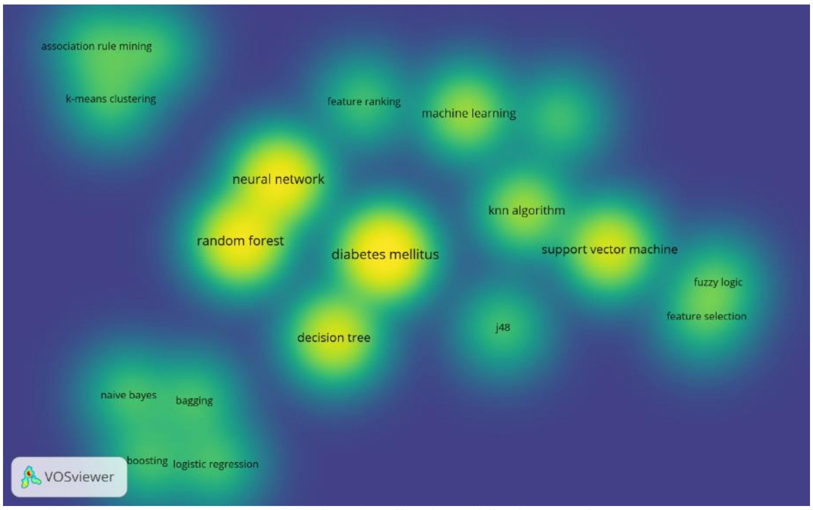


Fig. 3: Density visualization of the keywords

TABLE I. SUMMARIZATION OF RESEARCH WORK CITED

No	Research title	Year	Dataset(s) used	Techniques used	Diabetes evaluated
1	Type 2 Diabetes Risk Forecasting from EMR Data using Machine Learning [10]	2012	EMR of Vanderbilt University Medical Center	Gaussian NB, LR, kNN, CART, RF, and SVMs	T2DM
2	An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network [11]	2012	PID	Initially - data Mining technique C4.5, for improvising - NN	
3	Diabetes Prediction: A Deep Learning Approach [12]	2019	PID	Deep learning	Diabetes mellitus
4	Predicting Diabetes Mellitus With Machine Learning Techniques [13]	2018	hospital examination data in Luzhou, China, PID	RF, Decision Tree, NN, mRMR [13]	
5	An improved early detection method of type-2 diabetes mellitus using multiple classifier system [14]	2014	PID, Indonesian Patients (IP)	NN, LR, SVM, C4.5, NB, MFWC [14]	T2DM
6	Prediction of Diabetes Using Artificial Neural Network Approach [15]	2019	PID	NN	Diabetes mellitus
7	The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes [16]	2019	Heinz-Nixdorf-Recall Study (HNR) [16] data (4814 participants)	Deep Learning, DNN, SVM	T2DM
8	Diabetes Prediction Using Different Machine Learning Approaches [17]	2019	PID	DT, NN, NB, SVM	Diabetes Mellitus
9	Comparison of Classifiers for the Risk of Diabetes Prediction [18]	2015	26 Primary Care Units (PCU) in Sawanpracharak Regional Hospital during 2012 – 2013 (30,122 people)	13 models including bagging and boosting, DT, NN, LR, NB	

10	A model for early prediction of diabetes [19]	2019	PID	NN, RF, k-means	
11	Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine [20]	2019	PID	Fuzzy SVM	Diabetes Mellitus
12	Predictive modelling and analytics for diabetes using a machine learning approach [21]	2018	PID	Linear SVM, radial basis function (RBF), kernel SVM, kNN, NN and multifactor dimensionality reduction (MDR)	
13	Performance Analysis of Classifier Models to Predict Diabetes Mellitus [22]	2015	PID	DT, J48, kNN, RF, SVM in 2 separate situations (before data prepossessing and after)	
14	Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records [23]	2019	EHR dataset from the US, by Practice Fusion in 2012 for a data science competition [23]	Wide and deep learning framework	T2DM

It also mentions that results from RF are not much human-understandable to be used by medical personnel at clinical facilities, and that results derived from CART is comparatively human-comprehensible [10]. However, their objective to identify that the EMR is feasible enough to forecast the risk of diabetes before 365 and 180 days was achieved.

S. Priya et al. in their study have used the PID dataset and NN to improve the classification model made using data mining. Their model when only using the C4.5 classifier has given 91% accuracy rate, while when using NN to improve the data mining model, was able to give a success rate of 96%-98%. They have also used Kappa statistics, which serves as an agreement between two different qualitative observations, to compare the two models. The Kappa statistics show that the NN model has performed better with a success rate of 0.953 while C4.5 classifier has only given a rate of 0.8249. Safial et al. have proposed a deep neural network (DNN) on the PID dataset, which has given the accuracy of 98.35%, among the other ML techniques used. Quan Zou et al. have used both PID dataset and also hospital examination data in Luzhou, China. They have tested the data five times and the results they have presented in the paper are the average of the five experiments. Among the ML models they used, namely, RF, NN, and data mining technique J48, the RF has given the best performance without considering blood glucose. Nevertheless, they have found out that using all features for the prediction still has a better result and that only using blood glucose as a feature to predict diabetes is not a good decision, especially with NN as a classifier.

Jia Zhu et al. have used multiple classifier systems (MCS) with a dynamic weighted voting scheme called multiple factors weighted combination for classifiers' decision combination (MFWC), to improve the accuracy of the detection of T2DM on the PID as well as an Indonesian dataset [14]. They say that combined methods have performed better than using individual classifiers to predict T2DM. Also states that, MFWC has a slightly larger training time compared to other models and that it takes around 0.2seconds to identify a T2DM patient, still fast enough for a real application. They have validated their MFWC with other complex diseases as well and have observed that this combination remains at the top position on identifying those diseases too.

Suyash et al.'s and Sonar et al.'s researches have both used PID dataset. Suyash et al. suggest that their model will be able to perform even better than 92% accuracy from NN if exposed to many data in the future. Sebastian et al. have used SVM and DL to predict T2DM. It is integrated to an AI-based clinical decision support system (CDSS) which they call as a virtual doctor [16], to provide a solution to the exact problem that is brought forward in this review paper. From the data that the CDSS takes and the answers given by the patient the virtual doctor, can predict whether the patient is prone to diabetes or not. Their training data was collected during a Heinz-Nixdorf-Recall study (HNR) [16]. Among SVM and DNN, DNN has performed better in predicting T2DM, therefore the prediction model is implemented using DNN. The prediction probability is checked for, and if it lies within 30-70%, HbA1C test is prescribed by the virtual doctor [16]. The researchers have identified that the DNN trained without HbA1C test results outperforms the SVM in terms of AUC curve, which is the main reason for choosing DNN for the initial prediction. AUC is an abbreviation for the area under the curve [24]. It is used in classification analysis to determine which of the used models predicts the classes best. As Sebestian et al. describe, there can be cities where the population is not inspect-able by the medical personnel in that area. In such situations, it is always better if an AI-based virtual doctor can be used.

Nai-arun et al. have conducted a comparison between various classifiers for diabetes prediction. They have collected the data from 26 Primary Care Units (PCU) in Sawanpracharak Regional Hospital during 2012 – 2013 [18], containing 30122 records. They have used bagging and boosting algorithms on DT, NN, LR and NB. Boosting algorithm reduces the error of a weak classifier while bagging algorithm avoids overfitting and reduce the variance of the predicting model [18]. They have also created a web application by integrating the model, where they have used RF for the prediction (highest accuracy of 85.558%). Their web application lets the patients enter their details and view the prediction regarding them, which is another application where predicting diabetes is automated.

Mahboob et al. have also used the PID dataset to predict diabetes using NN, RF and k-means clustering (data mining). The best accuracy is 75.7% from NN. Lukmanto et al. have used the same dataset, and an advanced version of SVM to do the prediction. Their fuzzy SVM model strengthens and optimizes the traditional SVM classifier [20]. Fuzzy SVMs can be used for classification analysis and it is aimed at finding the most optimal hyperplane [20]. The results show an accuracy of 89.02% when using the fuzzy SVM model.

Kaur and Kumari [21] have proposed 5 predictive models using linear SVM, RBF, kernel SVM, kNN, NN and MDR. The best accuracy on PID was given by SVM, 89% with a precision of 88%. A study by Kandhasamy and Balamurali on PID dataset using DT, kNN, RF and SVM with bagging to enhance the prediction models, have compared the models in two different situations as before pre-processing and after.

They conclude that before pre-processing, the DT with J48 classifier has achieved the highest accuracy while after, kNN where k=1 and RF have provided 100% accuracy [22]. It is seen that, after removing noisy data, the models provide better prediction results.

Nguyen et al.'s research also uses electronic health records (EHR) that was used by Practice Fusion, the United States in 2012 for a data science competition. This dataset contains records of 9948 patients with 1904 diagnosed with T2DM [23]. Their algorithm predicts diabetes onset based on the wide and deep learning frameworks [23]. They have used Synthetic Minority Oversampling Technique (SMOTE) to improve the performance of the dataset, and their ensemble of the classifiers RF and NB, when improved with SMOTE, has provided an accuracy rate of 84.28% and a specificity rate of 96.85%.

When considering all the above-cited work, it can be seen that the accuracy results obtained differs with the use of different classifiers and the datasets. For the convenience of comparison, models using PID dataset can be considered. *Table II* is a summary of the results of the 14 research work.

TABLE II. BEST PERFORMED MODELS WITH THEIR ACCURACIES

No	Accuracy on PID	Accuracy on other datasets
1		Over 70% - RF on EMR of Vanderbilt University Medical Center [10]
2	96-98% - NN	
3	98.35% - DL	
4	77.21% mRMR [13]	80.84% - RF on hospital examination data in Luzhou, China
5	Almost 95% - MFWC [14]	97-98% MFWC [14] - IP dataset
6	92% - NN	
7		DNN – 0.703 AUC on HNR[16]
8	85% - DT	
9		85.558% - RF (highest ROC) 26 PCU in Sawanpracharak Regional Hospital [18]
10	75.7% - NN	
11	89.02% Fuzzy SVM	
12	89% SVM with 88% precision	
13	Before preprocessing, J48 73.82% accuracy. After, kNN (k=1) + RF 100% accuracy.	
14		Ensemble (RF+NB) 84.28% acc. - EHR dataset from the US released by Practice Fusion 2012 for a data science competition – 2020 [23]

When considering *Table I* and *II*, it is seen that 10 out of the 14 pieces of research have used the PID dataset for diabetes prediction. Only 2 out of 10 have used PID for T2DM prediction, while the other 8 research work predicts diabetes in general. Moreover, there are a total of 5 research work that has the hypothesis of predicting T2DM using ML techniques, and the other 9 are predicting diabetes mellitus. On the other hand, when focusing on PID dataset, it contains only female patient records 21 and above, which is doubtful whether this dataset is suitable enough to predict diabetes among patients in general, across genders, moreover to predict T2DM, a specialized version of diabetes. This can be the reason that there are only 2 researches among the 5 have used the PID dataset to predict T2DM, and the other 3 pieces of research use their custom datasets that have a higher number of records.

In general, the research work predicting T2DM specifically using ML techniques is scarce, which is why a concrete conclusion cannot be given by this review paper as to which method would be the best for T2DM prediction. From the observations above, RF and NN have the best overall performance in predicting diabetes, while there are 2 out of 14 research work obtained better results using SVM and 1 using DT. According to *Table II*, RF has successfully predicted when it comes to custom datasets, other than PID dataset. When analysing deeper on the researches that predict T2DM, NN and DNN have performed best. The difference between NN and deep learning lies with the depth of the model [25], [26]. The MFWC has also done a considerably good prediction. The MFWC has been more successful than the ensembles of RF and NB that has only obtained an accuracy of 84.28%.

The derivation obtained earlier, from analysing the data extracted from the full-text read of the research work indicates that RF, NN, SVM and DT have given the best results, which was previously illustrated in *Fig. 3*, created using the reference data of the research work, from VOSviewer [7].

RF and NN are better ML techniques for T2DM prediction when considering the above research work, each reserving 5 out of 14 research work with best results. In research no. 2 and research no.3, 2 hidden layers are used in the NN. Research no. 6 has used 1 hidden layer while research 7 (the virtual doctor) [16], has used from 1-3 hidden layers, and 1-20 neurons per each hidden layer. In research 10, a single hidden layer is used, while changing the no. of neurons. According to their study, the no. of neurons has an inversely proportional nature to the accuracy. They have indicated that when the no. of hidden neurons was increased from 50 to 100, the accuracy has worsened. Therefore, when considering the findings in research 7, we can assume that with the number of hidden layers, the accuracy rises, while reduces inversely with the no. of neurons in the hidden layers. But according to Bala Vignesh (PhD Computer Science from Yale University), in his answer to "Is accuracy proportional to the number of the hidden layers and the units of each layer in Neural Network?" [27] on Quora suggests that, even if we obtain a near-perfect accuracy for training sets, the test set error can shoot up when the number of hidden layers is increased [27]. He also mentions that the accuracy of the neural network depends on the performance of NN on its test data [27]. Therefore, what can be derived is that the accuracy of a neural network model does depend highly on the scenario and cannot be exactly pointed whether it would increase or decrease with relative to the number of hidden layers and hidden neurons used in the model.

VI. Conclusion and Suggestions for Further Work

Diabetes is a non-commutable disease with a high fatality rate. This review paper's objective was to review the existing research work that predicts T2DM with the use of ML techniques. Although this particular research area is scarce, among the work cited, it is seen that RF and NN have the best prediction accuracies. When comparing results, the accuracies tend to depend highly on the datasets used. Even though many research works have used the PID dataset since it is freely available, there is a doubt of its appropriateness to predict T2DM or diabetes among patients of both genders.

The recommendation from this review is to use a custom dataset for future T2DM predictions with both female and male patients. If a researcher is predicting diabetes in general, a dataset containing patient data from all age groups is preferred, while if the research is focused on predicting T2DM, it is preferred for the researchers to use a dataset with more adult patients, as T2DM is prone in adults.

Many research-work that have predicted diabetes and T2DM using ML ensembles, RF and NN have succeeded in their prediction with high accuracy. Therefore, it can be suggested that using RF and NN combined ML models will be able to give high prediction accuracies in future researches and integrating such models with CDSSs will be an added advantage for clinical facilities in predicting T2DM.

References

[1] "Prediabetes can be associated with both type 1 and type 2 diabetes," *MSU Extension*. [Online]. Available: https://www.canr.msu.edu/news/prediabetes_can_be_associated_with_both_type_1_and_type_2_diabetes. [Accessed: 27-Feb-2020].

[2] *Diabetes Type 1 and Type 2, Animation*.

[3] "Type 2 diabetes - Diagnosis and treatment - Mayo Clinic." [Online]. Available: https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/diagnosis-treatment/drc-20351199. [Accessed: 21-Jan-2020].

[4] CDC, "Prediabetes - Your Chance to Prevent Type 2 Diabetes," *Centers for Disease Control and Prevention*, 08-Jan-2020. [Online]. Available: http://bit.ly/2hMpYrt. [Accessed: 23-Feb-2020].

[5] "Type 2 diabetes and life expectancy: Risk factors and tips." [Online]. Available: https://www.medicalnewstoday.com/articles/317477. [Accessed: 24-Feb-2020].

[6] "PRISMA Flow Diagrams - YouTube." [Online]. Available: https://www.youtube.com/watch?v=H3H4KHWqnhI. [Accessed: 24-Feb-2020].

[7] "VOSviewer - Visualizing scientific landscapes," *VOSviewer*. [Online]. Available: https://www.vosviewer.com//. [Accessed: 28-Feb-2020].

[8] "Kaggle: Your Machine Learning and Data Science Community." [Online]. Available: https://www.kaggle.com/. [Accessed: 28-Feb-2020].

[9] "Pima Indians Diabetes Database." [Online]. Available: https://kaggle.com/uciml/pima-indians-diabetes-database. [Accessed: 28-Feb-2020].

[10] S. Mani, Y. Chen, T. Elasy, W. Clayton, and J. Denny, "Type 2 Diabetes Risk Forecasting from EMR Data using Machine Learning," *AMIA. Annu. Symp. Proc.*, vol. 2012, pp. 606–615, Nov. 2012.

[11] S. Priya and R. R. Rajalaxmi, "An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network," p. 4, 2012.

[12] S. I. Ayon and M. M. Islam, "Diabetes Prediction: A Deep Learning Approach," *Int. J. Inf. Eng. Electron. BusinessIJIEEB*, vol. 11, no. 2, p. 21.

[13] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front. Genet.*, vol. 9, 2018, doi: 10.3389/fgene.2018.00515.

[14] J. Zhu, Q. Xie, and K. Zheng, "An improved early detection method of type-2 diabetes mellitus using multiple classifier system," *Inf. Sci.*, vol. 292, pp. 1–14, Jan. 2015, doi: 10.1016/j.ins.2014.08.056.

[15] P. Rahimloo and A. Jafarian, "Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them," *Bull. Société R. Sci. Liège*, vol. 85, p. 17, 2016.

[16] S. Spänig, A. Emberger-Klein, J.-P. Sowa, A. Canbay, K. Menrad, and D. Heider, "The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes," *Artif. Intell. Med.*, vol. 100, p. 101706, Sep. 2019, doi: 10.1016/j.artmed.2019.101706.

[17] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 367–371, doi: 10.1109/ICCMC.2019.8819841.

[18] N. Nai-arun and R. Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," *Procedia Comput. Sci.*, vol. 69, pp. 132–142, Jan. 2015, doi: 10.1016/j.procs.2015.10.014.

[19] T. Mahboob Alam *et al.*, "A model for early prediction of diabetes," *Inform. Med. Unlocked*, vol. 16, p. 100204, Jan. 2019, doi: 10.1016/j.imu.2019.100204.

[20] R. B. Lukmanto, Suharjito, A. Nugroho, and H. Akbar, "Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine," *Procedia Comput. Sci.*, vol. 157, pp. 46–54, Jan. 2019, doi: 10.1016/j.procs.2019.08.140.

[21] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Appl. Comput. Inform.*, Dec. 2018, doi: 10.1016/j.aci.2018.12.004.

[22] J. P. Kandhasamy and S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus," *Procedia Comput. Sci.*, vol. 47, pp. 45–51, Jan. 2015, doi: 10.1016/j.procs.2015.03.182.

[23] B. P. Nguyen *et al.*, "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records," *Comput. Methods Programs Biomed.*, vol. 182, p. 105055, Dec. 2019, doi: 10.1016/j.cmpb.2019.105055.

[24] "classification - What does AUC stand for and what is it?," *Cross Validated*. [Online]. Available: https://stats.stackexchange.com/questions/132777/what-does-auc-stand-for-and-what-is-it. [Accessed: 28-Feb-2020].

[25] "About Artificial Intelligence, Neural Networks & Deep Learning," *Ayima*, 24-Jan-2017. [Online]. Available: https://www.ayima.com/blog/artificial-intelligence-neural-networks-deep-learning.html. [Accessed: 01-Mar-2020].

[26] M. A. Nielsen, "Neural Networks and Deep Learning," 2015.

[27] "(1) Is accuracy proportional to the number of the hidden layers and the units of the each layer in Neural Network? - Quora." [Online]. Available: https://www.quora.com/Is-accuracy-proportional-to-the-number-of-the-hidden-layers-and-the-units-of-the-each-layer-in-Neural-Network. [Accessed: 02-Mar-2020].

IoT Based Learning Enhanced Smart Parking Management System: A Smart City Initiative

LB L Senevirathne
Undergraduate
Department of Information Technology,
Faculty of Computing,
General Sir John Kotelawala Defence University.
Rathmalana, Sri Lanka.
34-1t-014@kdu.ac.lk

R P S Kathriarachchi
Senior Lecturer
Department of Information Technology,
Faculty of Computing,
General Sir John Kotelawala Defence University.
Rathmalana, Sri Lanka.
pathum@kdu.ac.lk

Abstract— *Smart City is a new concept that emerged from the last decade. It is a revolutionary way of thinking about how the new technological trends and Internet of Things (IoT) can be applied to a city. If a city can monitor and integrate, conditions of all of its critical infrastructures, including roads, bridges, tunnels, rail/subways, airports, seaports, communications, water, power, even major buildings, can better optimize its resources, plan its preventive maintenance activities, and monitor security aspects while maximizing services to its citizens.*

Parking has always been a predicament in large cities. A person cannot find a parking space inside of the busy urban area with ease, it takes a huge effort to find parking and It creates problems from traffic congestion to other problems like environmental damage to wastage of fuel and time.

The study focuses on the parking infrastructure of a city. Parking management systems are already available solutions. The available systems provide the solution to the parking payments and the other management facilities only. The future system will address the problem of finding a suitable parking space that is suitable to the precise user. The new system acts as a single network of parking spaces throughout the city so the user can find a parking space near to the location. To provide a better parking experience to the user, the study adopts machine learning techniques, the system will use user data, feedback, ratings, and behavior to provide a better parking experience. It is a single system to manage every parking infrastructure throughout the city using real-time data transmission and android mobile platform with the use of IoT sensor network and while managing it recommends suitable parking spaces using user feedback, rating, and behavior.

Keywords— **IoT, Smart City, Android, Smart Parking Management System, Real-Time Database, Sentiment Analysis.**

I. INTRODUCTION

This study aims to aid the users to find a parking spot near to their work using IOT devices and using the android platform, by using real-time data transmission techniques. The study adopts the same concept as Uber Taxi Service and applies that concept to the problem. The study proposes a simple yet effective methodology for vehicle parking space management[1] using real-time data transmission and geo-positioning systems. The proposed system aims to use sentiment analysis to analyze user feedback[2],[3] and aid in providing a better parking space according to each user, by using their feedback, rating, and behavior. The proposed system acts as a single network of parking spaces so the user can find a parking space near to his work. The main features are, finding a parking space close to their workplace without creating a lot of traffic congestion, without wasting time and fuel. The system identifies available and allocated parking spaces using IOT devices and uses that data to display nearby available parking spaces[4], notifies the users availability of the parking spaces and where he/she can park their vehicle. The system also utilizes user feedback and behavior to recommend suitable parking spaces. This approach does not stop at a single parking lot/structure, it creates a network connecting with other parking lots/structures available throughout the city and uses all the data as a huge pool to manage the parking infrastructure[5]. Retrieving real-time data, Processing the data, transmit the processed information throughout the parking space network, and transmit information back to users

II. LITERATURE REVIEW

Smart Cities are a new concept that emerged from the last decade. It is a revolutionary way of thinking about how the new technological trends and Internet of Things (IoT) can be applied to a city concept[6].

A city that monitors and integrates conditions of all of its critical infrastructures, including roads, bridges, tunnels, rail/subways, airports, seaports, communications, water, power, even major buildings, can better optimize its resources, plan its preventive maintenance activities, and monitor security aspects while maximizing services to its citizens. Emergency response management to both natural as well as man-made challenges to the system can be focused and rapid. With advanced monitoring systems and built-in smart sensors, data can be collected and evaluated in real-time, enhancing city management's decision-making[6].

To establish what a smart city is, it is an urban area that uses different types of electronic Internet of Things sensors to collect data and then use insights gained from that data to manage assets, resources and services efficiently[7].

Transportation of both the private and public sector is an integral part of any city, whether it is a smart city or a regular city. In a city, people use many types of vehicle for their day to day tasks and the number of vehicles keeps increasing in the huge margin with the increase of vehicles it allows few problems to rise throughout the city. With the increased number of vehicles, the main problem is increased vehicle traffic congestion[8]. The traffic can cause more problems in related infrastructures. When having a closer look into this problem reveals the main cause that it is not enough parking space and poor management of parking space[9].

In simple terms with the increased number of vehicles in the city, drivers having difficulty with finding a parking space, so they go through the city more than few times until they find an available parking space. This causes unnecessary traffic congestion, unnecessary consumption of fuel, environmental damage, and waste of time. As we can see in a city, one problem in a single infrastructure that supports the city can lead to a cascade of failures and problems throughout various infrastructures. It can easily cripple a city and bring it to its knees. also, by 2023, market spending for smart parking products and services is expected to grow at a Compound Annual Growth Rate (CAGR) of 14% and surpass $3.8B according to an IoT Analytics report[10]. The growth of market spending is good news because it will force people to try to find a solution to these traffic problems instead of taking no action. Parking management systems are already available systems. But does it address above mentioned problems, no because currently available solutions only address one part of the problem, the current systems look after only the charging the payment from the customer and the parking infrastructure capacity management only, it does not address the problem of finding a suitable parking space that is suitable to the specific user[11]. The proposed system acts as a single network of parking spaces so the user can find a parking space near to his work. To provide a better parking experience to the user, the study adopts machine learning techniques, the system will use user data, feedback, ratings, and behavior to provide a better parking experience. The study adapts the same concept as Uber Taxi Service and applies that concept into the problem[5].

To tackle this problem the study proposes an IOT based Learning Smart Parking Management System. It is a single system to manage every parking infrastructure throughout the city using real-time data transmission and android mobile platform with the use of IoT sensor network and while managing it recommends suitable parking spaces using user feedback, rating, and behavior. It is a System with 3 major Components,

1. IoT Module
2. Mobile Application
3. Admin Console/ Server.

IoT Device comprises of Arduino Mega Board That has WIFI Capabilities (ESP8266) and Ultra Sonic sensors, the system utilizes ultrasonic sensors to detect if a vehicle is parked in the parking space or not [Appendix A, *Diagram IV]*. To achieve a perfect detection system uses 3 ultrasonic, use of 3 sensors in a zig-zag pattern eliminates the fake detection alarms. all 3 sensors need to register an object within a valid distance range to identify the object as a vehicle. Then the detected Data is Transmitted to Firebase Realtime Database via the ESP8266 WIFI module.

IoT module transmits data directly to the real-time database. The System Utilizes 2 Types of NoSQL Databases, Firebase Realtime Database, and Firebase FireStore, System uses FireStore Database to Store the User Related Data and Feedback Analysis Data.

For All the Location Related Querying and Realtime Location, the System Utilizes GeoFire. GeoFire is an open-source JavaScript library that allows you to store and query a set of items based on their geographic location. This enables the Geographic location transfer between User, Server, and Real-Time Database.

Mobile Application provides an interface to the User so that the user can interact with the system. Using the mobile application user can see how many parking spaces available near the user, User can reserve a parking space, user can pay for the parking space he used, Early reservation system, User can check where he parked his/her vehicle, Provide feedback and rating, Getting recommended parking space information, If the near parking spaces are full, the application will suggest a parking space closest to the user.

Recommendation systems are widely available systems but most of them are single-dimensional only and do not use improved methods like machine learning, Natural Language Processing, and sentiment analysis. Feedback Analysis is done Using Google Natural Language API[12]. The API will analyze User Feedback and provide a score then the score will be added to the relevant parking space. Using this score and users' attributes, work location and current location, the system provides parking space recommendations.

Server controls the data flow and the user activity via the Administrative Web Application, it can control and Inspect every aspect of the user, Add, Remove, Block, Update user, Manage user behaviors, User Feedback Analysis, and recommendation system.

Currently, available solutions do not involve city-wide management of parking spaces. The proposed system can alleviate traffic congestion happening throughout the city because of poor parking space management and the problems occurring with it.

III. METHODOLOGY

The methodology as follows, Quantitative data is needed for the study because the study addresses a problem related to city infrastructure. Data gathering was done using a carefully designed questionnaire and requesting information from relevant authorities. The research studied how IoT technology being used in similar situations and other projects related to smart city concepts. So, the previous studies have revealed that even though IOT uses is heavy in these kinds of projects but, a city-wide application is rarely used.

To achieve city-wide parking management and give specific characteristics like learning and smartness it is necessary to achieve 2 objectives.

The first objective is a Mobile application to Interact with the user and letting the user interact with the system and let the user search and reserve parking spaces. For to achieve this study utilizes Android Studio to develop the Android app and uses Firebase SDK, Google Map SDK, and Google Natural Language API. Firebase is used to Store and Retrieve data in real-time. Google Map SKD is used to display real-time geolocation data, give coordinates to the user, and retrieve location updates from the user. Google Natural Language API is used for user feedback sentiment analysis then adds the score to the used parking space. When next time a user search for parking, his/her parking choices will be affected positively or negatively by this rating depending on the rating score.

The second objective is an IoT Module (Arduino Module). The study utilizes this IoT module to detect parking space state and parking space queue management. IoT Module contains the Arduino Mega Board, ESP8266 WIFI Module, LCD Display Module, and 3 Ultrasonic Sensors. Using Ultra Sonic sensors to detect whether a vehicle is parked in the parking space. The IoT Module achieves this by checking whether an object is near a certain distance (Appendix A Diagram III), by using not 1 but 3 sensors. it gives much-needed accuracy to the module and it further enhances the detection of various vehicles regardless of brand or type. As for queue management, the IoT module is programmed to update the Firebase real-time Database about the parking space state, by doing so the mobile app does not need to send any request(ping) to update parking space states. Both the Mobile Application and IoT Module act Independently but seemingly helps each other to achieve the goal of Parking Management without Bottlenecks or any other drawbacks.

IV. LIMITATIONS

The Study Identified 2 Major Limitations to implement,

I. City-wide user load balancing and reservation.
II. User Data Collection Related Privacy Issues.

Since the proposed system has many modules and a large user base, it is imperative to Load balance the users and the information flow. The system should be reliable and fluid. To achieve this, proper load balancing is a must. Proper load balancing will result in a system that is Scalable, Flexible, Efficient, and reduced down times[13]

as a solution for this challenge, it is proposed to use Firebase Realtime Database and let the IoT Module directly handle the parking queue processing. For example, if a parking space is available IoT module updates the Realtime Database, the system nor the app sends any requests to the IoT module to Update the Parking queue only thing the App and The System does is read the Realtime database and display the data and related tasks.

In recent times, user data and privacy has become a sensitive topic. privacy is the right of an individual to be free from uninvited surveillance. To safely exist in one's space and freely express one's opinions behind closed doors is critical to living in a democratic society. It is very much important that the system collects only the needed information with the user's consent. Also, to enforce user privacy the system should implement security measures when processing and transferring data[14]

as a solution for user data privacy, the system does not keep any user sensitive data like user real-time location system only use the real-time location but it does not store it in the database once the user closes the App location will be deleted from the system.

V. IOT ARCHITECTURE AND PLATFORM

The IoT Module has Comprised of,

I. Arduino Mega Board.
II. ESP8266 Wi-Fi Module
III. 3x Ultrasonic Sensors
IV. LCD Display Module

The Module uses all 3 Ultrasonic sensors simultaneously to detect whether the parking space is Full or Not. After detecting the state by using the ESP8266 Wi-Fi Module it updates the parking queue in Firebase Realtime Database. (Appendix A Diagrams IV, V, and VI).

VI. TESTING AND IMPLEMENTATION

Testing Was Completed Successfully and the system performs without any errors. The summary of the test cases is below.

Test Case	Description	Result
Parking Reservation Component	This component includes viewing of currently available parking spaces in real-time closer User's vicinity and Reserving Them.	Mobile Application displays available parking spaces and selects the best suitable parking space according to user input.
Parking Space State Detection (IoT Module)	This Module Includes Parking Space Detection and Update Parking Queue in Firebase Realtime Database.	IoT Device automatically detects when a vehicle is parked in space and updates the real-time database.

VII. SYSTEM EVALUATION

The System evaluation ensures that the built system can satisfy all end-user problems and can perform its functions according to objectives. First Objective is Parking Space State detection and Parking Spaces Queue Management, the objective has been successfully achieved by implementing IoT Module (Appendix A Diagrams V and VI). the second objective is to reserve the best parking space and user feedback sentiment analysis. The Objective has been successfully implemented into the Mobile Application.

VIII. CONCLUSION AND FUTURE ENHANCEMENTS

The current system has proven to be a solid solution and a foundation for the parking predicament. but these improvements are small compared to the further enhancements and improvements to the system. The current system has much more potential untapped. Further enhancements like,

- Predicting the traffic in nearby urban areas.
- Predicting parking availability
- Improve Vehicle Detection with Image Processing Techniques.

With these kinds of improvements, the system can become much more powerful and effective.

According to the literature review and gathered data, the system has the potential to improve city efficiency to its maximum. This is a completely new approach to the parking management concept. Using IoT Based Learning Enhanced Smart Parking Management System is an optimal solution to alleviate traffic-related problems in a City. Furthermore, having a clear knowledge about Driver patterns the system can identify better parking locations and personalize the user experience. By automating the complete process from the level of identifying parking space state, to the level of reserving parking space, the system optimizes the city experience and it will be a major cause to reduce Accidents, Traffic Congestion, Pointless fuel consumption, and time waste.

REFERENCES

[1] "Smart parking systems: A survey," *ResearchGate*. https://www.researchgate.net/publication/312241375_Smart_parking_systems_A_survey (accessed Feb. 17, 2020).

[2] M. P. Anto, M. Antony, K. M. Muhsina, N. Johny, V. James, and A. Wilson, "Product rating using sentiment analysis," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Mar. 2016, pp. 3458–3462, doi: 10.1109/ICEEOT.2016.7755346.

[3] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, no. 1, p. 5, Jun. 2015, doi: 10.1186/s40537-015-0015-2.

[4] J. Cynthia, C. Priya, and P. A. Gopinath, "IoT based smart parking management system," *International Journal of Recent Technology and Engineering*, vol. 7, pp. 374–379, Jan. 2019.

[5] J. Cramer and A. B. Krueger, "Disruptive Change in the Taxi Business: The Case of Uber," *American Economic Review*, vol. 106, no. 5, pp. 177–182, May 2016, doi: 10.1257/aer.p20161002.

[6] R. Hall, B. Bowerman, J. Braverman, J. Taylor, H. Todosow, and U. Wimmersperg, "The vision of a smart city," *2nd Int. Life* ., Jan. 2000.

[7] K. Su, J. Li, and H. Fu, "Smart city and the applications," in *2011 International Conference on Electronics, Communications and Control (ICECC)*, Sep. 2011, pp. 1028–1031, doi: 10.1109/ICECC.2011.6066743.

[8] Z. (Sean) Qian, F. (Evan) Xiao, and H. M. Zhang, "Managing morning commute traffic with parking," *Transportation Research Part B: Methodological*, vol. 46, no. 7, pp. 894–916, Aug. 2012, doi: 10.1016/j.trb.2012.01.011.

[9] A. Aoun, M. Abou-Zeid, I. Kaysi, and C. Myntti, "Reducing parking demand and traffic congestion at the American University of Beirut," *Transport Policy*, vol. 25, pp. 52–60, Jan. 2013, doi: 10.1016/j.tranpol.2012.11.007.

[10] T. Lin, H. Rivano, and F. Le Mouël, "A Survey of Smart Parking Solutions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3229–3253, Dec. 2017, doi: 10.1109/TITS.2017.2685143.

[11] R. Wang and Q. Yuan, "Parking practices and policies under rapid motorization: The case of China," *Transport Policy*, vol. 30, pp. 109–116, Nov. 2013, doi: 10.1016/j.tranpol.2013.08.006.

[12] "Cloud Natural Language | Google Cloud." https://cloud.google.com/natural-language (accessed Oct. 01, 2020).

[13] X. Sun and N. Ansari, "Traffic Load Balancing Among Brokers at the IoT Application Layer," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 489–502, Mar. 2018, doi: 10.1109/TNSM.2017.2787859.

[14] M. U. Iqbal and S. Lim, "Privacy Implications of Automated GPS Tracking and Profiling," *IEEE Technology and Society Magazine*, vol. 29, no. 2, pp. 39–46, Summer 2010, doi: 10.1109/MTS.2010.937031.

[15] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, "Collaborative filtering and deep learning based recommendation system for cold start items," *Expert Systems with Applications*, vol. 69, pp. 29–39, Mar. 2017, doi: 10.1016/j.eswa.2016.09.040.

[16] S. Stumpf *et al.*, "Toward harnessing user feedback for machine learning," in *Proceedings of the 12th international conference on Intelligent user interfaces*, Honolulu, Hawaii, USA, 2007, pp. 82–91, doi: 10.1145/1216295.1216316.

[17] "(PDF) Collaborative Filtering and Deep Learning Based Recommendation System For Cold Start Items," *ResearchGate.* [Online]. Available: https://www.researchgate.net/publication/309224171_Collaborative_Filtering_and_Deep_Learning_Based_Recommendation_System_For_Cold_Start_Items. [Accessed: 21-Feb-2020].

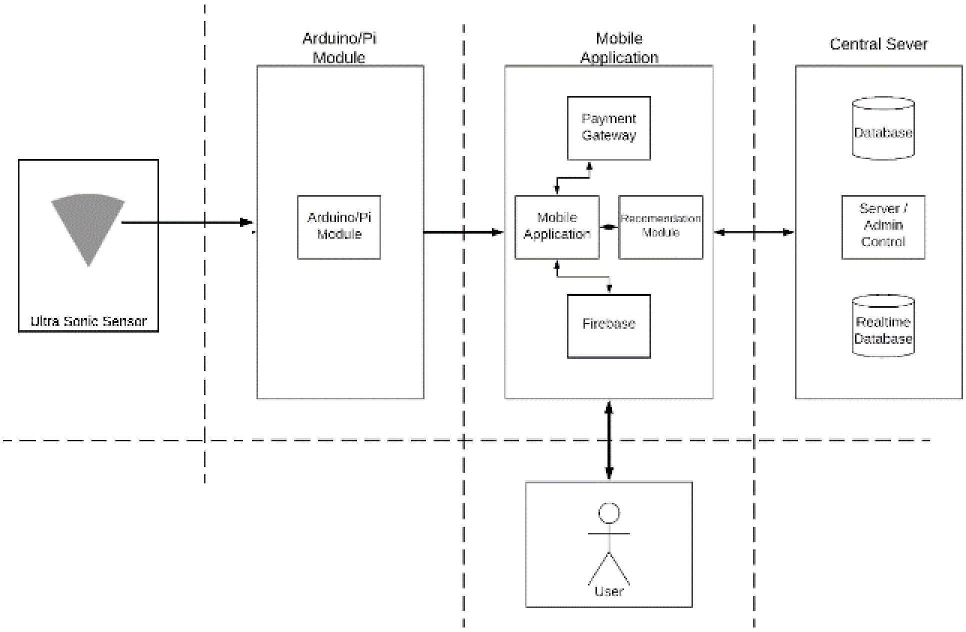
[18] "The Value of User Feedback," *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/221397271_The_Value_of_User_Feedback. [Accessed: 21-Feb-2020].

[19] D. Teodorović and P. Lučić, "Intelligent parking systems," *European Journal of Operational Research*, vol. 175, no. 3, pp. 1666–1681, Dec. 2006, doi: 10.1016/j.ejor.2005.02.033.

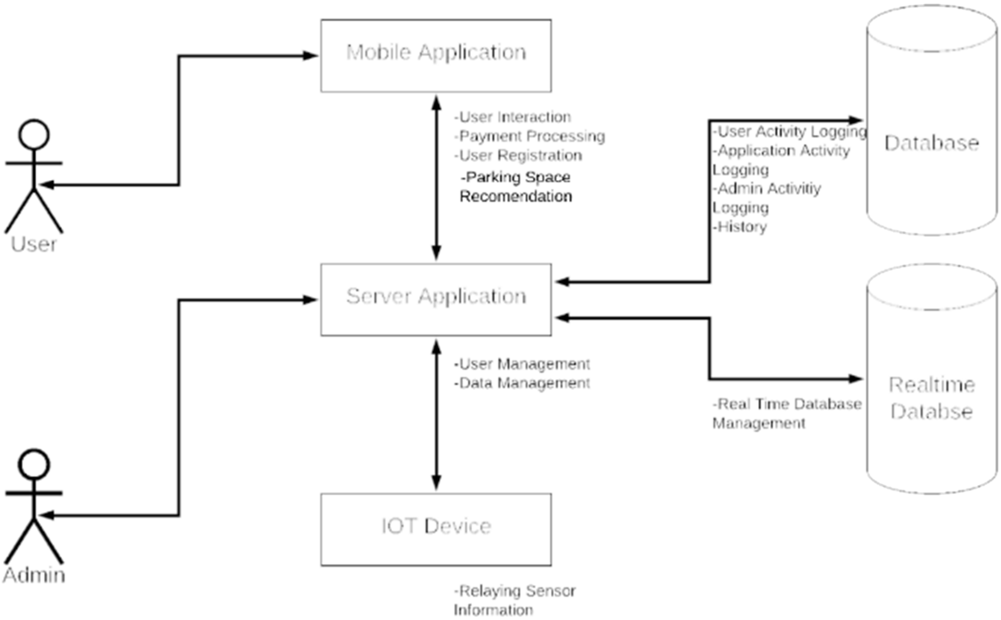
[20] S. I. Mohd Mustari *et al.*, "IoT Based Smart Parking System," *2nd International Conference on Advance & Scientific Innovation*.

[21] L. Nils and J. J. E. Brown, "Vehicular parking systems," US3166732A, 19-Jan-1965.

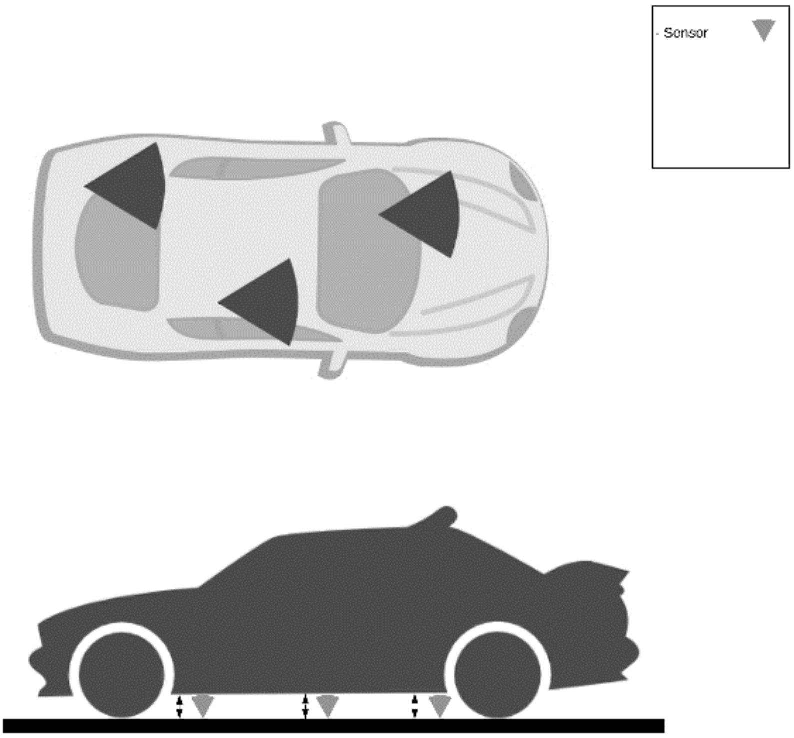
Appendix A



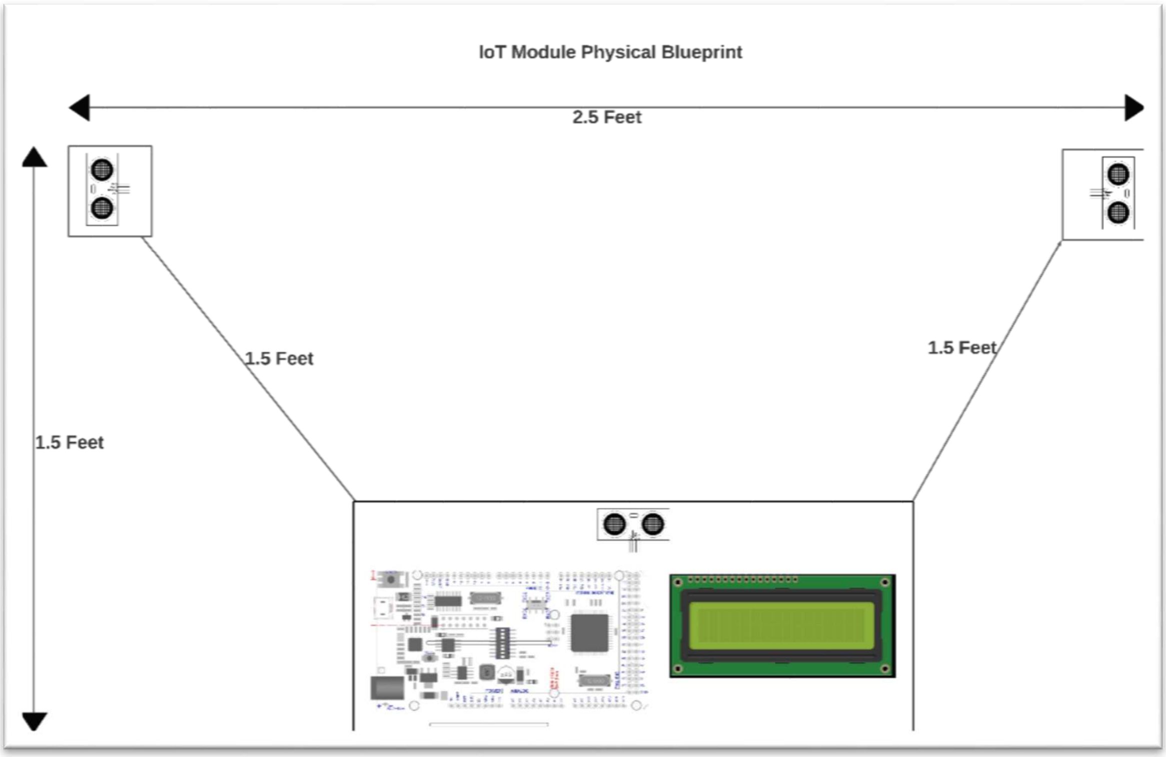
Basic System Architecture
Diagram I (Source: Author created)



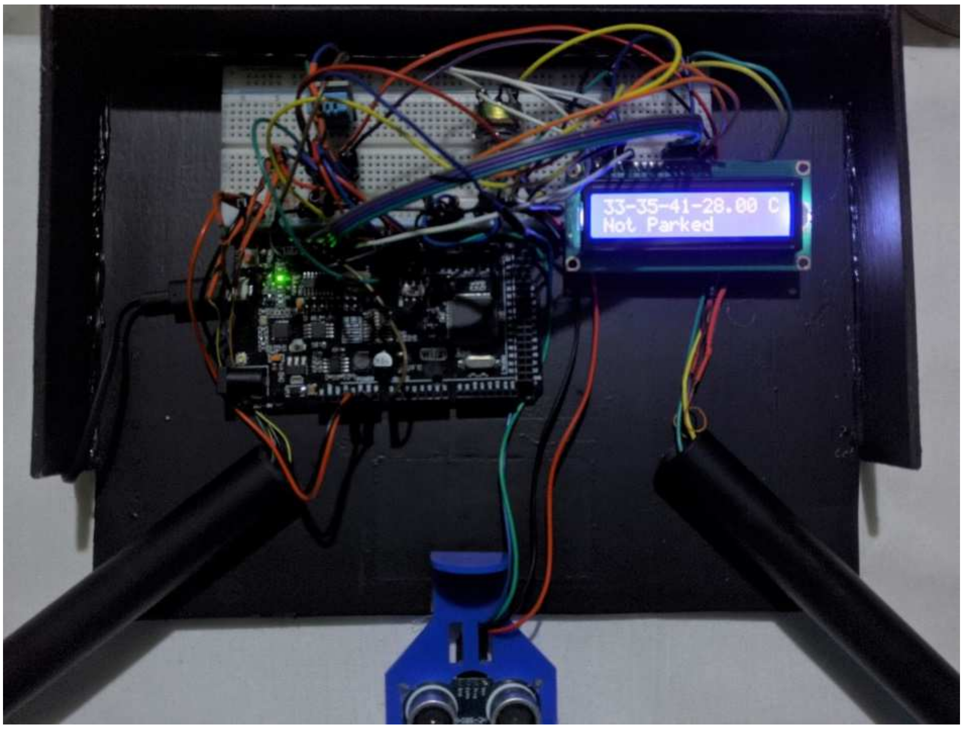
Parking Management System (Data Flow)
Diagram II (Source: Author created)



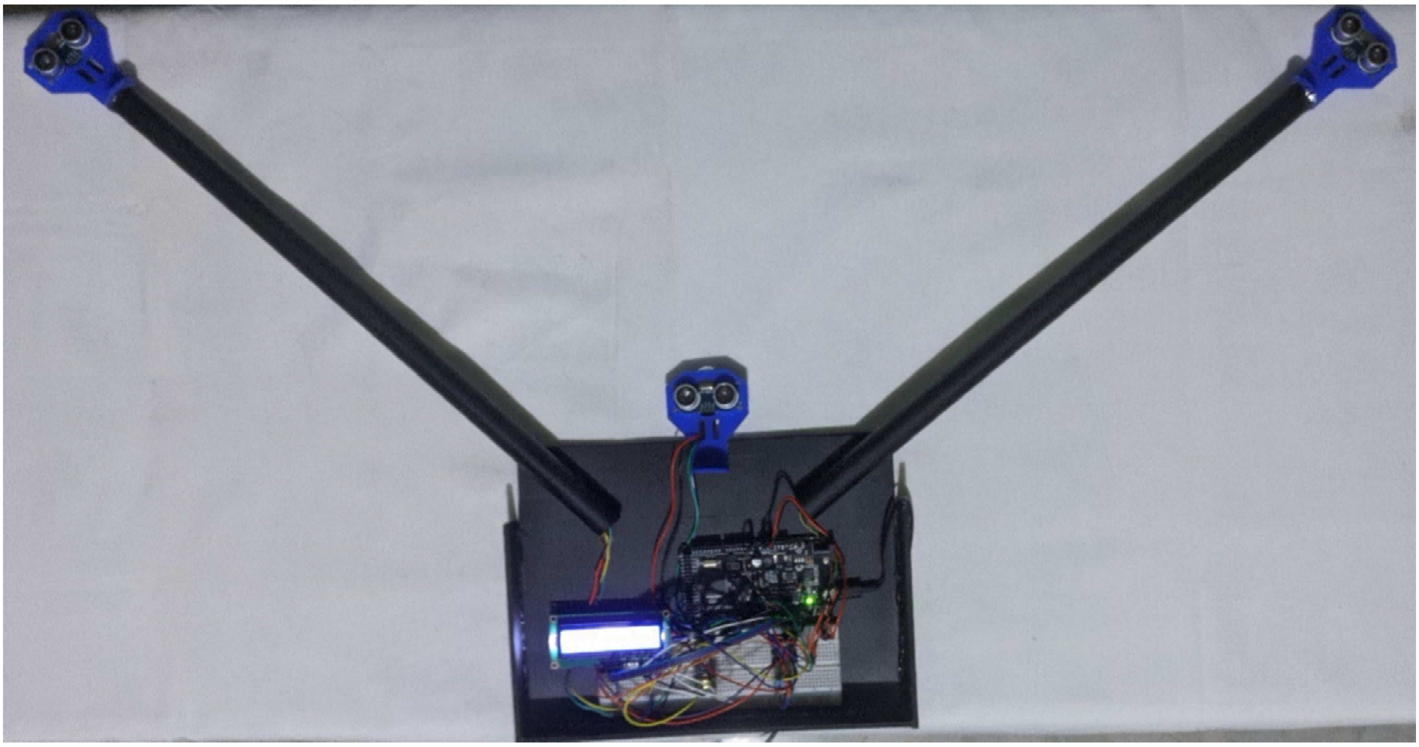
How the IoT Module Works
Diagram III (Source: Author created



Arduino Physical Blueprint
Diagram IV (Source: Author created)



Arduino Device
Diagram V (Source: Author created)



Arduino Device
Diagram VI (Source: Author created)

Personality Prediction Review on Text Modality Dataset

Tharsika Thurairasa
InformationTechnology
SLIIT
Colombo, SriLanka
tharsikasliit@gmail.com

Abstract—**Automatic personality detection of an individual's character qualities has numerous significant useful applications. Personality assessment is used to evaluate the individual on different aspects. With regards to assumption examination, for instance, the items and administrations prescribed to an individual ought to be those that have been emphatically assessed by different persons with a comparable personality type.**

Social media usages have been on an ever-increasing exponential rise. These days sage of social media sites, such as Twitter and Facebook, for social interaction has been becoming a popular trend. Personality trait can be predicted using text modality or visual modality or audio modality or by combining visual and audio then it's called bimodal or combining text, visual and audio then it's called trimodal. Many approaches have been proposed on text modality. This paper gives the reader an overview of the advancement of personality detection used on text modality datasets from text via social media.

Keywords—personality prediction, social media, machine learning, deep learning, big five model

I. Introduction

In these days computerized personality detection is an important field. There have not been numerous extensive writing reviews done in character location and our paper is the first which gives the reader a bird's-eye perspective on the ongoing patterns and advancements in the field. Texts often reflect different aspects of the author's personality [4]. There is no ongoing work which gives the reader a general point of view of the advances in machine learning and deep learning based mechanized personality detection specially using text dataset. As indicated by the Merriam-Webster dictionary the social media is characterized as types of electronic communication through which users make online networks to share thoughts, information, individual messages, and other substance, (for example, recordings, videos). Online media is an inescapable aspect of the Internet, as insights show that individuals burn through 1 in like clockwork of their Internet use via social media. A perception with respect to Facebook use detailed that clients sign into their Facebook accounts from 2 to 5 times each day, with a normal of 5 to 15 minutes for every meeting [14].

II. Personality Measures

A. Personality Measures Theories

Personality Theories fall into fundamental 4 classifications as Psychoanalytic Theory (additionally alluded as psychodynamic), Trait hypothesis, Humanistic hypothesis, and Social comprehension hypothesis. In the exploration reason, the quality hypothesis is broadly utilized [1].

Psychoanalytic Theory: - As indicated by Freud, the per sodality is comprised of three segments as id, inner self and superego. id alludes to the drive vitality that is liable for the human needs like sustenance, thankfulness and urges like scorn, desires and so on. The superego or soul, represent profound quality and Social standards, speak to what an individual need to be. Conscience the third part chips away at the guideline of reality that intervenes between the requests of the primary segment id and the second segment superego and afterward picks the most reasonable answer as long as possible [1].

Trait Theory: - Present-day attribute hypothesis attempts to show character by setting of various grouping measurements (generally following a lexical methodology) and developing a poll to measure them. Analysts have utilized different plans for personality demonstrating, for example, 16PF, EPQ-R and three quality character model PEN where there are super-factors Psychoticism, Extraversion, and Neuroticism (PEN) at the head of the pecking order. The Myers–Briggs Type Indicator (MBTI) is one of the most generally managed individuality tests, given a huge number of times each year to representatives in a great many organizations. The MBTI character measure arranges individuals into two classifications in every four dimensions: self-preoccupation versus extraversion, detecting as opposed to intuiting, thinking as opposed to feeling, and judging as opposed to seeing. The most mainstream measure utilized in the writing on mechanized character discovery is by a wide margin the Big-Five-character qualities, which are the accompanying paired (yes/no) values [2].

Humanistic Theory: - Maslow accepted that character depends on close to home decisions not on nature or sustain. He recommended that individuals have and are inspired to assist them with pursuing their necessities or want that was spoken to in and the last level: self-realization that is creating and developing to arrive at genuine potential [1].

Social Cognition Theory: - The social insight hypothesis see character in the type of social communications. An individual's conduct is influenced by nature in which he remains [1].

The Trait hypothesis is most broadly utilized in examining the character in the field of Psychology. Dissimilar to different speculations this depends on finding the contrasts between the characters of people. The blend of different characteristics frames a character that is consistently one of a kind for each person. Most examinations on character forecast have zeroed in on the Big Five or MBTI personality models, which are the two most utilized personality models.

Big Five: The present Researchers accept that there are 5-character qualities. Large Five recommends that the attributes can be ordered on 5 unique classes. Careful marks for these 5 characteristics are as yet hard to concur for some of them. The famous abbreviation is OCEAN for traits.

Extraversion (EXT): - Individuals with high Extroversion quality have high certainty, Positive vitality, and positive feelings, friendly and inclination to communicate more with others. They are garrulous in nature. It repudiates saved conduct. Components identified with this quality are vitality, garrulity, carefree, well disposed, helping and so forth. These individuals like themselves just as about their general surroundings. Individuals with low extroversion are saved, calm.

Is the individual friendly, loquacious, and vivacious versus held and lone?

Neuroticism (NEU): - It is repudiated sure or secure nature. individuals with high neuroticism touchy or apprehensive. This characteristic is described by trouble, grouchiness, and passionate insecurity. They experience negative feelings and feelings effectively, similar to outrage, uneasiness, sorrow, pessimism and so forth. It alludes to the inclination to encounter negative enthusiastic states and see oneself and the world around contrarily. Variables like volatile, on edge on edge and so on are some related attributes.

Is the individual delicate and anxious versus secure and certain?

Agreeableness (AGR): - This is the propensity to be agreeable with others as opposed to being dubious. They are benevolent and enjoyed by their associates just as individuals encompassing them. They don't care to battle or contend as opposed; they are harmony creators. Humble, amenability, supportive, understanding, kind, touchy and so forth are the qualities that go under the umbrella of suitability.

Is the individual reliable, clear, liberal, and unassuming versus temperamental, muddled, small and pretentious?

Conscientiousness (CON): - It alludes to the fitness of being consistent, self-taught, capable, zeroing in on accomplishing objectives, and organizes designs rather than unconstrained conduct. It contrasts indiscreet conduct. It means how cautious, careful, genuine an individual is. It is an approach to control driving forces and act in a manner that is adequate socially by everybody around. These individuals are extraordinary at arranging and sorting out viably. These incorporate components as arranging, capable, difficult work, assurance, eager, control and so forth. They are acceptable in administration characteristics.

Is the individual proficient and sorted out versus messy and care-less?

Openness (OPN): - It mirrors the scholarly degree of an individual. How inquisitive, innovative novel an individual is. It likewise reflects how creative or autonomous an individual is. Transparency is identified with individuals' enthusiasm to attempt to new things, the capacity to be defenseless, and the ability to consider new ideas. Basic characteristics identified with transparency are: Imagination, different interests, Originality, Daring, Cleverness, Intellect, Creativity, Curiosity and so forth.

Is the individual innovative and inquisitive versus one-sided and careful?

III. Approaches Used

There is a critical developing enthusiasm for computerized personality prediction utilizing web-based media among specialists in both the Natural Language Processing and Social Science fields. Up until now, the utilization of customary personality tests has generally been restricted to clinical brain research, advising and human asset the board. Be that as it may, computerized character forecast from web-based media has a more extensive application, for example, online media advertising or dating applications and sites [3].

There have been numerous techniques utilized for personality detection as demonstrated as follows.

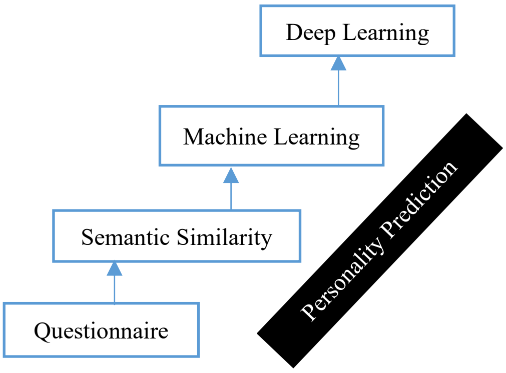


Fig. 1. Techniques Used

Questionnaire: - The most punctual type of approach utilized for character expectation was in the type of inquiries. Clients were posed a few inquiries that had various options, from which the client needed to choose one. These Questions were distinctive for various character qualities. In view of the client determination of alternative, it was appraised on some scale. Along these lines, help to foresee the last score for every characteristic by adding the absolute scores identified with that question.

Semantic Similarity: - In this for the characteristics there are pre-characterized jargon or word reference words. The client's words present in the posts are checked for the semantic comparability, for example, comparative implications have the same score. The separation is discovered and hence the attribute was anticipated.

Machine Learning: - Classical approaches can't deal with the vast amount of data. This is one of the upsides of Machine learning calculations. Machine learning algorithms utilize computational techniques to "learn" data legitimately from information without depending on a foreordained condition as a model. The ML algorithms adaptively increase their presentation as the quantity of tests accessible for learning increments. ML can likewise discover the examples from the information that probably won't be the obvious by the people.

Deep Learning: - In deep learning, a computer model figures out how to perform grouping assignments legitimately from pictures, text, or sound. deep learning models can be accomplished best in class exactness, now and again surpassing human-level execution. Deep learning models are prepared by utilizing an enormous arrangement of marked information and neural network architectures that contain numerous layers. Deep learning can be utilized to foresee the character attributes with more precision. It measures a similar path as human minds do. The component extraction measure is and there is no over-burden [1].

IV. Applications

There are different various modern utilizations of mechanized personality acknowledgment systems in the current day situation. We derive that a tremendous market will open up soon and if models can quantify character precisely and reliably, there will be an immense interest for robotized personality acknowledgment programming in the business. As a research advances in this field, before long better personality expectation models with a lot higher exactness and unwavering quality will be found. Counterfeit personality can be coordinated with practically all human PC cooperation going ahead.

Any computational gadgets can be furnished with a type of personality which empowers it to respond distinctively to various individuals and circumstances. For instance, a telephone can have various modes with various configurable characters. This will clear the path for additionally fascinating and customized connections. Another chance is to utilize character attributes as one of the contributions for accomplishing higher exactness in different errands, for example, mockery identification, lie recognition or word extremity disambiguation frameworks [2].

Employment screening: - In human assets the executives, character attributes influence one's reasonableness for specific positions. For instance, perhaps an organization needs to enroll somebody who will persuade and lead a specific group. They can limit their screening by eliminating up-and-comers who are exceptionally anxious and touchy, i.e., those having high estimations of neuroticism characteristic. Examines the activity up-and-comer screening issue from an interdisciplinary perspective of analysts and AI researchers.

Criminology: - If the police know about the character qualities of the individuals who were available at the wrongdoing scene, it might help in lessening the hover of suspects. Character identification additionally helps in the field of computerized duplicity discovery and can help in building lie indicators with higher exactness.

Specific medical care and guiding: - As of 2016, almost 33% of Americans have looked for proficient directing for psychological well-being connected issues. This is one more region where enormous useful utilization of character quality expectations exists. As per a person's character, suitable computerized directing might be given, or a psychiatrist may make utilize this data to offer better advising guidance.

Word extremity location: - Personality recognition can be misused for word extremity dis-ambiguation in estimation vocabularies, as a similar idea can pass on various significance to various kinds of individuals. Additionally, joining client character qualities and profile for dis-ambiguation among snide and non-mocking substance gives improved exactness results.

Suggestion frameworks: - People that share a specific character type may have similar interests and diversions. The items and administrations that are prescribed to an individual ought to be those that have been decidedly assessed by different clients with a comparative for each personality type. For instance, propose to show the vehicle buy goals of clients dependent on their pastimes and character. have built up a framework for prescribing games to players dependent on character which is displayed from their visits with different players.

Upgraded Personal Assistants: - In this Present-day robotized voice aides, for example, Siri, Google Assistant, Al exa, and so forth can be made to naturally recognize the character of the client and, subsequently, give tweaked reactions. Likewise, the voice colleagues can be modified so that they show various characters dependent on the client character for higher client fulfillment.

V. Personality Prediction from text review

Personality detection is where data about a person's character attribute is recognized, given a lot of information. There have been a few ways to deal with automated character prediction dependent on various types of datasets, for example, social media post, face Tube, speech, smartphone, video, essays, handwriting, travel pattern, gender, age. This paper will mainly focus to review personality prediction using text dataset.

A. Baseline Methods for Text

The following subsections summarize the models, dataset and techniques which had been used in machine learning, deep learning-based personality detection on text modality.

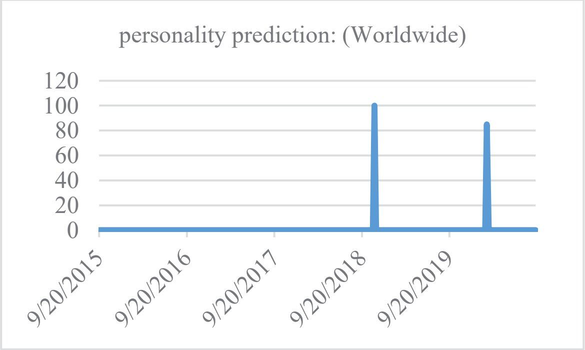


Fig. 2. Graph Personality prediction interest over time - Google Trends

a) Twitter - text Dataset

Most of the personality prediction research studies was attempted on Twitter. In the year of 2016, the authors in this paper [5] This study was used text classification to predict personality based on text written by Twitter users. Dataset consists of last 1, 000 texts in the form of tweets and re-tweets. Collection of tweets from users is also made into a single document/ one long string, then it was preprocessed and labeled according to Big Five personality dimensions. The languages used for this study are English and Indonesian. Classification methods was implemented. Those are Naive Bayes, K-Nearest Neighbors and Support Vector Machine. Finally testing conducted using 10-fold cross-validations. Testing results showed Naive Bayes method was slightly outperformed the other methods.

In the year of 2017, the authors in this paper [6] The aim of this research is to analyze how twitter (dataset) can be utilized to improve the user experience in character assessment. propose a manner by which the client's character can be anticipated through information mapping accessible to general society on their own twitter utilizing DISC (Dominance, Influence, Compliance, Steadiness) assessment. Text mining and sentiment analysis were performed for every user dependent on his/her ongoing tweets. Downloaded more than 1,000,000 tweets utilizing catchphrases.

In the year of 2017, the authors in this paper [7] to build a personality prediction framework dependent on a Twitter user's data for Bahasa Indonesia, the native language of Indonesia. It's possible without a tool with predefined words (LIWC, MRC) but by assessing the user's choice of words. The personality prediction framework is based on Support Vector Machine and XGBoost prepared with 329 instances of users. Assessment results utilizing 10-fold cross-validation shows that the framework figured out how to arrive at the most elevated normal exactness with SVM and XGBoost. To build personality prediction used the five-factor model which also known as big five model. This framework built on XGBoost managed to perform significantly better than on SVM.

In the year of 2018, the authors in this paper [8] presented optimization techniques for automatic personality recognition based on Twitter in Bahasa Indonesia. Evaluated a progression of techniques implementing hyperparameter tuning, feature selection, and sampling to improve the machine learning calculations utilized. The personality forecast framework is based on machine learning algorithms and used big five model. There are three machine learning calculations utilized in this study, to be specific Stochastic Gradient Descent (SGD), and two ensemble learning calculations, Gradient Boosting (XGBoost), and stacking (super learner). By executing this arrangement of optimization strategies, the current examination's assessment results show immense improvement by accomplishing 1.0 ROC AUC score with SGD and Super Learner.

Year - Model	Methodology
[5] 2016- Big Five	Naive Bayes, K-Nearest Neighbors and Support Vector Machine
[7] 2017- Big Five	Support Vector Machine and XGBoost
[14] 2018- Big Five	XGBoost, Logistic Regression, SVM
[8] 2018- Big Five	Stochastic Gradient Descent (SGD), and two ensemble learning calculations, Gradient Boosting (XGBoost), and stacking (super learner)

Table 1. Popular Twitter Datasets, Model, Methodology

b) Facebook - text Dataset

Personality detection had attempted on Facebook. In the year of 2017, the authors in this paper [9] proposed approach deep learning based Personality Recognition from Facebook Status Updates. Author had been investigating several neural network architectures such as fully- connected (FC) networks, convolutional networks (CNN) and recurrent networks (RNN) on the *myPersonality* shared task and compared them with some shallow learning algorithms. Finally, experiments were showed that CNN with average pooling is better than both the RNN and FC [9]. For this study used big five personality traits and the dataset used for experiments is a subset 250 users with 9917 status updates. The dataset includes Facebook statuses in raw text, author information and gold standard Big-5 personality labels.

In the year of 2017, the authors in this paper [10] is suggested the approach Personality Prediction System from Facebook Users. In this research Big Five Model Personality is used. While other previous researches were used older machine learning model used in this research is Big Five Model Personality. Previous researches were used older machine learning algorithm in building their models but this research to implement some deep learning architectures to check the comparison by doing comprehensive analysis method through the accuracy result. The results are succeeded to outperform the accuracy of previous algorithm in building their models, this research tries to implement some deep learning architectures to see the comparison by similar research with the average accuracy of 74.17% [10].

In the year of 2018, the authors in this paper [11] is presented personality predictions based on user behavior on the Facebook social media platform and measures of the big five model. Author analyzed and compared machine learning models and performed the correlation between each of the feature sets and personality traits. The study is constructed with 250 users and 9917 status updates from the *myPersonality* sample and the dataset labeled according to the big five model. The outcomes for the prediction accuracy showed XGBoost algorithm performed significantly better than which built on logistic regression, the gradient boosting classifier or the SVM.

Year - Model	Methodology
[9] 2017- Big Five	Neural network architectures such as fully- connected networks, convolutional networks (CNN) and recurrent networks (RNN)
[10] 2017- Big Five	Deep learning architectures
[11] 2018- Big Five	XGBoost, Logistic Regression, SVM

Table 2. Popular Facebook Datasets, Model, Methodology

c) Essay - text Dataset

Another personality detection had also attempted on essays. In the year of 2017, the authors in this paper [4] is presented an approach to extract personality attributes from stream-of-consciousness essays utilizing a convolutional neural organization (CNN). It used James Pennebaker and Laura King's stream-of-consciousness essay dataset. These datasets have 2,468 anonymous essays tagged with the authors' personality traits was experimented with the remaining 2,467 essays. Prepared five distinct networks, all with a similar architecture, for the five (big five model) personality attributes. Each network was a binary classifier that anticipated the comparing quality to be positive or negative. Finally built up a novel document modeling technique dependent on a CNN feature extractor. Took care of sentences from the essays to convolution filters to acquire the sentence model as n-gram include vectors then represented to every individual essay by amassing the vectors of its sentences. Then is concatenated the acquired vectors with the Mairesse features, which is extracted from the texts legitimately at the preprocessing stage; this improved the technique's performance. Finally, for the classification, this document vector is fed into a fully connected neural network with one hidden layer. Results outperformed the current state of the art for all five traits [4].

For the textual modality, data preprocessing is a significant advance and choosing the right technique will yield essentially better outcomes. Generally, features from text are extracted, for example, Linguistic Inquiry and Word Count (LIWC) , Mairesse, Medical Research Council (MRC), and so forth which are then taken care of into standard ML classifiers, for example, Sequential Minimum Optimizer, Support Vector Machine (SVM), Naive Bayes, and so on. Learning word embeddings and speaking to them as vectors (with GloVe or Word2Vec) is likewise a normally followed approach. These word vectors may likewise be made by taking care of the word character-wise into a Long Short-Term Memory (LSTM) or a Gated Recurrent Unit (GRU). It was seen that consolidating text features (LIWC, MRC) with something different, for example, practical information, convolutions, and so forth brings about better execution [2].

Year - Model	Methodology
[4] 2017- Big Five	Novel document modeling technique based on convolutional networks (CNN)

Table 3. Popular Essay Datasets, Model, Methodology

VI. Conclusion

This paper is gave reader an understanding on existing endeavors of the task of personality prediction specifically from text dataset to-date, alongside the different sorts of twitter, Facebook social medias which have been used for said task. In future can be combined different social media dataset to filter fake account. A portion of these strategies utilize a closed-vocabulary approach with psycholinguistic devices, for example, LIWC, while different techniques utilized an open-vocabulary approach by extricating n-grams and topics. While most personality detection studies to-date require a dataset to perform supervised learning, it is exorbitant to acquire a dataset labelled with personality traits. Recently have been taken a stab at applying semi- supervised and unsupervised learning out how to handle this issue. Further enhancements to the current condition of personality prediction can be made by extending the objective language, applying more appropriate algorithms or preprocessing strategies to accomplish to get higher accuracy. In future robust deep learning models can build then accuracy can be improved.

References

[1] Hetal Vora, Mamta Bhamare, and Dr. K. Ashok Kumar, “Personality Prediction from Social Media Text: An Overview,” *Int. J. Eng. Res.*, vol. V9, no. 05, pp. 352–357, 2020, doi: 10.17577/ijertv9is050203.

[2] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, “Recent trends in deep learning based personality detection,” *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2313–2339, 2020, doi: 10.1007/s10462-019-09770-z.

[3] hernandez and knight, “Predicting MBTI from text.”

[4] N. Majumder, I. P. Nacional, A. Gelbukh, and I. P. Nacional, “Deep Learning-Based Document Modeling for Personality Detection from Text,” 2017.

[5] B. Y. Pratama and R. Sarno, “Personality classification based on Twitter text using Naive Bayes, KNN and SVM,” *Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015*, pp. 170–174, 2016, doi: 10.1109/ICODSE.2015.7436992.

[6] N. Ahmad and J. Siddique, “Personality Assessment using Twitter Tweets,” *Procedia Comput. Sci.*, vol. 112, pp. 1964–1973, 2017, doi: 10.1016/j.procs.2017.08.067.

[7] V. Ong *et al.*, “Personality prediction based on Twitter information in Bahasa Indonesia,” *Proc. 2017 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2017*, vol. 11, pp. 367–372, 2017, doi: 10.15439/2017F359.

[8] G. Y. N. N. Adi, M. H. Tandio, V. Ong, and D. Suhartono, “Optimization for Automatic Personality Recognition on Twitter in Bahasa Indonesia,” *Procedia Comput. Sci.*, vol. 135, pp. 473–480, 2018, doi: 10.1016/j.procs.2018.08.199.

[9] J. Yu and K. Markov, “Deep learning based personality recognition from Facebook status updates,” *Proc. - 2017 IEEE 8th Int. Conf. Aware. Sci. Technol. iCAST 2017*, vol. 2018-Janua, no. January, pp. 383–387, 2017, doi: 10.1109/ICAwST.2017.8256484.

[10] T. Tandera, Hendro, D. Suhartono, R. Wongso, and Y. L. Prasetio, “Personality Prediction System from Facebook Users,” *Procedia Comput. Sci.*, vol. 116, pp. 604–611, 2017, doi: 10.1016/j.procs.2017.10.016.

[11] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Personality Predictions Based on User Behavior on the Facebook Social Media Platform,” IEEE Access, vol. 6, pp. 61959–61969, 2018, doi: 10.1109/ACCESS.2018.2876502.

[12] Hetal Vora, Mamta Bhamare, and Dr. K. Ashok Kumar, “Personality Prediction from Social Media Text: An Overview,” *Int. J. Eng. Res.*, vol. V9, no. 05, pp. 352–357, 2020, doi: 10.17577/ijertv9is050203.

[13] K. Yamada, “Incorporating Textual Information on User Behavior for Personality Prediction,” pp. 177–182, 2019.

[14] V. Ong, A. D. S. Rahmanto, Williem, and D. Suhartono, “Exploring personality prediction from text on social media: A literature review,” *Internetworking Indones. J.*, vol. 9, no. 1, pp. 65–70, 2017.

Retinal Image Analysis to Detect and Classify the Stages of Diabetic Retinopathy

H.A.T Uthapala
Department of Computing & Information System,
Sabaragamuwa University of Sri Lanka
Belihuloya,Sri Lanaka
thiliniuthpala207@gmail.com

Dr.R.M.K.T Rathnayaka
Department of Physical Science & Technology,
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
kapila.tr@gmail.com

Abstract— **Diabetic retinopathy (DR) is an eye disease caused by elevated blood glucose. Most of the working-age people are suffering from diabetics. In certain DR patients, early diagnosis and adequate care may decrease vision loss. The seriousness of the condition should be determined for the delivery of the proper treatment until the signs of DR are identified. The manual process to detect the DR severity stage is needed an experienced clinician & it is taken lots of time. Most of the researchers are compared their prediction accuracy using different classification but this paper predicting the accuracy rate using various image enhancement techniques. This paper introduces a novel method that can detect the DR severity stages without expertise & less time consumption. Using a different type of image enhancement techniques to improve the image quality & using a convolutional neural network to classify the diabetic retinopathy stage accurately in minimum time consumption. In the preprocessing stages, using the CLAHE image contrast technique gamma correction to improve the quality of the image dataset. This paper using CNN used VGG 16 architecture to predict the best results in each type of stage. According to a different type of pre-processing stage novel, image enhancement image is provided a 95.12% accuracy rate rather than others.**

Keywords— Image Processing, Machine Learning Techniques, Contrast Limited Adaptive Histogram Equalization

I. Introduction

Diabetes is the seventh deadliest disease, according to the World Health Organization (WHO). In 1980, only 108 million patients were diabetic, while in 2018 the community of people with diabetes was 4 times as high as 422 million. In diabetic patients, over 18 years of age an increase has been reported, from 4.7 percent to 8.5 percent, as per survey data. The chief factor for diabetic suffering in Sri Lanka[1].In Sri Lanka, around 4.4% of human beings have been diagnosed as diabetic sufferers within the age group of 20 – 79.

Diabetes is caused by a blood glucose boom inside the blood stage. This abnormality will lead to irreversible damage to the blood vessels for a long period. A diabetic person is susceptible to kidney failure, vision failure, bleeding gums, seizure of the limbs, foot injuries, and nerve damage. Among most diabetic patients, there is an excessive risk of coronary heart attack and stroke.[2]

Diabetic retinopathy is clinically validated into four stages.

1.Mild Nonproliferative Retinopathy: During this previous stage, MA is produced. These Mas are small red spots due to the weakening of the blood vessels and are characterized by the presence of MA and HA "spot" and "spots" on the retina during the eye examination. Overall, 40% of diabetic patients have mild signs of DR.[3]

2.Moderate Nonproliferative Retinopathy: A few internal organs which nourish and optic nerve are obstructed as the infection progresses. The characteristic features of this stage are HA and hard exudates. PDR would develop soon for patients with moderate NPDR.[3]

3.Severe Nonproliferative Retinopathy: The blood vessels in the retina are blocked and therefore send signals to the body for new blood vessels to grow for food. During this stage, intraregional MA, HA, venous rims, and intraregional microvascular abnormalities occur in at least two quadrants of the retina. About 50% of severe NPDR would progress to PDR in one year.[3]

4.Proliferative diabetic retinopathy (PDR): When blood vessels proliferate, severe NPDR enters an advanced stage called PDR. Lack of oxygen triggers the growth of new, thin, and fragile blood vessels on the surface of the retina called neovascularization. Due to the fragile walls, blood vessels can leak blood into the vitreous and cause blindness.[3]

Detecting Diabetic Retinopathy disease stages manually from retinal images is a very difficult task and time-consuming work for Doctors. Developing an automated system is to minimize the period to determine the stages of the disease from retinal images. Manual detection requires more specialists' opinions and processes before arriving at a decision. By this automated system, we can reduce the labor Intensive and period that the Doctor must spend. Early treatment can be done if the disease is identified in the initial stage which reduces the patient's life risks. This proposed system reduces the time taken for detecting the disease manually. In some instances, during the treatment process ambiguity may occur. A current method dynamically accurately detects the different diseases.

II. Related work

Almost all research activities have been carried out to identify diabetic retinopathy at an initial point. Researchers tried, first of all, to make a solution to this problem by using traditional computer vision and machine learning methods. For example, Priya [4] proposed a computer-based vision approach to the detection by colored background images of diabetic retinopathy. They tried using image processing technologies to extract features from the image and supplied them with the binary classification SVM, achieving 98 percent susceptibility, 96% specificity, and 97,6 percent accuracy in a 250-image test set. Furthermore, [5]other researchers tried to fit other multi-class models for images and adjustments, for example, with PCA. Decision trees, naive Bayes, and k-NN (2012) with better results 73.4% and 68.4% for measuring F while taking 151 different-resolutions images with a dataset. Different methods use CNN to solve this problem, as deep learning approaches become ever more widely known.[3]have developed a net with CNN architecture and data extension which can detect and automatically and without user intervention diagnosis of the complex functions involved in the classification process, like micro-aneurysms, exudates, and retinal hemorrhages. Sensitivity reached 95% and precision 75% with 5,000 Validation pictures. Additionally, other researchers have been working on CNN[3]. It should be noted that Asiri [6] reviewed a considerable number of available methods and data sets, showing their advantages and disadvantages. They also noted the challenges to be tackled in the design and training of efficient and robust drill-down algorithms for various DR diagnostic problems and highlighted suggestions for future work.

III. Methodology

A. DataBase

We collected a 3,587 dataset to Train, Test, and Validate the Neural Network. The five relatively unstable Diabetic Retinopathy image classes are divided into five subclasses. The table and diagram below show the number of images. Our data set includes images taken with different lights. Five training courses take place in Labels [0,1,2,3,4] with the patient's right and left eye names "right" and "left" and with unique patient identification. Names are marked with normal DR, moderate DR, severe DR, and PDR. The data set we use is collected from a contest held at kaggle.com[7],[8].

TABLE I. NUMBER OF IMAGES ON EACH STAGE

Class	Name	Number of Images
0	Normal	799
1	Mild DR	810
2	Moderate DR	800
3	Serve DR	399
4	PDR	769

A big part of evaluating the machine learning model is the separation of data into preparation, testing, and validation sets. Training Dataset which is the data sample for the model. This is the data set we use to train the algorithm (in the case of a neural network, weights, and biases). This data is interpreted and taught by the developer. A Validation Data sample is used to evaluate a model that fits into the training dataset without prejudice while setting model parameters. Data sampling the assessment becomes more prejudicial when expertise becomes integrated into the model configuration on the validation dataset.

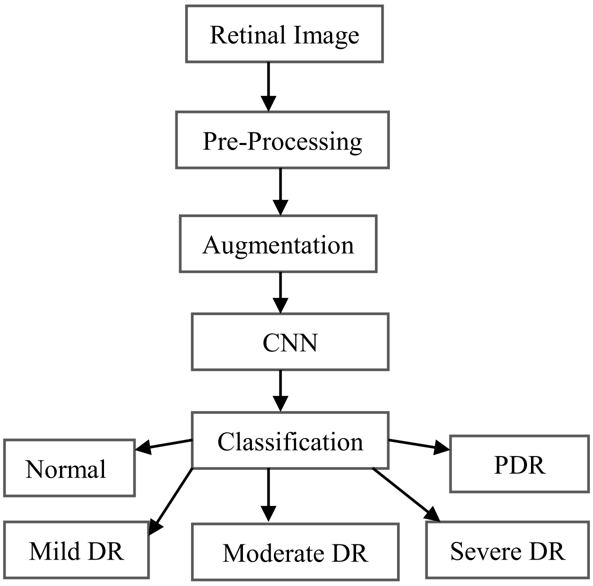


Fig. 1. Proposed Methodology

Preprocessing is the main and important part of this research. Because we can only get the best accuracy results within clear images. Most of the research paper convert retinal image into RGB channel and get the green channel image within it and training the CNN model using the green channel. But Grayscale image provides a better background contract than an RGB image. Most of the researchers did not use the Adaptive Histogram Equation after converting to the grayscale image and we also used gamma correction after CLAHE. It improves the image quality. So, we can get the best result using CNN base on the Inception V3 model and VGG 16 model. Using the Inception of V3 and VGG 16 architecture of Google Net is designed to perform well even under strict constraints on memory and computational budget. Plan to get more accurate results than other research. It takes less time consumption to segment the Diabetic Retinopathy stages.

B. Pre-Processing

Preprocessing is a crucial phase in the handling of retinal fundus images, as these images usually encounter difficulties such as bad contrasts, light changes, and noise effects. The main aim of preprocessing is to improve the input images by reducing intensity variations to present normalized pictures. Improved images are less bruising, making analysis, and displaying easier. We used Kaggle's dataset with 3,587 pictures in diabetic retinopathy. The data collection includes images of different sizes. We have also employed different methods for data collection preprocessing.

We used different types of pre-processing techniques.it is shown in the following diagram.

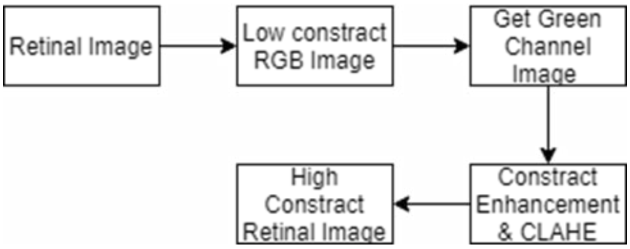
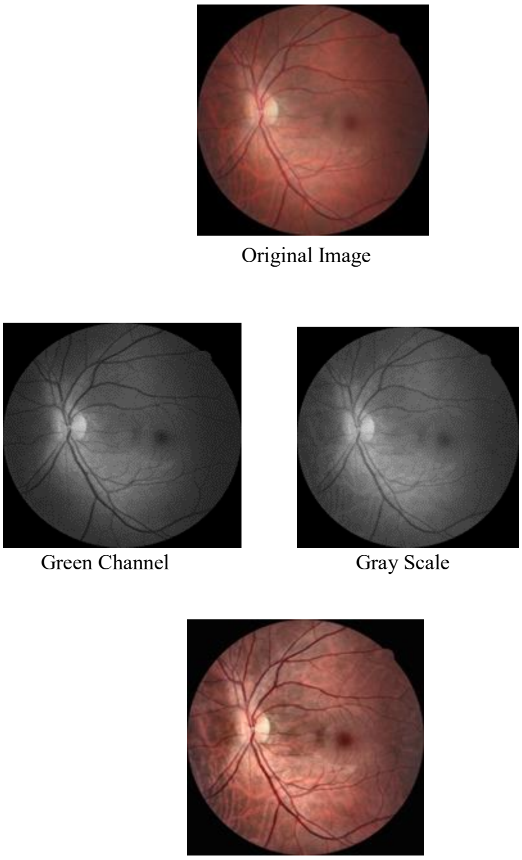


Fig. 2. Step of one preprocessing technique


Original Image

Green Channel

Gray Scale

CLAHE Image

Fig. 3. Above Images provided different pre processing stages Images

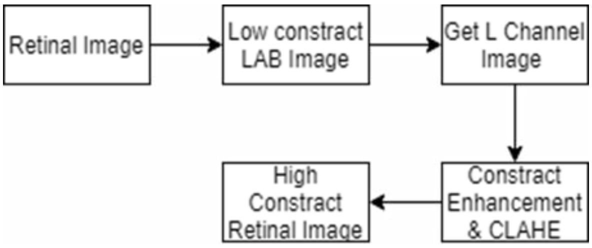
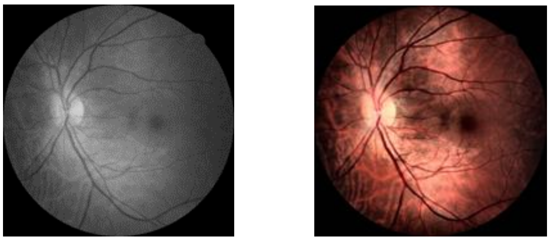


Fig. 4. L Channel image enhancement


L Channel

CLAHE Image

Fig. 5. Novel preprocessing method output Images

Then we used CNN Neural Network Architecture to predict better accuracy among various types stage of images. Using the VGG 16 model to test the accuracy level of the image construction.

C. Convolutional Neural Network

Neural networks of convolution have proven to be effective in the recognition and classification of images in a class of neural networks. By adding convolution, non-linearity, and sub-samples, CNN extends its regular neural networks.[9] The aim is to extract functionalities from the images in the feedback. By mixing the images with specially selected small square matrices, certain processing effects like edge, it may be possible to detect, sharpen, and blur. Another operation could be called Rectified Linear Unit (ReLU)[10].
The nonlinear operation used after each converting operation will be replaced by zero in the feature map for all negative pixel values. ReLU aims to establish nonlinearity. The third operation called pooling or subsampling reduces the size of each characteristic map while keeping the information most important[11]. For instance, the max, average, or sum of a subregion in the function map can be taken and stored. The output feature map will be connected to a traditional neural network to complete the classification task after adding an appropriate number of layers of these operations.

1) VGG 16 Model

VGG16 is a neural network model that evolves with an average 138 million of weighted parameters and 16 weighted layers when compared with AlexNet.VGG16 explores the effect of the convolutional network depth in the large-scale image recognition setting on classification accuracy[12]. VGG16 has the advantage that the construction of hidden layers is simple and standardized: all convolution layers employ three filters with steps 1 and the same. The padding of all pool layers with 2 to 2 filters with phase 2 is a maximum pooling layer[13]. Furthermore, the number of filters in a convolution layer is 2: starting from 64 up to 512.VGG 16 model architecture is shown below.

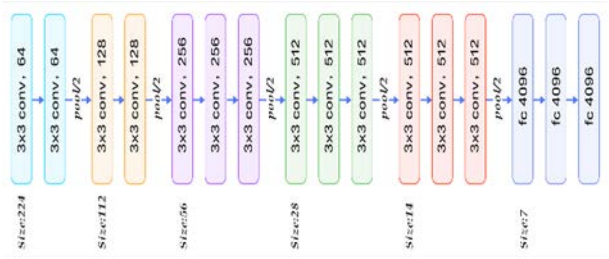


Fig. 6. VGG 16 Model Architecture

Input five stages of the dataset into the VGG 16 model & training the model to identify the different stages of features and inputting the testing dataset to the VGG 16 model to extract the characteristic of the images store it in an appropriate model to classify the different stages. Then input the retinal images in different pre-processing stages & predicting the accuracy of each stage. It is predicted the stage of the diabetic retinopathy according to enhanced retinal image.

IV. Result and Discussion

TABLE II. Results of Varies image prediction accuracy

Pre-Processing Stage	Prediction Accuracy
Green Channel Image	87.37%
Green Channel Gray Scale Image	87.43%
CLAHE Image	93.36%
L channel CLAHE Gamma correction Image	95.12%

As a VGG 16 model result, L channel Gamma correction Image is provided better prediction results than other pre-processing stages retinal images. CLAHE is provided 93.36% accuracy that is closely near to the new developing image enhancement retinal image accuracy rate. Green Channel is provided less accuracy than other Images. we converted all retinal image dataset into L channel CLAHE Gamma correction Images and training the VGG 16 model using those images to get better results. Novel image enhancement image provided a better accuracy rate.

TABLE III. SUMMARY OF THE VGG 16 MODEL ACCURACY

Stage of DR	Precision	Recall	F1-Score
Normal	0.82	0.93	0.87
Mild	0.58	0.34	0.43
Moderate	0.61	0.49	0.55
Serve	0.70	0.75	0.72
PDR	0.80	0.61	0.69

According to VGG 16 accuracy, the Normal DR class is provided more precision than other classes. Mild DR is provided less accuracy than other classes and Normal DR is provided high accuracy.

F1-Score than other classes. If the classification retinal image has not DR it provides better accuracy than others. Precision, Recall, and F1-score is highest in the normal retinal eye.

V. Conclusion

The proposed new image enhancement technique helps to identify the severity stage of the Diabetic Retinopathy that affects the patient who is a victim of diabetes. This algorithm helps to void the blindness of the patient and Doctors can be treated the patient earlier is saved the time slot that was spent in the Doctors who wanted to identify the Diabetic retinopathy stages. Proposed image enhancement techniques are provided better construction of eye images and giving better feature details of the retinal images. Proposed model train and validate using enhanced retinal images data set and it provided better accuracy performance than other models. Most of the researchers focused on the different types of classification mythology we are focused on the various type of preprocessing methods. The case whole algorithm is dependent on the retinal image and if a person who can't identify the feature of the images classification result is inaccurate. So, our effort is successful and provided better prediction results.

Acknowledgment

I would like to express my gratitude for the support of the Department of Computing and Information System, Sabaragamuwa University of Sri Lanka, which helped to find this paper internally.

References

[1] S. Kumar and B. Kumar, "Diabetic Retinopathy Detection by Extracting Area and Number of Microaneurysm from Colour Fundus Image," *2018 5th Int. Conf. Signal Process. Integr. Networks, SPIN 2018*, pp. 359–364, 2018, doi: 10.1109/SPIN.2018.8474264.

[2] S. Ekatpure and R. Jain, "Red Lesion Detection in Digital Fundus Image Affected by Diabetic Retinopathy," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–4, 2018, doi: 10.1109/ICCUBEA.2018.8697387.

[3] Y. H. Li, N. N. Yeh, S. J. Chen, and Y. C. Chung, "Computer-Assisted Diagnosis for Diabetic Retinopathy Based on Fundus Images Using Deep Convolutional Neural Network," *Mob. Inf. Syst.*, vol. 2019, no. 1, 2019, doi: 10.1155/2019/6142839.

[4] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," *Procedia Comput. Sci.*, vol. 90, no. July, pp. 200–205, 2016, doi: 10.1016/j.procs.2016.07.014.

[5] C. Grosan and A. Abraham, *Machine Learning*, vol. 17. 2011.

[6] T. Shanthi and R. S. Sabeenian, "Modified Alexnet architecture for classification of diabetic retinopathy images," *Comput. Electr. Eng.*, vol. 76, pp. 56–64, 2019, doi: 10.1016/j.compeleceng.2019.03.004.

[7] "Search | Kaggle." https://www.kaggle.com/search?q=diabetic+retinopathy+in%3Adatasets (accessed Sep. 10, 2020).

[8] "Diabetes Asia." https://www.diabetesasia.org/ (accessed Jul. 30, 2020).

[9] S. Dutta, B. C. S. Manideep, S. M. Basha, R. D. Caytiles, and N. C. S. N. Iyengar, "Classification of diabetic retinopathy images by using deep learning models," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 1, pp. 89–106, 2018, doi: 10.14257/ijgdc.2018.11.1.09.

[10] C. Mahiba and A. Jayachandran, "Severity analysis of diabetic retinopathy in retinal images using hybrid structure descriptor and modified CNNs," *Meas. J. Int. Meas. Confed.*, vol. 135, pp. 762–767, 2019, doi: 10.1016/j.measurement.2018.12.032.

[11] M. Arora and M. Pandey, "Deep Neural Network for Diabetic Retinopathy Detection," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 189–193, 2019, doi: 10.1109/COMITCon.2019.8862217.

[12] J. Krause *et al.*, "Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018, doi: 10.1016/j.ophtha.2018.01.034.

[13] G. Kumar and P. K. Bhatia, "A detailed review of feature extraction in image processing systems," *Int. Conf. Adv. Comput. Commun. Technol. ACCT*, pp. 5–12, 2014, doi: 10.1109/ACCT.2014.74.

Video Data Preprocessing for Soccer Video Highlight Summarization

Asitha Nanayakkara
Faculty of Information Technology
University of Moratuwa
Sri Lanka
asitha.15@itfac.mrt.ac.lk

C.R.J. Amalraj
Faculty of Information Technology
University of Moratuwa,
Sri Lanka
amalraj@uom.lk

***Abstract*—This Automatic soccer video summarization is a computer vision-related task that has taken a greater interest in terms of accuracy and speed in computation that would rather replace human labour. A system incorporating computer vision-related techniques for understanding video sequence and extracting high-dimensional data to yield useful information for decision making is a huge challenge with respect to expert systems. The three main modules that the proposed system incorporates are Video shot segmentation module, Keyframe extraction module, and Audiovisual descriptor module. The video shot segmentation module partitions a whole video into separate camera shots. The keyframe extraction module extracts important frames within those shots as processing every frame is time-consuming and not essential. The audiovisual descriptor module consists of tools to analyze and extract a set of low and mid-level audio and video descriptors for every video-shot. The proposed system is designed to output a database that consists of relevant data for every video shot with attached meaning to it. In future, that data can be analyzed to create a soccer video summarization according to relevance interest and the likelihood of the shot by keeping the certain video shots in highlight package or removing it. The application of image processing techniques has paved the way for the proposed system to become realistic.**

***Keywords*—key frame extraction, audiovisual descriptors, Computer vision, and video summarization.**

I. Introduction

Producing highlights of a soccer match is a labour-intensive task and requires specialization in the field. The automatic summarization of soccer video data is a field that has marked interest in the past decade. With a large number of international and domestic (La Liga, UEFA) soccer matches played throughout the world, it is difficult for soccer fans to be up to date with all the news and to watch live streaming. Therefore, techniques incorporating content-based video analysis and processing play an important role in addressing a variety of problems including semantic analysis, video retrieval, and video summarization [1] [2]. Manual highlight generation needs professional editing skills and expert sports knowledge and most importantly it is a process that consumes a lot of time. Therefore, the need for a system that automatically does soccer video summarization is fueled.

A soccer match is a series of successive plays of attack and defense, which can be divided into series of video shots each captured from a single camera angle. They make a whole soccer match when added together. The semantically meaningful cues, which are important to be added to the highlight video, need to be distinguished from unimportant cues. The Task may be challenged by the complexity of the game which has a lot of rules and thereby creating a large range of events containing important circumstantial factors. Since the game bears the most popular title on the planet, it is crucial to summarize the sports video sequence in order to minimize the time duration and maintain the quality of the game for the best spectator experience.

Automatic video summarization is done by combining event-driven features where specific rules of the games are taken into consideration and excitement-based actions of players, the crowd, and commentators. We propose this system to generate highlights by identifying video shot boundaries, key frame extraction, zoom detection, replay detection, and loudness detection. These identifications can be reliably recognized by using modern solutions for computer vision, and by video processing to generate a fully functional solution for automatic soccer video summarization.

Football is the most popular sport in the world with a huge fan base across the globe. Among these fans, a considerable amount tries to watch only the important events of a game without consuming too much of their time as the sport is played more often. And a certain amount of the fans wants to watch the same match repeatedly as it brings them enthusiasm and excitement with results and main events of the match. Therefore, highlights would be a better solution for them to seek what they want by spending less time. By considering the above-mentioned facts, the need for a summarized version of a single match is triggered. This background information points out a major factor that motivated us to implement a system to deliver a summarization of soccer matches automatically. Nowadays, people who create highlights have to select important events from the whole match video which needs expert knowledge on soccer scenarios and high editing skills. An automatic software solution can be used to address this problem.

The defects in the prevailing systems motivated the way to propose a system that could address those issues and generate a shorter video clip to have a better experience. The evaluation process of the proposed system is to be done by comparing the broadcaster's highlights and seeking feedbacks of fans on highlights generated by the system.

Soccer fans seek news updates on the matches and players frequently, and they expect them within a shorter time. Fans are not always able to watch the whole match as the playtime of live matches vary with the countries where the matches are played, and they may not be able to watch all the important events of the match with the busy lifestyle. Therefore, most of the fans tend to watch the highlights of the matches whenever they want. Also, when fans seek highlights in the World Wide Web, they often find unreliable highlight providers who do not cover all important events, consider the time duration of the output video but not the quality of the highlights. The need for a quick and reliable highlighting package is motivated to find a solution for this problem domain.

The following objectives were defined to achieve the aim of the required solution.

- To partition the full match video into video shots and extract important cues.
- To analyze derived cues and generate a summary of the match.

The system incorporates a video-shot segmentation module that identifies shot boundaries and key frames. Key frame identification and shot analysis tasks are carried out using a set of audio-visual descriptors. Zoom detection is done through background subtraction technique. Technique. Binarized frames are obtained using the background subtraction, and a derived golden area of each frame is been processed to analyze whether the video shot is the close-up, long view, median shot, or audience shot. Replay detection is done through logo identification which is at the beginning and the end of a replay. Loudness is also detected finally, and the loudness information is attached as true or false to generate a fully analyzed data set for every video shot.

The rest of the paper is arranged as follows; section II provides a critical analysis of the similar video processing systems for highlight generation of games and their respective utilized techniques. Section III explains the methodology which we followed in implementing the system. Then, section IV provides the evaluation of the system and explains how accurate and optimized the provided system is. Finally, section V and VI explain the conclusion of our research and further work to be done on the suggested methodology.

II. LITERATURE REVIEW

Video partitioning and Key frame extraction

Video shot segmentation is the process of partitioning a whole video into several video shots with different time durations. Here the shot boundaries are identified through the change of the camera angle. A video is a combination of sequence of frames where different videos have different numbers of frames per second. These frames are images; therefore, a video is derived as a sequence of images [3]. These images change their pixel representation when going along the video from the beginning to the end. In this case, change in the camera angles could be detected according to the change in pixel-level [13]. A greater change in pixel intensity between two frames will identify if there is a change in the camera angle. If the video is at the same camera angle, the successive frames would not differ in pairs when going along the video in the perspective of pixel intensity. Therefore, video shot segmentation can be defined as obtaining the positions of the video where camera angle changes. A soccer match is recorded from several cameras and the feedbacks from those cameras are added to complement a full match video.

This problem of finding the video shots (the time intervals where camera angle changes) are addressed in the majority of researches for different projects [3]. The partitioning process acts as the basis for many video processing projects as it helps to modify, simplify and understand the overall presentation of images with pixels. By analyzing these researches an optimized system is proposed as a solution which is efficient and effective.

Key frame extraction is the basis for other tasks like video analysis and content-based video recovery. The use of Key frames which gives an identity to each video shot can be seen [4]. The key frames can be defined as the most important frames of a single video shot which can be used to apply image processing techniques to annotate a semantic to whole videoshot. In recent years, many key frame extraction algorithms have focused on the original compressed video stream [3]. This can increase the complexity of the video when the decomposition is needed before the video is processed. The key frame is the image, and it can be a prototype of the content and information of the shoot. The key frames of the video should be chronologically summarized by the key elements of the video. Therefore, there is an essential need for an efficient and secured key frame selection technique in an efficient video analyzing system [3].

Audio-visual Descriptor module

The most interesting events of interest in a football match are usually presented from different platform points of view or slow motion. Therefore, re-identification approaches can provide credible details in the process of summarizing stimuli. The TV production style in soccer media usually identifies the beginning and the end of a replay using two identical logos [10]. Therefore, the strategy to detect a video sequence soccer replay can be based on logo detection and is generally implemented in four stages such as searching for video frames that are candidates to contain a logo, detection of the logo pattern used in the soccer media, matching of the logo pattern along with the soccer video and pairing the detected logos to identify the replays as shown in Fig. 1.

The algorithm used in the first step uses the difference in brightness between images to detect peaks along with video images, which is a characteristic of a logo [12]. This method is suitable for difficult logo transitions; however, the algorithm has been modified to support the firing range detector, as it is poorly programmed for the transition. An algorithm based on the k-means cluster algorithm for the brightness and variation of raster images is employed to identify the logo template. Once the logo templates are detected, the pixels are done by converting the logos and candidate images to gray and by reducing the pixels. The sum of all the differences gives an idea about their similarity [9]. Finally, a threshold is applied to detect the correct matches that present the lowest difference values. A replay is identified if only its beginning and ending logos are detected [8]. In TV production styles that only use single logos, and the pairing logos stage is not applied.

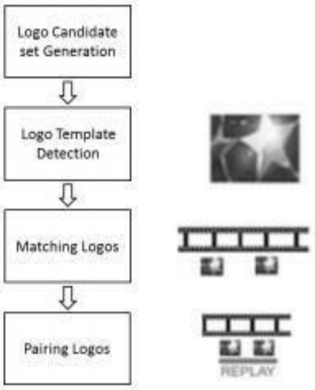


Fig. 1. Replay detection

Zoom Detector

The zoom operation usually indicates an important instant in a soccer match. It is essential to have a zoom detector that may help to find the highlights of the match. Several techniques for zoom detection exist in the prevailing industry; the most common methods involve computing the optical flow, fitting it into a parametric motion model and assessing the corresponding coefficients of the model. Unfortunately, the majority of these steps include iterative techniques that require high computational costs [11]. The research document suggests that the proposed algorithm is derived from the MPEG 360x288 video sequence in the exploitation of motion compensation vectors derived directly from the MPEG stream and the magnification using the motion vector field. Motion vector emissions are eliminated using a 3x3 median filter. Then the algorithm divides the motion field into the golden sections as shown in Fig. 2.

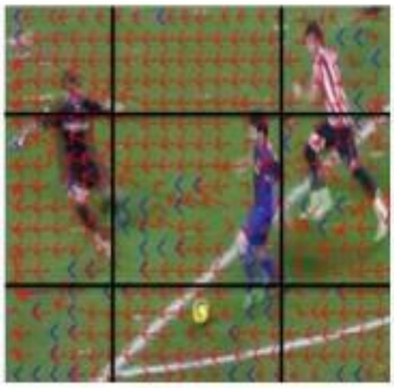


Fig. 2. Zoom detection

Section Hypothesis for zoom detection

After the motion field is divided into the golden sections (Fig. 2), the motion vector that expresses the dominant direction of each area is calculated to find the average of the motion vector coordinates within each section [6]. Finally, each of the leading vectors is configured in a line. And if there is a common focal point between the intersections of these lines, a magnification operation is revealed. This zoom detector has been tested to contain 120 soccer zooms, with 78.33% recall results and 88.68% accuracy. Given that in the application in the research paper a higher precision rate than a higher recall is preferred, the achieved results are considered satisfactory.

Long Shot Detector

The most significant highlights need to be identified. Method to simplify this process is to remove the non-relevant occurrences. Video shots that are low in interest (Fig. 3) for the highlight video of a soccer game mostly provide a panoramic view of the ground. This type of shooting is considered a shot and green is the dominant colour. Dominating green colour can be found also in a mid-shot, which is defined as an intermediate shot between a close-up shot and a long shot.

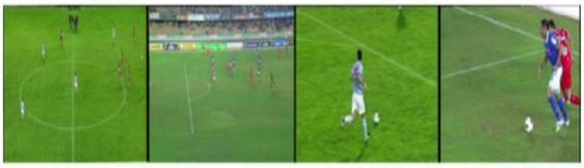


Fig. 3. Long shot detection

The most common approaches in the literature to distinguish the use of long-haul aircraft as a feature of differentiating the percentage of grass in the whole image or the percentage of grass in different parts of the image [4]. The approach consists of analyzing three previously acquired main image stages: first, the relevant information is deleted, then the green dominance is evaluated, and finally, the colour homogeneity is evaluated. The first step is to remove the majority of pixels that are not grass [14]. To do this, one-third of the top of each image is cut, because this part of the image is not modified for analysis and usually corresponds to the general public. The next step is filtering the playing area with green supremacy. This is done by defining the dominant RGB colours of the image and the percentage thresholds corresponding to them. This information is directly obtained from leading MPEG-7 colour descriptors. The third step evaluates colour consistency by analyzing the parameters of the MPEG-7 colour layout description [7]. This colour descriptor reduces the input image to a very small 8x8 YCrCb image and finds the discrete cosine transform coefficient (DCT). In the long-term view, soccer players are shown with a few pixels; so, players can be deleted. This procedure, unlike the mid-plane case, provides a uniform green image in the long-range view. As a counter DC, the system analyzes the homogeneity of the green colour by computing the variance of the first 9 coefficients of the DCT chromosome and comparing them to a low threshold. The deviation of the key frame below the threshold is marked as a distance view [6].

After going through several system implementations and researches, we proposed a method that achieved higher accuracy in detecting features with higher speed.

III. METHODOLOGY

Approach for video shot partitioning module

The input to the proposed system is full match videos of international soccer plays. Each and every video consists of a sequence of video shots that resembles a whole match. Here the term shot is referred to as the feed obtained from a single camera angle. Each shot is composed of a series of continuous frames of visual data. Python and OpenCV are used to accomplish this task. The program implements a threshold-based scene detection algorithm by generating a threshold compared to the average pixel intensity of the frames. That is, the pixel intensities are vastly changed between two successive frames. For each fade detection, relevant frame numbers and the timecodes are obtained.

Abrupt/hard-cut shot detection will be useful during the game and cross-dissolves are used before or after the game. These boundaries must be tackled correctly in order to acquire maximum accuracy in shot segmentation (Fig. 4). Hard cuts are the location at which camera angle changes rapidly. We calculate the Chi-square distance of consecutive histograms going through the whole video [7]. Chi-square distance is 0 if two exactly equal images are compared. This similarity measure helps us in finding the consecutive frames which are vastly different, suggest the change in camera angle. We use a threshold of 40, and if the Chi-square difference is greater than 40, we take output as hard cuts detected in that frame (frame number) and the timestamps are gained relative to the frame numbers and start and the end of each video shot are classified according to the timestamps using them in pairs. These boundaries must be tackled correctly in order to acquire maximum accuracy in shot segmentation.

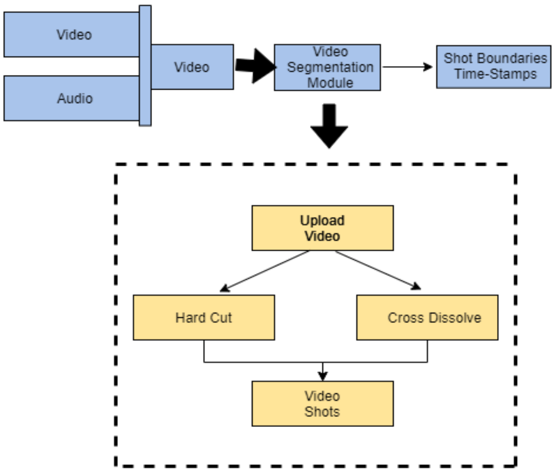


Fig. 4. Video shot partitioning

Key frame extraction

Once the video shots are defined, key frames are extracted. This is because image processing of every frame is not practical enough and it will reduce the performance. This task is implemented in our research to achieve high performance relevant to time as it is a time-intensive task to process every frame. Therefore, we detect representing frames (Fig. 5) of video shots. So, they can be processed and analyzed to annotate meanings to video shots. Frames with high motion intensity are used to analyze the colour and frames with similar colours are discarded.

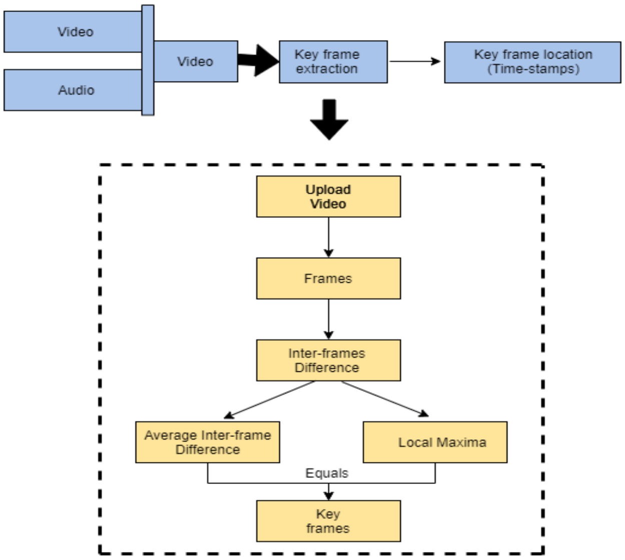


Fig. 5. Key frame extraction

Approach for Audio-visual descriptor module

Replay detector

Broadcasters generate replays to focus on interesting events of a match from different perspectives or in slow-motion. The broadcaster’s style for replays would be displaying an identical logo in the start and at the end of the replay [5]. These instances can be tackled using a transition algorithm aided by a shot boundary detector algorithm.

Zoom detector

Zoom operation defines an important action in the game which would help to find highlights of the video. A motion vector field of 23×18 is extracted first and then the motion vector outliers are removed. After that the motion field is divided into the golden section hypothesis and motion vector along the dominant direction is calculated. A zoom operation (Fig. 6) is detected once there is a focal point at the intersection of the dominant vectors that have been parameterized.

Excitement detection

Excitement detection is a very important aspect when generating gaming highlights. Excitement tends to increase when interesting parts of the video are played by means of audience loudness and commentator's loudness [6]. Loudness, silence and pitch are effective measures of excitement. Our approach is to find the volume of each frame. Here louder, less silence and higher pitch audio frames are identified as excitement intensive frames. Consecutive excited frames are joined to find the starting and ending timestamps of excitement period.

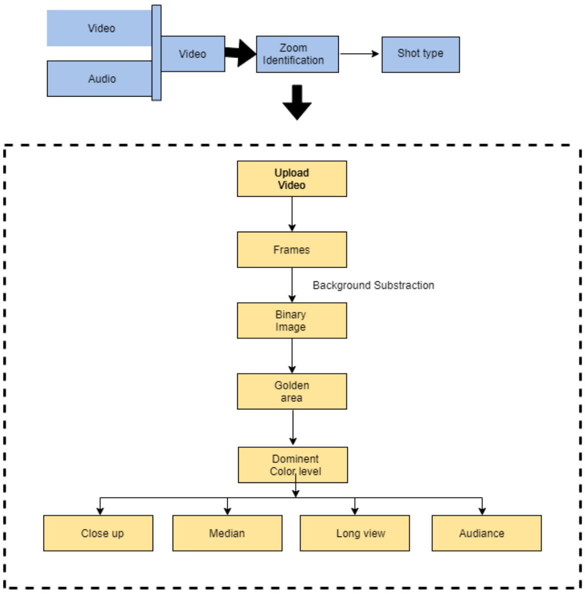


Fig. 6. Zoom detection

IV. EXPERIMENTS AND RESULTS

Experiments have been carried out to check the accuracy of the system using 10 video samples of 15-minute duration "Laliga" soccer league matches. Initially, we manually recorded the change in video shots where there are hard cuts in the video and then compared with the results obtained by the system. The shot detection rate was more than 90% all the time and also it detects less false video shots and misses less true video shots (Table 1).

Table 1. Evaluation of Video shot partitioning

Video sample	Total detected	false	miss	recall	Precision (%)
Soccer_a.mp4	1289	78	32	97.51	93.94
Soccer_b.mp4	1138	65	25	97.80	94.28
Soccer_c.mp4	1309	70	19	98.54	94.64
Soccer_d.mp4	1025	48	24	97.65	95.31
Soccer_e.mp4	754	32	8	98.93	95.75
Soccer_f.mp4	1023	43	14	98.53	95.79
Soccer_g.mp4	988	45	11	98.88	95.44
Soccer_h.mp4	1259	52	26	97.93	93.63
Soccer_i.mp4	496	15	6	98.79	96.97
Soccer_j.mp4	689	16	10	98.54	97.67

Then we carried out an experiment on the accuracy of keyframes extracted as whether they are relevant key frames or the system misses important key frames during the processing of these keyframes is important. Keyframes detected are mostly relevant when compared with frames which are generated through the separation of all the frames from sample videos. It also misses a small number of important frames (Table 2).

When experimenting for replay detection it gave an accuracy of almost 100% all the time as replays are detected perfectly with logo detection. This is the most optimized module in the system which is faster as well (Table 3).

Table 2. – Evaluation of key frame extraction

Video sample	Total detected	false	miss	recall	Precision (%)
Soccer_a.mp4	8936	398	985	88.9	95.5
Soccer_b.mp4	9586	285	654	93.1	97.0
Soccer_c.mp4	10025	852	895	91.1	91.5
Soccer_d.mp4	8965	658	1029	88.5	92.6
Soccer_e.mp4	13548	925	1258	90.7	93.1
Soccer_f.mp4	11985	585	1356	88.6	93.4
Soccer_g.mp4	9654	789	1485	84.6	91.8
Soccer_h.mp4	12658	1003	1658	86.9	92.0
Soccer_i.mp4	11352	1254	1325	88.3	88.9
Soccer_j.mp4	10544	1123	989	90.9	89.3

For the evaluation purpose of zoom detection, we created a matrix which explains the data of the system generated outputs vs manually detected outputs. Table 4 shows the accuracy of four types of video shots such as close-up, median shot, long shot and audience shot. It also shows the number of correct outputs and the number of false outputs regarding the types of video shots. You can go through the main diagonal of the matrix to detect the number of correct outputs obtained.

Table 3. Evaluation of Replay extraction

Video sample	Total detected	false	miss	recall	Precision (%)
Soccer_a.mp4	8	0	0	100	100
Soccer_b.mp4	10	1	1	90	90
Soccer_c.mp4	7	0	0	100	100
Soccer_d.mp4	11	0	0	100	100
Soccer_e.mp4	8	0	0	100	100
Soccer_f.mp4	10	0	0	100	100
Soccer_g.mp4	14	1	1	92.87	92.87
Soccer_h.mp4	12	0	0	100	100
Soccer_i.mp4	7	0	0	100	100
Soccer_j.mp4	9	0	0	100	100

These experiments have been done mostly by splitting the video samples into the frames and comparing with the system generated outputs. According to the expected results and occurred results we have calculated the precision percentages, and they prove that the research we carried out is a success as rates are high enough to come to a conclusion.

Table 4. Evaluation of Zoom Detection and shot type detection

Shot Type	long	medium	close-up	audience
long	925	97	46	15
medium	83	807	17	19
close-up	0	13	450	25
audience	0	3	12	87
recall	91.8	87.1	85.7	59.6
precision	85.4	84.9	86.9	85.3

V. CONCLUSION

This comprehensive research report presents a novel framework to preprocess the soccer video data for generating automatic video summarization of soccer broadcasting video sequences. The approach presented was to identify the video shots generated through partitioning and give them a meaning so that finally they could be given a relevance measure in future. The multiple modules for annotating semantics for video shots would increase the efficiency and accuracy of our work. The semantics is given to the video data using several techniques and results are processed in an optimal way, not to neglect any important highlights. After the evaluation of separate modules, it concluded that the accuracy and efficiency of the proposed framework are at an optimum level suggesting a successful video shot data generation for highlight video creation.

VI. FUTURE WORK

For further work, this work can be developed in such a way that the performance of the system can be further improved as it still takes considerable time to summarize the results. Also, this research is done for certain tournaments of soccer matches and it can be tested with a large number of soccer matches played around the globe. This product is targeted for the people who are creating highlights for other spectators to watch. Television broadcasters of soccer matches and video editors are the main users of this software solution. This can be further developed to make a software product which can be used by common people who are interested in creating soccer match highlights on their own. After doing this comprehensive project based on soccer video summarization through event-driven and excitement-based techniques, the problem of creating automatic video highlights will be widely discussed. Yet no one has released any commanding commercial project. Therefore, this product can be commercialized with proper management and maintained for common people to use according to their interest in the future.

REFERENCES

[1] Hu W, Xie N, Li L, Zeng X, Maybank S (2011) A Survey on visual content-based video indexing and retrieval. IEEE Transactions on Systems, Man, and Cybernetics. Part C: Applications and Reviews Vol. 41, no. 1, pp. 797-819.

[2] Datta R, Joshi D, Li J, Wang J Z (2008) Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, Vol. 40, no. 2, Article 5.

[3] Zawbaa, Hossam & El-Bendary, Nashwa & Hassanien, Aboul Ella & Kim, Tai-Hoon. (2012). Event Detection Based Approach for Soccer Video Summarization Using Machine learning. International Journal of Multimedia and Ubiquitous Engineering (IJMUE). 7.

[4] Harb, Hadi. (2009). Highlights detection in sports videos based on audio analysis.

[5] Arxiv.org. (2020). [online] Available at: https://arxiv.org/ftp/arxiv/papers/1411/1411.6496.pdf [Accessed 29 Feb. 2020].

[6] Zawbaa, Hossam & El-Bendary, Nashwa & Hassanien, Aboul Ella & Kim, Tai-hoon. (2011). Machine Learning-Based Soccer Video Summarization System. Communications in Computer and Information Science. 263. 10.1007/978-3-642-27186-1_3.

[7] Openaccess.thecvf.com. (2020). [online] Available at: http://openaccess.thecvf.com/content_cvpr_2018_workshops/papers/w 34/Shukla_Automatic_Cricket_Highlight_CVPR_2018_paper.pdf [Accessed 29 Feb. 2020].

[8] People.cs.vt.edu. (2020). [online] Available at: http://people.cs.vt.edu/~zhaozhao/papers/vip.pdf [Accessed 29 Feb. 2020].

[9] Python, B. (2020). Guide to Automatic Highlight Generation in Python. [Online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2019/09/guide-automatichighlight-generation-python-without-machine-learning/ [Accessed 29 Feb. 2020].

[10] Eldib, Mohamed & S.Abou Zaid, Bassam & Zawbaa, Hossam & El-zahar, Mohamed & El Saban, Motaz. (2009). “Soccer video summarization using enhanced logo detection.” Proceedings / ICIP ... International Conference on Image Processing. 10.1109/ICIP.2009.5413649.

[11] D.Yow, B.L.Yeo et al., “Analysis and presentation of soccer highlights from digital video,” in Proc. ACCV, 1995.

[12] M. E. Anjum, S. F. Ali, M. T. Hassan and M. Adnan, "Video summarization: Sports highlights generation," INMIC, Lahore, 2013, pp. 142-147

[13] Ahmet Ekin, A. Murat Tekalp, Fellow, IEEE, and Rajiv Mehrotra,” Automatic Soccer Video Analysis and Summarization”, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 12, NO. 7, JULY 2003

[14] Tabii, Youness & Rachid, Oulad haj thami. (2009). A Framework for Soccer Video Processing and Analysis Based on Enhanced Algorithm for Dominant Color Extraction. International Journal of Image Processing. 3.

PLATINUM SPONSER